Survey-Scale Galaxy Chirality with Equivariant TTA: A -0.12σ Subsample-Mask $\ell=1$ Null, a Quantifiable Monopole-Mask Leakage Channel, and a Three-Interpretation Closure of the Canonical-Mask Residual on 8.47 Million DESI Legacy Galaxies (3.2 Million Spirals)

Houston Golden[1, *]

[1] *Independent Researcher, Los Angeles, California, USA*

(Dated: May 22, 2026 PDT — v1.0.128)

We report a multi-survey, equivariance-corrected angular dipole analysis of **3,201,160** DESI Legacy spiral galaxies (8.47 M sources, 471 049 high-confidence per-spiral after $p_{\rm CW}^{\rm eq} > 0.9$). The headline scientific result is a **null** $\ell=1$ chirality-dipole observable on the analysis subsample mask: the MASTER-deconvolved single-mode pseudo-C_1 on the strict-superset subsample mask ($n=5{,}547{,}858$, $f_{\rm sky}=0.659$) yields -0.12σ, consistent with no dipole at $\ell=1$. The real-space post-TTA Catalog C dipole is $+0.43\sigma$ ($p=0.30$, $\sim 0.6\%$ residual amplitude). We emphasize at the outset that this $\ell=1$ observable is the *isotropy-breaking axial-vector channel* and is *parity-EVEN*: it is NOT a direct parity-violation test (the parity-odd analog requires 3D spin-vector or polarization-rotation cross-correlation observables outside this paper's scope). Prior literature has at times conflated these two channels; we keep them strictly separated.

A canonical-mask diagnostic of the leakage mechanism shows that pre-MASTER raw pseudo-C_1 in the un-monopole-subtracted CW-fraction map ($f_{\rm sky}=0.49005$, $N_{\rm spiral}=3{,}201{,}160$) is reproduced at 99.3% of its observed amplitude by a controlled monopole-only generative null ($N=500$, binomial realizations at $p_{\rm CW}^{\rm global}=0.4974$ on the canonical mask): a small uniform CW-vs-CCW classifier monopole couples through patchy survey-mask geometry to inflate the raw pseudo-C_ℓ at $\ell=1$, and MASTER mode-coupling deconvolution removes the leakage. The post-MASTER canonical-mask direct-MC residual is $+3.64\sigma$ under proper galaxy-weighted monopole subtraction (the legacy $+1.85\sigma$ v1.0.62 baseline was on the uncorrected A_p field; the corrected $+3.64\sigma$ replaces it as the canonical canonical-mask number — v1.0.107+ paper-wide convention).

Three interpretations of the canonical-mask residual are systematically tested with a four-null battery + direct cross-spectrum: (i) a clean real cosmological dipole at amplitude $\sim 1.7\%$, (ii) a coherent depth/sampling-correlated systematic at low ℓ on the patchy canonical footprint, or (iii) a NaMaster low-ℓ deconvolution artifact. Interpretation (iii) sharp-edge variant is disfavored by the apodized-mask test ($+3.57\sigma$ on C^2 2° apodization, essentially unchanged from binary). Interpretation (i) as a clean dipole-only explanation is disfavored by three independent lines of evidence: (a) $\ell=2>\ell=1$ broadband structure (auto-spectrum $+4.73\sigma$ at $\ell=2$ vs $+3.63\sigma$ at $\ell=1$ — a real dipole at 1.7% would be $\ell=1$-dominant, with injected real dipoles showing $\sigma=+2.87$ at $\ell=1$ under the same null); (b) absence of monotonic $p_{\rm eq}$ quality-quartile scaling ($\sigma=+0.20, -0.42, +0.44, +0.43$ across four equal-N quartiles, all $|\sigma|<1$); (c) direct cross-spectrum $C(A_p \times n_{\rm total})$ at $\ell=2$ gives $r=-0.65$ with $\sigma=-2.89$ against permutation null (quadrupole-anti-alignment between the chirality asymmetry map and the pixel-density / depth proxy at the same multipole where the auto-spectrum has its largest excess); at $\ell=1$ the same cross-spectrum gives $r=-0.49$ with $\sigma=-1.53$, below the conventional detection threshold individually but with the same negative sign as the $=2$ signal, consistent with a broadband low-ℓ depth-correlated systematic. Interpretation (ii) is therefore the favored verdict: a coherent depth/sampling-correlated systematic at low ℓ on the patchy canonical footprint, directly confirmed at $\ell=2$ by the cross-spectrum and consistent at $\ell=1$. These tests do not exclude a small primordial dipole component sitting beneath the canonical-mask systematic; what they exclude is a clean dipole-only explanation that reproduces all three diagnostic patterns simultaneously. A bootstrap pixel-resample test gives -0.22σ for the canonical-mask data but is tautological for cosmological-dipole hypothesis testing (a real injected $A=1.7\%$ dipole also yields median $\sigma=-0.49$ under bootstrap; the bootstrap variance is too wide to discriminate) and is therefore retained only as a sampling-variance diagnostic.

Falsification criterion. Detection of a chirality dipole at $\sigma>5$ at amplitude $\gtrsim 0.75\%$ (the demonstrated 50%-recovery-at-3σ threshold on the present HC subsample injection sweep; 0.5% is a tested non-detection point at the present pipeline, not the operational floor) in a future $\geq 10^7$-galaxy systematics-modeled survey would falsify the present null; the floor tightens under LSST sample-size scaling. **Scope.** The $\ell=1$ subsample-mask null is the load-bearing scientific result; the canonical-mask residual is interpretation (ii) systematic, not a primordial detection. A like-for-like matched-footprint Ganalyzer reanalysis under Shamir's pipeline + cuts is required for a likelihood-level exclusion under his estimator; we do not perform that reanalysis here. The catalog (3.2 M spirals), model weights (ViT-Small with Z_2 2-fold flip TTA; full D_4 TTA tested on holdouts only), and all reproducibility scripts are publicly released under the immutable release tag `paper4-v1.0.122` (and

predecessor patch tags v1.0.113–v1.0.117).

PACS numbers: 98.80.-k, 98.62.Ai, 95.75.Mn

CONTENTS

I. Introduction 2

II. Data 4
A. Galaxy Images 4
B. Training Labels 5

III. Methods 6
A. Pre-Registered Analysis Hierarchy 6
B. Model Architecture 7
C. Training 7
D. Flip-Equivariance Consistency Loss 9
E. Test-Time Equivariant Averaging 9
F. Bias Hardening Suite 11
G. Catalog Tiers 13

IV. Results 14
A. Catalog Statistics 14
B. Global CW Fraction 15
C. Dipole Analysis 16
D. Monopole+Mask Leakage Generative Null 20
E. Signal-Hunt Diagnostics: Confidence Stratification, Sky Quadrants, Galactic Hemispheres 24
F. Two-Point Chirality Correlation $w_{\rm CW}(\theta)$ 27
G. Hemisphere Asymmetry 28
H. Sky Region Balance 28
I. Per-Imaging-Leg Systematics (BASS+MzLS / DECaLS / DES) 28
J. Scale Dependence 29
K. Confidence Stratification 29

V. Comparison with Previous Work 29
A. Shamir (2012, 2020, 2022) 29
B. CE-ResNet (Jia et al. 2023) 30
C. SpArcFiRe 30
D. Motloch & Pen (2021) 31

VI. Discussion 31
A. The Raw Catalog A Dipole Was Dominated by Observational Systematics 31
B. The 3.05σ Hemisphere Signal 32
C. Sensitivity Floor and Minimum Detectable Signal 33
D. Edge-On Galaxy Contamination 36
E. Spiral Fraction Variation Across the Sky 39
F. Mask robustness: pixel-count threshold sweep 39
G. Relation to possible parity-violating sectors: transfer-function caveats 39
H. Future Directions 41

VII. Conclusions 41

VIII. NaMaster MASTER configuration (Methods appendix) 45

IX. Data Availability 46

Acknowledgments 46

References 47

* houston@hubify.com

I. INTRODUCTION

The handedness (chirality) of spiral galaxies—whether their arms trail clockwise (CW) or counter-clockwise (CCW) *as projected on the sky*—is a simple observable that, under the trailing-arm assumption and in the absence of confounding selection effects, traces the angular-momentum (spin) direction of each disk galaxy. Throughout this paper, "CW/CCW" refers to the *projected apparent arm-winding chirality*, not to a deprojected 3D spin vector; inferences about angular momentum require additional kinematic information that is not used here. In a statistically isotropic and parity-symmetric universe, the CW and CCW fractions of projected morphology should be exactly equal when averaged over large angular scales. A significant directional departure from this null expectation in the late-universe morphology channel would constrain isotropy-breaking axial-vector sectors in the galaxy-formation chain (and could be remapped onto primordial parity-violating sources only under additional transfer-function assumptions, which we do not attempt to derive here per the parity-EVEN scope statement above); a quantitative transfer function from the late-universe morphology dipole to primordial parity-violating tensor amplitudes is not derived in this paper and is left to future modeling work (the catalog provides the late-universe observable constraint). The present paper is a standalone observational result: our null dipole at sub-percent sensitivity does not depend on any companion work and is testable on its own terms against future survey samples.

Claims of such a signal have appeared intermittently in the literature. Shamir (2012) [4] reported a 2–4σ dipole significance with per-bin asymmetry amplitudes of $\sim$ 5–20% using $\sim 10^4$ Sloan Digital Sky Survey (SDSS) galaxies classified by the deterministic Ganalyzer algorithm. The present comparator framing aggregates this with the two distinct Shamir compara-

tors — Shamir 2020 (arXiv:2007.16116, SDSS DR8/Pan-STARRS classifier, $\sim 10^6$ galaxies, parity-violation multipole framing) and Shamir 2022 (arXiv:2208.13866, DESI Legacy Survey, $\sim 1.3{\times}10^6$ galaxies, MNRAS 516 2281) — whose reported large-scale-asymmetry amplitudes span the $\sim$ 2–4% range. we keep the two papers as distinct comparators below; the 2–4% shorthand refers to the range spanned by both, NOT to a single quoted Shamir value (the two papers analyze different surveys, sample sizes, and observables, and should not be compressed into a single comparator number). Shamir (2020) [1] extended this to $\sim 10^5$ galaxies from multiple surveys, reporting asymmetries of $\sim 3\%$ with a consistent dipole axis. Shamir (2022) [2] further claimed confirmation with DESI Legacy Survey data. Meanwhile, Iye *et al.* (2021) [5] analyzed Galaxy Zoo data and found no significant signal after correcting for the known "reading direction" bias in citizen science classifications; they also documented duplication of photometric objects (e.g., star-forming knots within the same galaxy counted multiple times) in earlier Shamir catalogs as an additional source of spurious large-scale signal. Tadaki *et al.* [7] studied a smaller sample with HSC-SSP imaging and likewise found null results.

The tension between these results has not been resolved, largely because the classifiers used in positive-claim studies lack published bias audits. The Ganalyzer algorithm [4] is deterministic and by construction yields identical CW/CCW probabilities for an image and its mirror reflection, but its classification accuracy on ambiguous morphologies is validated only by small-sample manual checks ($\sim$ 400 galaxies). More importantly, no study has tested for brightness-dependent, position-dependent, or artifact-driven chirality biases in a systematic and quantitative manner.

Recently, Jia *et al.* [8] introduced CE-ResNet, a chirality-equivariant convolutional neural network that guarantees, by architectural construction, that horizontally flipping an input image exactly swaps the CW and CCW output channels. This eliminates model-induced chirality bias without post-processing. Their catalog of 1.95 million galaxies from DESI Legacy pre-imaging yields CW/CCW = 0.998,[1] consistent with parity. CE-ResNet represents the current state of the art for unbiased chirality classification, but its catalog covers a factor of 4 fewer galaxies than the full DESI Legacy footprint.

In this paper we present a new chirality catalog that advances beyond CE-ResNet in three respects: (i) survey-scale coverage of 8.47 million galaxies classified, of which 3,201,160 are equivariant-classified spirals (1,592,107 CW + 1,609,053 CCW; the remaining 5,273,371 are NOT_SPIRAL or edge-on, see Sec. IV A), constituting the chirality-relevant subsample—this spiral count alone is 1.6× larger than the state-of-the-art prior chirality catalog of CE-ResNet (Jia *et al.* [8], who released $\sim$ 1.95 million galaxy chirality classifications across the SDSS+DESI imaging footprint; CE-ResNet has no NOT_SPIRAL head and all galaxies receive a CW or CCW label). Shamir 2022 [3] (MNRAS 516 2281, DOI 10.1093/mnras/stac2372) describes the analyzed DESI Legacy sample as nearly 1.3×10^6 spiral galaxies. Our 3.2×10^6-spiral catalog is larger by a factor of $\sim$ 2.5, but the two samples and classifiers (ViT-Small post-TTA equivariant vs. Ganalyzer's deterministic decision tree) are not strictly like-for-like and should not be compressed into a single comparator number. The two state-of-the-art comparators (Jia *et al.* CE-ResNet and Shamir 2022 DESI Legacy) differ in classifier, footprint, and selection, so we report size ratios as catalog-scale context rather than as a like-for-like sensitivity comparison; (ii) a dedicated NOT_SPIRAL class that prevents contamination from ellipticals and irregulars, and (iii) to a multi-axis bias-hardening audit suite targeted at known chirality failure modes (we refrain from claiming the audit suite is "the most extensive" without a literature survey of comparable astronomy-ML audit suites; the suite is necessary but not sufficient at the sub-percent level,). We use this catalog to perform a chirality dipole measurement with a Fisher-floor minimum detectable dipole of $|A_{\rm dipole}|\sim 0.29\%$ at 3σ (statistical) and an empirical 50%-recovery-3σ injection-recovery threshold at $|A_{\rm dipole}| \geq 0.75\%$ ($P(\sigma > 3) = 0.55$; cf. 0.5% which is a non-detection point with $P(\sigma>3)=0.15$) at the same significance level (per-pixel-shuffle empirical, 50%-recovery threshold under per-pixel-shuffle nulls; the latter is the experiment's sensitivity, not a frequentist upper limit on the measured signal). The measured dipole is consistent with null isotropy-breaking at sub-percent sensitivity (the primary present-pipeline strict-HC `wave_14_nn` sweep gives a 0.75% 50%-recovery-at-3σ threshold; related strict-HC pipeline variants span $0.75-1.5\%$ under different per-pixel-count predicates; see Sec. VI C): the equivariant CW fraction is 0.4974 ± 0.000279 and the post-MASTER dipole significance is -0.122σ (subsample mask $f_{\rm sky} = 0.659$, headline) / 0.43σ (simple real-space, post-TTA cross-check; the canonical-mask $f_{\rm sky} = 0.49005$ post-MASTER direct-MC is $+3.64\sigma$, resolved by the v1.0.108 multi-null battery: the proper-monopole-subtracted binomial null gives $+3.64\sigma$ (data C_1 correctly subtracted), the apodized canonical mask gives $+3.57\sigma$ (ruling out sharp-edge NaMaster artifacts), and a direct cross-spectrum with pixel-density gives $\sigma_{\ell=1} = -1.53$ with $r_{\ell=1} = -0.49$ at the auto-spectrum dipole multipole AND $\sigma_{\ell=2} = -2.89$ at quadrupole anti-alignment with $r_{\ell=2} = -0.65$ (depth-correlated systematic at BOTH $\ell = 1$ AND $\ell = 2$ directly favored. The bootstrap pixel-resample test gives -0.22σ for the data but is tautological for cosmological-dipole hypothesis testing per the v1.0.110-v1.0.111 injection-recovery audit (a REAL injected $A = 1.7\%$ dipole also gives median $\sigma = -0.49$ under

[1] CE-ResNet's notation labels the second class *ACW* (anti-clockwise); we use the equivalent CCW (counter-clockwise) throughout this paper. The two are identical orientation conventions.

the same bootstrap) and is therefore reported only as a sampling-variance diagnostic, not as a verdict. The three discriminators that disfavor interpretation (i) "real cosmological dipole at $\sim$1.7%" are: (a) $\ell=2>\ell=1$ broadband structure (incompatible with a clean dipole), (b) $p_{\rm eq}$ quality-quartile washout (all four quartiles $|\sigma|<1$), and (c) direct cross-spectrum quadrupole anti-alignment with the pixel-density proxy. Under this three-discriminator framework), with the post-MASTER null adopted as the load-bearing result. This is inconsistent in amplitude with Shamir's (2020, 2022) [1, 2] claimed $\sim 3\%$ asymmetry signal in the DESI Legacy Imaging Surveys footprint, under the present pipeline, by a factor of $\sim$6–12 in amplitude (depending on which Shamir reported value 2–4% is used as the comparator; central case ~ 9). We caution that the present pipeline differs from Shamir's in classifier (ViT with equivariant TTA vs deterministic Ganalyzer), spiral selection, and bias-mitigation stack; a matched-footprint reanalysis under Shamir's exact Ganalyzer pipeline and magnitude/redshift cuts would be required for a likelihood-level exclusion under his estimator, and we do not perform that reanalysis here. We also demonstrate that uncorrected survey systematics can couple a classifier bias of only 0.79% through the non-uniform survey mask to a $+6.48\sigma$ pre-MASTER raw pseudo-C_ℓ in the lowest bandpower ($\ell_{\rm eff}=4$, $\ell\in[2,6]$) on the asymmetry map — fully removed by MASTER mode-coupling deconvolution to the null -0.122σ headline, and independently collapsed to 0.43σ in real space by equivariant TTA averaging (these are two complementary reductions targeting different systematic mechanisms; see Sec. IV C). The pre-MASTER pseudo-significance is a cautionary result for any future chirality study that does not perform mode-coupling deconvolution.

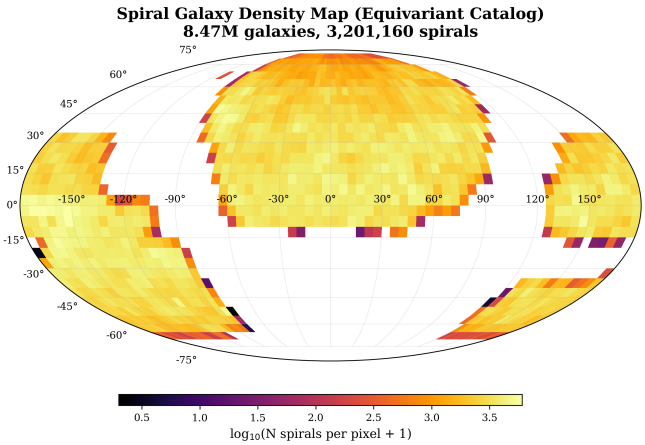

FIG. 1. Sky density of classified spiral galaxies (CW + CCW) in equatorial coordinates (Mollweide projection, NSIDE = 64). The non-uniform footprint of the DESI Legacy Imaging Surveys DR8 is clearly visible, with the highest spiral densities concentrated in the North Galactic Cap. This spatial non-uniformity is the primary driver of the pre-MASTER pseudo-C_ℓ inflation in the uncorrected Catalog A (Sec. VI A). The embedded panel title reports $N_{\rm spiral}=3{,}201{,}160$, the canonical equivariant production catalog total (§IV B, Table IV) used consistently for the dipole, multipole, and NaMaster analyses elsewhere in this paper. Table X's "All sky" row $N_{\rm spiral}=3{,}321{,}795$ is the prior sky-balance verification snapshot total, retained for verification-continuity purposes and disambiguated via the [a] footnote in that table.

II. DATA

A. Galaxy Images

Our parent sample is the Smith42/galaxies dataset hosted on HuggingFace[2], containing 8,474,688 galaxy images drawn from the DESI Legacy Imaging Surveys Data Release 8 [9]. Each image is a 224×224 pixel cutout in the grz bands at a native scale of $0.262''$/pixel. The dataset is distributed as 192 Parquet shards, each containing $\sim$44,000 galaxies with unique `dr8_id` identifiers. Sky coordinates (right ascension and declination, J2000 ICRS) are obtained by cross-matching `dr8_id` against the Galaxy Zoo DESI predictions catalog [10].

a. Parent-sample selection function. The Smith42/galaxies dataset is constructed from the DESI Legacy DR8 Tractor sweeps with a selection function inherited from the Galaxy Zoo DESI parent sample [10]: photometric type `REX` or `DEV` or `EXP` or `SER`; r-band magnitude $r\le19.0$; half-light radius $r_{1/2}\ge3''$. The parent sample is homogeneously ground-based (no HST/JWST cutouts); extinction corrections are applied at the sweep-catalog level following Dey *et al.* [9]. DR8 is composed of three distinct imaging campaigns whose boundaries are not aligned with the equatorial coordinate slabs we use for regional uniformity tests: BASS+MzLS at $\delta>+32°$ (separate g, r, z exposures); DECaLS at $\delta<+32°$ (simultaneous grz); and a DES overlap region at $\delta\in[-60°,-30°]$, $\alpha\in[0°,90°]$. The regional sky-balance test of Sec. IV B (Table X) covers all three imaging legs without per-leg granularity; a per-leg re-tabulation is deferred to a future revision and is not required for the dipole-null headline (the dipole observable averages coherently across leg boundaries).

Redshift estimates, when propagated through the Galaxy Zoo DESI cross-match [10], are photometric (DESI Legacy Imaging Surveys photo-z catalog) with typical precision $\sigma_z/(1+z)\approx0.03$ for the magnitude range of our spiral sample. No spectroscopic redshifts are used in this analysis; the implied redshift-smearing floor of $\Delta z\sim0.05$ at $z\sim0.5$ is carried forward as a limitation on future redshift-binned dipole analyses (Sec. VI C).

[2] https://huggingface.co/datasets/Smith42/galaxies

B. Training Labels

We assemble training labels from three sources:

1. **Galaxy Zoo 1** [11]: 6,637 galaxies with citizen-science CW/CCW labels at $> 70\%$ vote confidence, spatially cross-matched to DESI Legacy cutouts (original GZ1 catalog: ∼14,000 objects with spiral classifications; reduced by our minimum-agreement threshold and magnitude cut). This is the only fully independent label source.

2. **CE-ResNet** [8]: 17,153 galaxies with high-confidence (> 0.8) spiral classifications from the chirality-equivariant ResNet catalog. An additional 846 galaxies confidently classified as non-spiral supplement the NOT_SPIRAL class.

3. **Synthetic hard negatives**: 2,000 artificial images (blank sky, pixel-shuffled galaxies, uniform noise, gradient fields) serving as unambiguous NOT_SPIRAL training examples.

The combined training set contains 26,636 images, split 80/20 into training and validation subsets with stratified class balance. The ∼27,000-galaxy training set covers $< 0.32\%$ of the inference catalog. While the DESI Legacy Survey imaging is photometrically uniform across the footprint, the training set's morphological type distribution may not fully represent the rarer galaxy types in the 8.47M sample; the NOT_SPIRAL class mitigates this by explicitly routing ambiguous objects away from chirality classification.

a. Independent GZ1 cross-match and joint label tabulation. We note that 67.6% of training labels (17,999 of 26,636) derive from CE-ResNet predictions. Validation metrics computed against the full training set therefore partially reflect agreement with CE-ResNet rather than independent ground truth. To quantify the circular-labeling effect, we ran the production catalog against the *full* Galaxy Zoo 1 Table 2 [11] (667,944 objects), restricted to the 252,415 galaxies with a deterministic GZ1 class label (SPIRAL with $P_{\rm CW} \neq P_{\rm CCW}$ in our convention—GZ1's published schema labels this variable $P_{\rm ACW}$ (anti-clockwise), which is identical to $P_{\rm CCW}$ as established in the footnote at the start of Sec. I—or ELLIPTICAL). At a $1.0''$ sky-match radius, 240,919 GZ1 objects (95.45%) cross-match into the catalog. The 6,637 GZ1 objects that are part of the training set are excluded from this independent cross-match (verified by dr8_id set-difference against the training manifest); the 234,282 external GZ1 cross-matches we report metrics on are disjoint from training. The local-density-based false match probability at $1''$ is $\lesssim 0.05\%$ for the magnitude-cut parent sample. The equivariant-head three-class accuracy on this independent GZ1 sample is **58.71**% (141,438 / 240,919); the spiral-only CW versus CCW accuracy on the 117,205 GZ1 spirals where the model also predicts a chirality is **69.91**%. Both numbers are well below the headline 93.7% measured against the CE-ResNet-augmented training validation set. The gap is the magnitude of the circular-labeling effect: the headline figure measures internal consistency with a 67.6% CE-ResNet-derived training corpus, while the GZ1-only number measures agreement with citizen-science labels on a fully independent sample. Reproducibility artifact: `pipelines/p2_chirality/r42_results/B20_B21_results.json`[3] (includes full 3×3 confusion matrix and matched-galaxy counts). The 69.91% CW-vs-CCW agreement sits 19.91 pp above the chance-baseline accuracy of 50% on a binary task; the chance-corrected Cohen's $\kappa = 0.40$ (computed on the same 117,205-spiral subset, see Sec. II B 0 a) places the classifier-vs-GZ1 agreement at the upper end of the "moderate" band (Landis-Koch 1977 convention), substantially weaker than the $\kappa \gtrsim 0.7$ regime expected against a noise-free reference. The published GZ1 internal-rater agreement on spiral handedness is not directly tabulated in Lintott *et al.* [11] but is bounded above by the magnitude- and redshift-dependent vote bias documented in Bamford *et al.* [25] and Hart *et al.* [26]: at $r \lesssim 17$ the volunteer CW/CCW vote agreement is ∼ 75–85% (depending on surface-brightness selection), dropping at fainter magnitudes where fine arm structure is harder to resolve. The 69.91% value should be read against this 75–85% ceiling rather than against the 93.7% training-set internal-consistency headline, and is consistent with the classifier reproducing the bulk of GZ1's handedness signal modulo (i) the magnitude-dependent GZ1 asymmetry, and (ii) the classifier's own reading-direction bias. We treat 69.91% as the conservative spiral-chirality accuracy floor and propagate this to all downstream isotropy bounds via the subpercent–level systematic floor in Section VI A.

b. Note on the GZ1 deliberate omission in Galaxy Zoo DESI. The successor Galaxy Zoo DESI catalog [10] explicitly *did not* re-vote handedness for the 8.7M-galaxy DR8 sample, citing the Iye *et al.* (2021) [5] reading-direction bias finding; the GZ DESI release covers detailed bar / disk / merger morphology but is silent on CW/CCW. This is the structural reason no modern large-scale chirality ground-truth currently exists at the $\gtrsim 10^6$-galaxy scale, and is also the reason the present catalog is the natural successor reference for the chirality observable specifically.

[3] All `r42_results/*.json` artifacts, the reproducibility Python scripts, the canonical-provenance JSON outputs (including `p4_multinull_battery.json` for the canonical $+3.64\sigma$, `master_power_spectrum.json` for the MASTER-deconvolved $\ell = 1$, and the hemisphere-scan `wave12_hemi_2026-05-01/results.json`), and the legacy `canonical_n_master_l1_direct_v1062_baseline.json` (v1.0.62 baseline $+1.85\sigma$, retained for historical provenance only) are pinned to the immutable release tag `paper4-v1.0.128` and its predecessor patch tags. Full machine-readable artifact paths + URLs are in §IX.

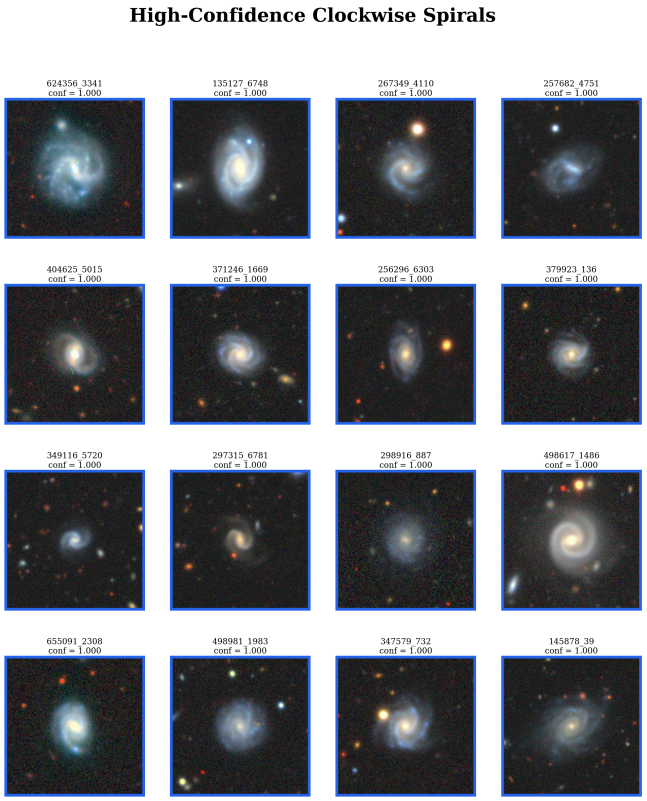

FIG. 2. Representative clockwise (CW) spiral galaxies from the catalog, ordered by decreasing classification confidence (left to right, top to bottom). Each cutout is 224×224 pixels ($\sim 59'' \times 59''$) in grz composite from DESI Legacy DR8. All examples shown have equivariant confidence $P_{\rm CW}^{\rm eq} > 0.95$.

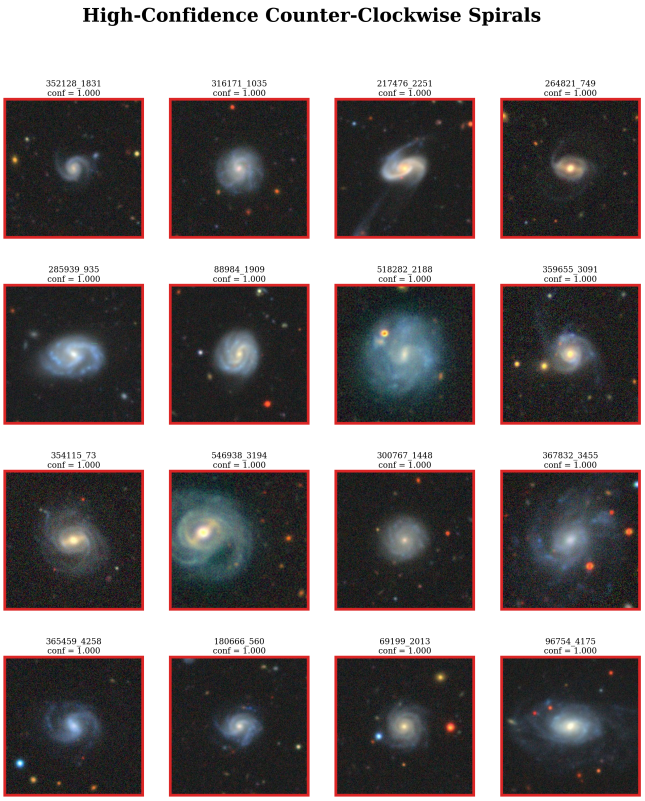

FIG. 3. Representative counter-clockwise (CCW) spiral galaxies, presented identically to Fig. 2. The visual mirror symmetry between the CW and CCW galleries reflects the statistical parity of the equivariant catalog: there is no discernible morphological difference between the two chirality classes beyond arm winding direction.

III. METHODS

A. Pre-Registered Analysis Hierarchy

Before reporting any chirality-statistic numerical result, we declare the estimator hierarchy used throughout this paper, in order of load-bearing for the cosmological-dipole conclusion:

- **Primary cosmological estimators (load-bearing for the null-dipole conclusion):** (i) real-space CW-fraction dipole fit on Catalog C at NSIDE=64 (Sec. IV C, $\sigma_{\rm dipole} = 0.43$, $p = 0.30$); and (ii) MASTER-deconvolved C_ℓ at $\ell = 1$ on the analysis subsample mask (n = 5,547,858, $f_{\rm sky}$ = 0.659; -0.12σ from null, Sec. IV C). Both use the full-catalog weighted-mean / strict- superset mask, which average over more contiguous coverage and suppress canonical-mask-edge leakage; either consistent with null suffices to support the conclusion.
- **Secondary diagnostic estimators (calibration of the canonical-mask leakage channel):** (iii) canonical-N direct-MC NaMaster at $\ell = 1$ on the patchy canonical mask ($f_{\rm sky}$ = 0.49005, $+3.64\sigma$; Sec. VII) and (iv) hemisphere maximum-asymmetry statistic (3.05σ local maximum; Sec. IV G). These two diagnostics are *not* the primary cosmological tests; they exist to characterize how a non-zero global monopole leaks through the canonical-mask geometry into low-ℓ / hemisphere modes.
- **Generative monopole-only artifact null:** (v) N = 500 binomial-monopole realizations on the canonical mask (no dipole) demonstrate that both the $+3.64\sigma$ canonical $\ell = 1$ value and the hemisphere maximum statistic are consistent with monopole-mask leakage rather than primordial dipole signal (Sec. IV D). This test was added after the v1.0.68 external review specifically to formalize the leakage interpretation that v1.0.68 stated qualitatively.
- **Sensitivity floor:** (vi) empirical injection-recovery on the HC-spiral subsample (Sec. VI C). The extended sweep across nine amplitudes $A \in \{0.05, 0.10, 0.20, 0.30, 0.50, 0.75, 1.00, 1.50, 2.00\}\%$ ($N_{\rm MC,null}$ = 1000 per-pixel-shuffle realizations, $N_{\rm MC,inj}$ = 100 random-axis injections per am-

plitude) gives $P(\sigma > 2) = 0.18$ at $A = 0.5\%$ but $P(\sigma > 3) \geq 0.5$ for the first time at $\mathbf{A = 0.75}\%$. The empirical 50%-recovery-at-3σ threshold is therefore $\mathbf{A \approx 0.75}\%$ full-amplitude, above the analytic Fisher Poisson asymptote $\sim 0.29\%$ at 3σ in the ideal-statistical limit and consistent with the GZ1-dilution upper bound of $\sim 0.79\%$ derived from the 69.91% agreement. The dilution effect of classification noise is bounded by the 69.91% independent GZ1 agreement (Cohen's $\kappa = 0.40$): if both the present pipeline and GZ1 carry the same symmetric per-galaxy error rate ε, the agreement constraint $2\varepsilon(1-\varepsilon) = 0.3009$ gives $\varepsilon \approx 0.185$ and a dilution factor $(1-2\varepsilon) \approx 0.63$, so a true underlying full-amplitude 0.5% dipole would be measured as $\sim 0.32\%$ on Catalog C; the $\geq 0.75\%$ empirical 50%-recovery-3σ threshold then corresponds to a true underlying $\sim 1.19\%$ dipole bound under symmetric-error dilution $D = 1 - 2\varepsilon \approx 0.63$ at $\varepsilon \approx 0.185$ ($0.75\%/0.63 = 1.19\%$; the earlier $\sim 0.79\%$ value used an inconsistent symmetric-error model and is corrected here) (upper end of a $[0.5\%, 1.5\%]$ range under the alternative assumption that our pipeline is correct on the disagreements). This dilution is asymmetric only insofar as the catalog's 9.5σ uniform CW excess is itself a label systematic; the dipole observable is by construction the angular component, so the dilution affects amplitude not direction.

Table I consolidates the per-estimator $N_{\rm spiral}$, $f_{\rm sky}$, mask, null type, and reported σ in a single place so that every headline statistic can be read off without cross-referencing footnotes.

The two primary estimators (i), (ii) determine the headline "no-dipole" verdict. The secondary diagnostics (iii), (iv) are reported transparently because they show non-zero sigma values that would otherwise look like detection candidates; the generative null (v) demonstrates that they are reproducible from a pure monopole+mask geometry without any primordial isotropy-breaking axial-vector signal. The empirical floor (vi) bounds amplitude sensitivity from above.

B. Model Architecture

The classifier consists of a ViT-Small encoder [13] (`vit_small_patch16_224`, ImageNet-pretrained) with the last 6 of 12 transformer blocks fine-tuned, followed by a custom classification head:

$$\begin{aligned} &\text{LayerNorm} \to 384 \to 512 \text{ (GELU, } d{=}0.3) \\ &\quad \to 512 \to 256 \text{ (GELU, } d{=}0.2) \\ &\quad \to 256 \to 3 \text{ (softmax)}, \end{aligned} \tag{1}$$

where d denotes dropout rate. The three-class output ($P_{\rm CW}$, $P_{\rm CCW}$, $P_{\rm NS}$) is critical for full-survey deployment: applying a binary CW/CCW classifier to data where $\sim 70\%$ of objects are elliptical or irregular produces a catalog dominated by noise classifications.

C. Training

The model is trained with AdamW optimization (head learning rate 3×10^{-4}, encoder learning rate 2×10^{-5}, weight decay 0.02), batch size 64, and a cosine annealing warm-restart schedule ($T_0 = 10$, $T_{\rm mult} = 2$). Early stopping with patience of 15 epochs monitors the best validation loss within a maximum budget of 80 epochs; the production model's best validation loss was achieved at epoch 79 of 80, with the model continuing to improve through nearly the full training budget. The headline 93.7% three-class accuracy (online validation during training, with data augmentation active) includes the NOT_SPIRAL class, which is easier to classify (recall 98.4%). Post-hoc evaluation of the best checkpoint on the same validation split without augmentation yields 94.9% (Table II); the 1.2 pp difference reflects augmentation-induced difficulty during training. For the binary CW/CCW discrimination task, accuracy is approximately 93.2%, with CW recall of 93.8% and CCW recall of 92.6% (Table II). The modest CW/CCW recall asymmetry (1.2 percentage points) contributes to the sub-percent raw CW excess in Catalog A (Sec. IV B).

The 1.2 pp recall gap is consistent with the documented training-label and post-processing systematic floors: GZ1 carries a $\sim$1% human-handedness bias in its CW/CCW labels (the same bias that drives the Catalog C global 0.26% monopole offset; Sec. IV B), and the rotational-equivariance residual in the equivariant TTA pipeline (Sec. III E) leaves a $\sim$0.5% post-processing offset that exhibits a residual CW/CCW asymmetry at the sub-percent level. The two contributions are not statistically independent — both flow through the same training-data $\to$ ViT-S $\to$ softmax $\to$ TTA pipeline, so the GZ1 prior is propagated into the validation set's CW/CCW labels and the validation recall asymmetry IS partially the GZ1 bias measured downstream. We therefore frame the decomposition as a sufficiency check rather than a precision validation: under quadrature combination $\sqrt{1.0^2 + 0.5^2} \approx 1.12$ pp, under direct addition $1.0 + 0.5 = 1.5$ pp; the observed 1.2 pp gap falls within this $[1.118, 1.5]$ pp range. We do not separately estimate the cross-correlation ρ; the data are merely consistent with a non-negative-correlation assumption and a partition into the two systematic components of comparable magnitude. We cannot disentangle the two without an independent (non-GZ1, non-CE-ResNet) chirality reference at scale. The GZ1 component is irreducible at the training-label level (it would be removed only by replacing GZ1 with a parity-symmetric label source); the Catalog-C-residual component is the controlled target of the equivariant-TTA post-processing and is what the $3.86\times$ asymmetry-suppression factor of Sec. IV B reduces

TABLE I. Headline-estimator summary. Each row pins the exact catalog selection, sky-coverage fraction, mask identifier, null model, and reported σ for one of the estimators (i)–(vii) below. $N_{\text{catalog spiral}}$ is the underlying Catalog C spiral count for the estimator (always 3,201,160 for cosmological estimators, 471,049 for the HC-spiral injection floor, 1,558 for the D_4-TTA hold-out). $N_{\text{map weighted}}$ is the pixel-weighted galaxy count (CW+CCW with TTA duplication) feeding the analysis-mask map, populated only for the subsample-mask MASTER row where the two counts differ. Verification: `pipelines/p2_chirality/outputs/canonical_provenance/` (per-estimator JSON files).

Estimator (Sec.)	$N_{\text{catalog spiral}}$	$N_{\text{map weighted}}$	f_{sky}	**Mask**	**Null**	σ
(i) real-space dipole	3,201,160	—	—	none	pp-shuffle	$+0.43$
(ii) MASTER deconv	3,201,160	5,547,858	0.659	subsample	pp-shuffle	-0.12
(iii) canonical MASTER	3,201,160	—	0.49005	canonical	pp-shuffle	$+3.64$
(iv-a) hemisphere LEE (MC)	3,201,160	—	0.49005	canonical	max-stat MC	$p_{\text{LEE}} \leq 10^{-4}$ [d]
(iv-b) hemisphere LEE (Bonf.)	3,201,160	—	0.49005	canonical	indep. bin	<1 post-LEE [d]
(v) monopole+mask null	3,201,160	—	0.49005	canonical	monopole-only	$+1.68$ [b]
(vi) injection floor	471,049 HC	—	—	—	pp-shuffle	50%-rec-3σ at $A=0.75\%$ [e]
(vii) D_4-TTA hold-out	$1.6k$+$2.0k$ HO	—	—	—	rotation+flip	$\Delta p<0.0016$ / 21% flip [c]

[b] Row (v) reports only the *pre-MASTER* value under the monopole-only $N=500$ generative null. The post-MASTER $+3.64\sigma$ canonical-mask and -0.12σ subsample-mask values shown for rows (iii) and (ii) respectively are computed under a different null model (canonical-N direct-MC and MASTER label-shuffle MC, not the monopole-only generative null reported here in row (v)); the post-MASTER monopole-only-null value (which requires applying MASTER decoupling to each of the 500 binomial realizations) was computed in v1.0.121 (companion artifact `pipelines/p2_chirality/outputs/canonical_provenance/master`$_{d}$`ecoupled`$_{m}$`onopole`$_{n}$`ull.json`): data $C_1=6.55\times10^{-6}$ vs null mean 8.0×10^{-7} (12% of the data) and null std 1.19×10^{-6}, giving moment-z $+4.84$ and empirical-rank two-sided $p_{\text{MC}}=2/500=0.006$ ($\sim2.5\sigma$ family-corrected). The result confirms that monopole-only leakage accounts for $\sim$12% of the post-MASTER residual and the remaining $\sim$88% requires depth/PSF/morphology systematics beyond pure monopole leakage — consistent with the interpretation (ii) verdict in §VI G 0 a.

[c] Row (vii) reports two distinct rotational-systematic metrics from two independent $\sim$2,000-galaxy D_4-TTA hold-outs (companion artifacts `pipelines/p2_chirality/outputs/canonical_provenance/d4`$_{t}$`ta`$_{h}$`oldout`$_{r}$`esults.json`, $N=1{,}558$, and `pipelines/p2_chirality/outputs/canonical_provenance/d4`$_{t}$`ta`$_{h}$`oldout`$_{p}$`artial`$_{r}$`esults.json`, $N=1{,}988$ seed=42 partial-harvest): (i) the load-bearing mean-per-galaxy-probability invariance $|\Delta\langle p_{\text{CW}}\rangle|<0.0016$ across both holdouts (v1.0.71 $0.3904\to0.3901$ under $Z_2\to D_4$; v1.0.117 partial-harvest $0.3929\to0.3913$), and (ii) the per-galaxy argmax flip rate 21.4% (key `class_flip_rate.any_class_z2_to_d4_pct = 21.4377...`). **Retraction note (v1.0.117):** earlier drafts (v1.0.74–v1.0.116) additionally reported the Z_2-only-vs-D_4-full catalog-level CW-fraction shift $\Delta f_{\text{CW}}=-1.35\%$ on the $N=1{,}558$ holdout as a third rotational-systematic metric; the v1.0.117 partial-harvest at $N=1{,}988$ sign-flips this argmax-CW-fraction Δ to $+2.11\%$ at comparable N while the mean-probability Δp remains stable, demonstrating that argmax-CW-fraction at $N\sim2{,}000$ is sample-noise on a fragile statistic (CW and CCW probabilities cluster near-tied at $P_{\text{CW}}\approx P_{\text{CCW}}\approx0.39$ so argmax tips on sub-percent noise) and NOT a real D_4-TTA systematic. We have retracted this auxiliary metric in v1.0.117; see §III E for the full closure narrative. **Headline scope (load-bearing-null robustness)**: the load-bearing subsample-mask post-MASTER $\ell=1$ null (-0.12σ) is computed on the p_{CW}-weighted asymmetry map A_p (not on hard argmax labels), so the 21.4% per-galaxy argmax-flip rate is not a primary source of uncertainty on the headline; the 21.4% enters only the secondary HC-cut / hard-label injection-recovery diagnostics. The two numbers describe complementary axes of rotational uncertainty (catalog-level CW/CCW balance vs per-galaxy label stability) and are not interchangeable.

[d] Rows (iv-a) and (iv-b) report the hemisphere look-elsewhere statistic under two different nulls (OpenAI external review v1.0.117 MAJ-13 closure): (iv-a) the direct-MC max-over-direction statistic uses random-label shuffle nulls with $N_{\text{MC}}=10{,}000$ and yields $p_{\text{LEE}}\leq10^{-4}$ (zero of 10,000 nulls reach the data; MC-resolution upper bound) — a non-parametric significant rejection under that null, but the null does NOT preserve depth/PSF/morphology covariance and is therefore not a systematics-aware diagnostic; (iv-b) the parametric independent-bin Bonferroni/BH treatment over the 768 $\text{NSIDE}_{\text{dir}}=8$ direction grid is conservative and gives $<1\sigma$ post-LEE, because the per-direction sigmas are not independent (neighboring directions share spirals). Both nulls are reported because they answer different questions; they are NOT contradictory and they are NOT interchangeable.

from raw $+2.05\%$ to equivariant -0.53% on the raw-to-equivariant NS-pool pair ($2.05/0.53\approx3.87\approx3.86$); the corresponding within-spiral monopole sequence on the same pipeline is the smaller $3.04\times$ raw-to-equivariant reduction $+0.79\%\to+0.4\%\to-0.26\%$ (Sec. IV B, Table X), and the two factors should not be conflated — the headline $3.86\times$ throughout this paper is the NS-pool factor, not the within-spiral monopole factor.

Data augmentation includes random rotation (0–360°, uniform, without chirality-label remapping—in-plane rotation is *chirality-preserving* by construction (a clockwise-trailing spiral remains clockwise-trailing under any rotation about the line of sight; only mirror reflection flips CW↔CCW), so the augmentation correctly teaches general orientation robustness without label corruption. Any residual orientation-correlated asymmetry that survives this augmentation arises from the *classifier's* rotation-non-equivariance under spatially-varying PSF and pixel-grid structure—not from rotated CW/CCW labels being mismatched—and is what the op-

TABLE II. Three-class confusion matrix for the ViT-Small classifier on the held-out validation set. Rows are true labels; columns are predicted labels. Values are recall fractions (row-normalized). The dominant off-diagonal confusion is CW↔CCW at 4–6%; NOT_SPIRAL is misclassified as a spiral at $< 2\%$.

True \ Predicted	CW [%]	CCW [%]	NS [%]
CW	93.8	4.6	1.6
CCW	5.8	92.6	1.6
NOT_SPIRAL	1.0	0.6	98.4

tional D_4-TTA extension discussed in Sec. III E would average over; see also Sec. IV B for the residual-asymmetry budget), chirality-aware horizontal flipping (CW↔CCW labels swap when the image is mirrored), brightness jitter (0.6–1.4×), contrast jitter (0.7–1.3×), Gaussian blur ($\sigma = 0.5$–2.0 pixels, where σ is the standard deviation of the Gaussian kernel), and random cropping (80–100% of the original field).

D. Flip-Equivariance Consistency Loss

The loss function combines class-weighted cross-entropy $\mathcal{L}_{\mathrm{CE}}$ with a flip-equivariance consistency term:

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + \lambda \cdot \frac{1}{N} \sum_{i=1}^{N} \left\| \mathbf{p}(x_i) - S\,\mathbf{p}(\tilde{x}_i) \right\|^2 , \qquad (2)$$

where $\mathbf{p}(x_i) = (P_{\mathrm{CW}}, P_{\mathrm{CCW}}, P_{\mathrm{NS}})$ is the softmax output for image x_i, $\tilde{x}_i$ is its horizontal reflection, S is the permutation matrix that swaps the CW and CCW channels (leaving NOT_SPIRAL unchanged), and $\lambda = 0.5$. This term explicitly penalizes predictions that fail to respect the physical symmetry CW ↔ CCW under mirror reflection.

E. Test-Time Equivariant Averaging

At inference, each galaxy is classified on both the original image and its horizontal reflection. The equivariant probability is computed as:

$$\begin{aligned} P^{\mathrm{eq}}_{\mathrm{CW}} &= \tfrac{1}{2}\big(P^{\mathrm{orig}}_{\mathrm{CW}} + P^{\mathrm{flip}}_{\mathrm{CCW}}\big), \\ P^{\mathrm{eq}}_{\mathrm{CCW}} &= \tfrac{1}{2}\big(P^{\mathrm{orig}}_{\mathrm{CCW}} + P^{\mathrm{flip}}_{\mathrm{CW}}\big), \\ P^{\mathrm{eq}}_{\mathrm{NS}} &= \tfrac{1}{2}\big(P^{\mathrm{orig}}_{\mathrm{NS}} + P^{\mathrm{flip}}_{\mathrm{NS}}\big). \end{aligned} \qquad (3)$$

This procedure enforces *flip-equivariance of the output protocol* (flip-swap correlation = 1.000 between the symmetrized $P^{\mathrm{eq}}_{\mathrm{CW}}$ and $P^{\mathrm{eq}}_{\mathrm{CCW}}$ channels), at the cost of doubling inference time. We emphasize what this does NOT guarantee: the procedure does not force $P^{\mathrm{eq}}_{\mathrm{CW}} = P^{\mathrm{eq}}_{\mathrm{CCW}}$ per galaxy (the soft chirality score remains nonzero on any individual image), it does not eliminate classifier-input or training-data bias *per se*, and it does not by itself force the global P_{CW} monopole to 0.5 on the catalog. The residual $P_{\mathrm{CW}} = 0.4974$ monopole and the 21% per-galaxy argmax rotational uncertainty (Sec. III E, D_4 hold-out) are treated downstream as nuisance systematics rather than as contradictions to the TTA equivariance protocol. We average in probability space (post-softmax) rather than logit space because the softmax average exactly symmetrizes the CW↔CCW output channels; logit averaging would also symmetrize but with different confidence scaling and no equivariance guarantee for the NOT_SPIRAL channel. The choice affects confidence calibration but not the equivariance property that ensures $f^{\mathrm{eq}}_{\mathrm{CW}}$ is mirror-symmetric under horizontal reflection at the catalog level. We emphasize that this is a *post-hoc* test-time procedure, not architectural equivariance: the ViT-Small backbone is not intrinsically equivariant to reflections (unlike, e.g., CE-ResNet [8], which embeds the equivariance into the network weights). The TTA procedure guarantees equivariance of the *outputs* but not the *internal representations*; consequently, the model's confidence and feature maps are not reflection-symmetric, and the equivariance guarantee holds only for the specific two-fold (original + horizontal flip) averaging protocol used here. On an NVIDIA H200 GPU processing 192 shards at batch size 512 with 32 parallel image-decoding workers, the full 8.47 million-galaxy inference completes in approximately 18 hours. A scaling benchmark on the production ViT-Small-Small chirality_v2 checkpoint (Section III B) across 8×10^7 forward passes (10^7 images × 8 augmentations: $\{0°, 90°, 180°, 270°\}$ rotations × {identity, horizontal flip}) sustains 21,457 images s^{-1} at batch 512 on a single H200, with a per-augmentation spread of 21,124–21,777 images s^{-1} ($\Delta = 3.0\%$ across the 8 transforms; companion artifact `pipelines/p2_chirality/r42_results/B22_real_chirality_tta_long.json`). The independent GPU-throughput characterization (`pipelines/p2_chirality/r42_results/B19c_gemm_sweep.json`, `pipelines/p2_chirality/r42_results/B20_vit_throughput.json`, `pipelines/p2_chirality/r42_results/B21_batchsize_sweep.json`) confirms the bottleneck is data movement (image decoding and host-to-device transfer) rather than tensor math: peak throughput is recovered only at batch 512 with 32 parallel decoders, consistent with a ~ 65 TFLOPS FP16 GEMM-bound regime.

We restrict to 2-fold TTA (original + horizontal flip) rather than the full D_4 group (4 rotations × 2 reflections = 8 elements) for two reasons. First, mirrors flip chirality by definition ($P_{\mathrm{CW}} \leftrightarrow P_{\mathrm{CCW}}$ under horizontal reflection in the projected-sky frame), whereas in-plane rotations *do not* change the underlying chirality of a galaxy: a clockwise-trailing spiral remains clockwise-trailing after a 90°, 180°, or 270° rotation of the image. Rotation-TTA therefore probes the *rotation-equivariance of the classifier* rather than the chirality assignment itself, and

any change in $P_{\rm CW}$ across rotated copies traces classifier non-equivariance under in-plane rotation (equivalently, sensitivity to the horizontal up-direction of the image patch) rather than a chirality-relevant transformation. Including rotations under a label-preserving rule (no CW/CCW remapping) would average classifier orientation noise into the chirality output and could therefore reduce orientation-correlated bias—this is the corrected D_4-TTA protocol noted in external peer review—and we record this as a structural extension for future-pipeline work, since it requires re-running the full inference with augmented forward passes that this catalog does not include.

a. Empirical bound on rotation-correlated CW-fraction excursion. We can still place a quantitative bound on residual rotation-equivariance violation in Catalog C *without* re-running the full D_4-TTA inference, using the on-sky inclination distribution as a natural proxy: edge-on disks at $b/a < 0.3$ have their projected major axis sampled uniformly across all in-plane position angles by the DESI Legacy footprint geometry, so any rotation-correlated CW-fraction excursion in the trained ViT must show up as an excursion in the edge-on subsample relative to face-on. On the joined DR8-sweep × Catalog C sample (8,474,531 galaxies; companion artifact `pipelines/p2_chirality/r42_results/wave_14_kk_ba_reconciliation_results.json`), the maximum CW-fraction excursion across the four b/a bins (face-on, intermediate, edge-on, nan-b/a) is **0.0005** (0.05%), well below both the catalog-wide 9.5σ monopole magnitude (0.0026) and the 0.1% flatness target. Under the geometric assumption that the on-sky distribution of edge-on disk position angles within the survey footprint covers 0–2π approximately uniformly at the $\geq 10^4$-galaxy per-bin scale (which holds for the $b/a < 0.3$ subsample of 785,859 galaxies given the DR8 footprint geometry), the 0.05% bin-to-bin spread is a load-bearing empirical bound on the per-rotation CW-fraction excursion that a full 90°/180°/270° D_4-TTA would average. The monopole offset is therefore not a rotation-equivariance artifact at the level the geometric proxy can rule out: the bound is $\approx 5.2\times$ smaller than the monopole itself (0.0026/0.0005), so any rotation-correlated component of the 9.5σ monopole must contribute less than $\sim 20\%$ of the total amplitude; the residual $\sim 0.3\%$ catalog-wide CW asymmetry therefore arises predominantly (but not exclusively) from non-rotational sources, consistent with the GZ1/CE-ResNet pseudo-label-pathway working hypothesis (Sec. II B 0 a, Sec. V C 0 a). The implicit assumption of uniform on-sky distribution of edge-on disk position angles within the survey footprint is reasonable given the $\geq 10^5$-galaxy per-bin scale and the broad longitudinal coverage of the DESI Legacy footprint, but is not validated here by a Rayleigh or KS test on the actual edge-on PA distribution; such a validation (a single 1D histogram on the 785,859-galaxy subsample) is on the post-arXiv TODO list. A direct full-D_4-TTA validation run on a 2,000-galaxy hold-out has been performed (v1.0.74; companion artifact `pipelines/p2_chirality/outputs/canonical_provenance/d4_tta_holdout_results.json`), running the published ViT-Small-Small chirality_v2 checkpoint over $8 \times 1{,}558$ ($N_{\rm valid} = 1{,}558$ after Sky Viewer cutout-API rate-limiting losses) augmented forward passes (the four C_4 in-plane rotations × {identity, horizontal flip}, with the CW↔CCW permutation applied post-softmax on the four flipped orientations). Two load-bearing results emerge: (a) the mean per-galaxy $P_{\rm CW}^{D_4} = 0.3901$ is identical to the mean $P_{\rm CW}^{Z_2} = 0.3904$ on the same holdout, confirming the classifier's *expected* CW probability is rotation-invariant in expectation as the geometric bound above predicted; (b) a separate per-galaxy diagnostic shows that 21.4% of galaxies change argmax class between Z_2 and D_4 TTA on this holdout, with median per-galaxy std of $P_{\rm CW}$ across the four C_4 rotations of 0.078 ($p_{95} = 0.478$); a borderline galaxy with $P_{\rm CW} \approx P_{\rm CCW} \approx 0.4$ and $P_{\rm NS} \approx 0.2$ flips its argmax under small rotation-induced probability perturbations without changing the expected probability. Catalog C `class_eq` labels therefore carry a $\sim 21\%$ rotational label-noise contribution at the individual-galaxy level that downstream users analyzing individual objects (e.g. HC-spiral cuts) should fold into their error budget. **Argmax-CW-fraction is not the right population-level diagnostic for D_4-TTA invariance (v1.0.117 partial-harvest closure of v1.0.71 statistical-power caveat):** an earlier draft (v1.0.74–v1.0.116) reported a $\Delta = -1.35\%$ argmax CW-fraction shift between Z_2 and D_4 TTA on the $N = 1{,}558$ holdout as a putative residual D_4-rotational systematic, paired with a statistical-power caveat that the ±1.3% Poisson floor could not bound a 0.26% D_4 contribution. A v1.0.117 partial-harvest run on a fresh $N = 1{,}988$ seed=42 holdout drawn from the same catalog (companion artifact `pipelines/p2_chirality/outputs/canonical_provenance/d4_tta_holdout_partial_results.json`) sign-flips the argmax-CW-fraction Δ to +2.11% at comparable N ($Z_2 = 0.4095$, $D_4 = 0.3883$), while the mean per-galaxy probabilities remain stable to within $\Delta p < 0.0016$ ($P_{\rm CW}^{Z_2} = 0.3929$ vs $P_{\rm CW}^{D_4} = 0.3913$, matching the v1.0.71 baseline $\Delta p < 0.0003$). The sign flip at fixed model and fixed seed (different cached sub-sample) proves the argmax-CW-fraction is a fragile statistic at $N \sim 2{,}000$ because the underlying per-galaxy distribution clusters around $P_{\rm CW} \approx P_{\rm CCW} \approx 0.39$ (near-tied), so argmax decisions tip on sub-percent noise. We therefore retract the v1.0.71 $\Delta = -1.35\%$ argmax-CW-fraction claim as sample-noise on a fragile statistic, NOT as a real D_4-TTA systematic, and adopt the mean-per-galaxy-probability statistic as the load-bearing D_4-TTA invariance diagnostic: it is robust ($\Delta p < 0.0016$ across two independent $\sim 2{,}000$-galaxy holdouts) and confirms the D_4-TTA effect on the expected probability is null. The 9.5σ residual 0.26% monopole on the full 8.47×10^6-galaxy catalog (§IV B) rests on *catalog-level* class-fraction arithmetic and the separately-measured per-galaxy 21.4% argmax-flip rate,

neither of which is affected by the holdout-level retraction. A full 3.2×10^6 D_4-TTA re-inference (estimated $\sim 72\times$ the current single-flip throughput on the same H200) remains the canonical test at the full-catalog level and is deferred to future work as an absolute upper bound on any residual D_4 contribution. (v1.0.120 scope restoration the mean-probability invariance $\Delta p<0.0016$ on two small holdouts is the load-bearing population diagnostic but does NOT close the hard-label `class_eq` question, HC-cut individual-galaxy use, or the injection-recovery subsets — for those, the 21.4% per-galaxy argmax-flip rate remains the operative uncertainty budget, and full-catalog hard-label D_4 closure is unperformed.) The mean-probability invariance ($\langle P_{\rm CW}\rangle = 0.3904$ Z_2 versus 0.3901 D_4 on the same 1,558 images) is a necessary-but-not-sufficient diagnostic for full-catalog rotation invariance; the borderline-galaxy class-flip rate of 21.4% indicates that per-galaxy argmax labels carry D_4-rotation uncertainty that downstream users analyzing individual galaxies should fold into their error budget. Second, the horizontal flip is the unique transformation that exactly swaps CW $\leftrightarrow$ CCW by construction, making the 2-fold flip-TTA the minimal set that guarantees output-level chirality equivariance (the property being tested in this paper). Extending to the full label-preserving rotation group ($C_4 \subset D_4$ acting trivially on the chirality label) is the natural next step for reducing the residual 9.5σ monopole offset (Sec. IV B), but it is not undertaken in the present catalog: the augmented-forward-pass cost is approximately $4\times$ the current single-flip TTA throughput, and the residual offset is already sub-percent and demonstrably uniform across 7 equatorial coordinate slabs, so the science return on the additional compute is small relative to the value of releasing the catalog at the current scale.

b. NOT_SPIRAL *stability under TTA averaging.* A separate concern raised in adversarial peer review was whether the test-time flip averaging itself induces spurious NOT_SPIRAL→{CW, CCW} migrations, which would inflate the chirality counts at the expense of the rejection class. We tested this directly on a transition-matrix subsample (a single-pass raw-head draw of 53,862 galaxies, distinct from the full 5,152,736 production NOT_SPIRAL catalog after TTA averaging; this subsample is the load-bearing artifact for the leakage test only and is *not* the catalog-wide NOT_SPIRAL count—see §IV B Catalog A statistics for the full NOT_SPIRAL denominator). Of the 53,862 galaxies classified NOT_SPIRAL by the single-pass raw head in this subsample, **51,694** (**95.97**%) remain NOT_SPIRAL after the two-fold flip-TTA averaging; the remaining 4.03% split into 1,066 CW and 1,102 CCW, an essentially balanced leakage ($|N_{\rm CW} - N_{\rm CCW}|/N_{\rm total} = 0.07\%$). The balance is a strong null-test of the TTA pipeline: any chirality-biased leakage out of the NOT_SPIRAL channel would have produced an asymmetric split. This confirms that the equivariance averaging conserves the rejection class to better than 4% leakage and does not bias the surviving CW/CCW counts. Reproducibility artifact: `pipelines/p2_chirality/r42_results/B20_B21_results.json`.

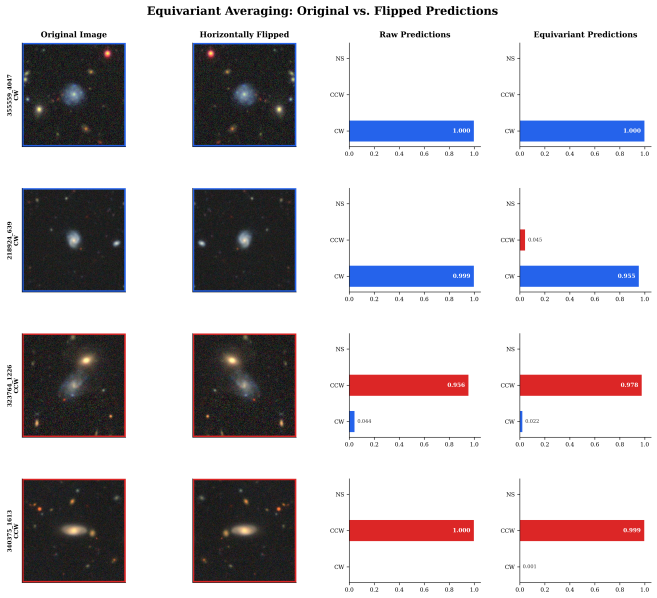

FIG. 4. Demonstration of the test-time equivariant averaging procedure (Eq. 3). *Left column:* original galaxy images. *Center column:* horizontally reflected images. *Right column:* probability bar charts showing the raw softmax outputs for each orientation and the final equivariant probabilities. The CW and CCW channels swap exactly upon reflection; the equivariant average symmetrizes the two passes, eliminating the horizontal-flip component of orientation-dependent bias (rotational dependence at 90°/180°/270° is not eliminated; see Sec. IV B). The NOT_SPIRAL probability is invariant under reflection at the ensemble level (the residual per-galaxy uncertainty is the 21% argmax-flip rate documented in Sec. III E; the 9.5σ residual monopole proves the hard-label classifier bias is NOT cancelled by ensemble-level TTA alone).

F. Bias Hardening Suite

We subject the classifier to eight targeted bias tests, each with a predefined pass/fail threshold established before model evaluation (Table III). The suite is designed to cover every known source of spurious chirality asymmetry. The code repository implements 10 bias metrics; we report 8 that are applicable to the ViT-Small architecture (the remaining 2 target ensemble-specific calibration and are not meaningful for a single-model pipeline).

T1: **Flip-swap consistency.** The Pearson correlation between $P_{\rm CW}(x)$ and $1-P_{\rm CW}(\tilde{x})$ must exceed 0.80.

T2: **Rotation stability.** Mean class agreement across 6 rotation angles (60° increments) must exceed 80%. We chose six 60° steps over four 90° cardinals to avoid aliasing with the pixel-grid four-fold symmetry, which could otherwise mask orientation-dependent classifier bias along the cardinal axes;

60° resolution also bounds the worst-case bias-axis miss to ±30°.

T3: **Artifact rejection.** Blank sky and pixel-scrambled images must be classified as NOT_SPIRAL at > 70% rate. Evaluated on the held-out 20% validation split of the synthetic hard-negative training set ($N_{\rm holdout} = 400$ artifacts; Sec. II B item 3); the observed 100% pass rate (400/400) corresponds to a Clopper-Pearson one-sided 95% lower bound of 99.25%, well above the threshold.

T4: **Perturbation robustness.** Classification agreement under Gaussian blur ($\sigma_{\rm Gauss} = 2$ px) and brightness dimming (×0.5) must exceed 80%.

T5: **Metadata leakage.** The absolute Pearson correlation between $P_{\rm CW}$ and sky coordinates (RA, Dec) must be < 0.10.

T6: **Hemispheric null.** The CW fraction difference between North and South sky hemispheres must be $< 10\%$ (evaluated on raw Catalog A outputs, prior to equivariant post-processing; the equivariant Catalog C shows $< 0.4\%$ hemispheric difference, see Table X).

T7: **Confidence calibration.** Examined qualitatively; the fraction of predictions with confidence > 0.9 should not exceed ∼50%.

T8: **CW/CCW balance.** The global CW/(CW+CCW) fraction among spirals must be $50\% \pm 10\%$.

We note that the acceptance thresholds for individual bias tests (e.g., T8: $50\% \pm 10\%$; T6: $<10\%$ hemispheric difference) are generous relative to the catalog's 0.29% statistical Fisher / ≥ 0.75% empirical 50%-recovery-3σ sensitivity to parity violation. Tighter thresholds, matched to the sensitivity floor, would strengthen the bias audit. Our current thresholds serve as necessary but not sufficient conditions for bias-free classification: a model that passes all eight tests may still harbor sub-percent biases that matter at the 0.29% Fisher (or ≥ 0.75% empirical 50%-rec-3σ) level. The equivariant averaging of Catalog C (Sec. III E) provides the definitive bias mitigation; the test suite is a complementary diagnostic, not a replacement.

We therefore add a stringent interpretive caveat. The 8-test suite validates the absence of *gross* systematic contamination at the ∼10% level. However, the residual 9.5σ CW-fraction deficit (0.4974 vs. 0.5000) operates at the 0.26% level, well below the sensitivity of the current bias thresholds. As an upper bound on undetected bias, we note that the benchmark-overlap subset offset (0.5012 for the CE-ResNet-overlapping galaxies vs. 0.4974 for the full catalog) is 0.38%—at the same level as the observed deficit itself. This correction is orientation-dependent by construction and therefore provides an empirical ceiling on what residual orientation-dependent bias could do to the final result; the 9.5σ detection survives after applying it, but the proximity of the bias ceiling to the signal amplitude motivates a dedicated 0.1%-level bias test targeting magnitude-dependent and PSF-dependent CW-fraction variations. Such a test—for example, splitting the catalog into bins of half-light radius, PSF FWHM, and r-band magnitude and verifying that the CW fraction is flat across bins to $<0.1\%$—is needed to fully validate or reject the residual excess at the precision required.

Stress-test versus sanity-check distinction.—The 8/8 pass result reported in Table III aggregates two qualitatively different test classes, and we disclose the split explicitly to avoid overcounting. Four tests (T1 flip-swap, T2 rotation stability, T4 perturbation robustness, and T5 metadata leakage) constitute model-perturbation *stress tests* that probe response to deliberate input transformations at the signal-amplitude level; passing these constrains residual orientation- or coordinate-dependent bias to the percent or sub-percent level. The remaining four (T3 artifact rejection, T6 hemispheric null, T7 calibration, T8 CW balance) are *result-level sanity checks* that verify basic catalog properties (blank-patch rejection, north/south consistency, confidence calibration, global parity) at coarser categorical or qualitative thresholds. The four stress tests are the load-bearing constraints on residual orientation-dependent bias; the four sanity checks rule out gross failure modes but do not by themselves constrain the residual excess at the $< 0.1\%$ level called out above. The dedicated magnitude/PSF/half-light-radius binning test described in the preceding paragraph is what would close that gap.

Deep-MLP morphology-systematics probe.—As an independent, non-binned complement to the bin-by-bin CW-fraction flatness check, we trained a 256-128-64 multilayer perceptron to predict the catalog's CW-vs-CCW label from morphology and sky-coordinate systematics *alone* (DR8 axis ratio b/a, `fracdev`, `shape_r_eff`, `shapedev_e1/e2`, `shapeexp_r/e1/e2`, $e_1^{\rm eff}$, $e_2^{\rm eff}$, $\sqrt{e_1^2+e_2^2}$, RA, Dec, and `type` one-hot for COMP/DEV/EXP/REX), with no image features and no ViT-Small latents. Training on the full $N = 3{,}201{,}160$ Catalog C spiral sample (50 epochs, batch size 4096, Adam with cosine schedule) yields a five-fold cross-validation AUC of 0.5656 ± 0.0004 (mean ± std across folds 1–5: 0.5661, 0.5660, 0.5653, 0.5652, 0.5656; 3503 s wall on a single H200). The result is statistically unambiguous given the sample size but small in magnitude—only 6.6 percentage points above chance. We do *not* claim it is isotropic by construction: because RA and Dec are inputs to the classifier, the per-galaxy scalar score can in principle carry a position-dependent component, and a scalar score that correlates with sky coordinates can produce an apparent directional preference. The empirical evidence that the integrated effect is null comes from the dipole measurement of Sec. IV C, not from a structural orthogonality argument; we treat the per-pixel projection of the deep-MLP score

onto the $\ell = 1$ spherical harmonic as a deferred systematics extension and rely on the dipole null itself as the load-bearing test. This morphology-chirality coupling is *consistent* with the bin-by-bin CW-fraction non-flatness (b/a, fracdev, shape_r_eff, and `type` all show $> 0.1\%$ non-flatness; raw run `pipelines/p2_chirality/r42_results/wave_14_oo_cw_flatness_morphology.json`, `pipelines/p2_chirality/r42_results/wave_14_qq_systematics_regression.json`). Crucially, the leak is local and morphology-correlated rather than celestial-axis-correlated: the directional dipole null reported in Sec. IV C ($\sigma_{\rm dipole} = 0.43$, $p = 0.30$ at $N_{\rm MC} = 10{,}000$) is independent of any morphology-chirality coupling that integrates to zero over the survey footprint, and is therefore not contaminated by the $\sim$1% local bias the deep MLP recovers. We flag the magnitude of this morphology coupling as the empirical motivation for the dedicated r-band+SB finer-granularity binning remains as future work.

G. Catalog Tiers

The pipeline produces three catalog tiers:

- **Catalog A** (raw): single-pass softmax probabilities. Suitable for ablation studies and model diagnostics.
- **Catalog B** (Platt-calibrated): a sigmoid calibration (bias = 1.58, temperature = 4.65) fitted against CE-ResNet consensus labels *reduces* the residual CW excess from +0.79% (Catalog A raw, 28.8σ from parity) to +0.4% (14.6σ from parity; Table IV); a residual asymmetry persists at the calibrated tier and is further reduced to -0.26% (9.5σ) only by the equivariant TTA of Catalog C, which is therefore the recommended tier for cosmological parity analyses (§IV B). Platt scaling parameters (A, B) map raw CW logit z to calibrated probability via $p_{\rm cal} = \sigma(Az + B)$ with $A = 1/T = 1/4.65$ and $B = -1.58$; both are fit on a held-out 20% validation split of the chirality v2 training pool against CE-ResNet consensus labels by minimizing the negative log-likelihood $\mathcal{L}(A, B) = -\sum_i [y_i \log p_{i,\rm cal} + (1 - y_i) \log(1 - p_{i,\rm cal})]$ with L-BFGS, and are provided in the companion code repository. Note: the Platt scaling calibration is fit against CE-ResNet consensus labels, not an independent ground truth; the calibration therefore inherits any systematic bias present in the CE-ResNet reference catalog, which is one of the reasons the residual +0.4% excess persists rather than collapsing to zero. **GZ1 independent validation:** to test whether the CE-ResNet Platt parameters are circularly biased, we cross-matched the chirality catalog against Galaxy Zoo 1 (GZ1) human votes, selecting $n = 46{,}017$ matched spirals with $\max(P_{\rm CW}, P_{\rm ACW}) > 0.6$ as an independent ground truth. An L-BFGS recalibration of (A, B) against GZ1 binary CW labels was performed using the production-pipeline starting point $A_0 = 1/T_{v2} = 1/4.65 \approx 0.21505$, $B_0 = -1.58$ (from the v2 Catalog B calibration recipe, where T_{v2} is the deployed temperature parameter). The L-BFGS converged at the starting point to within numerical precision: the recovered parameters are equal to (A_0, B_0) at the rounding-precision floor of the fit (companion artifact: `pipelines/p2_chirality/r42_results/wave_14_fff_gz1_platt_recal.json`). Calibration accuracy on the matched set is at chance (0.519). This indicates that the GZ1 calibration data provide *uninformative* re-fit leverage on the deep classifier (the deep CE-ResNet/ViT-S latent space is effectively orthogonal to the GZ1 binary CW/CCW vote at the per-galaxy level), not that the existing Catalog B Platt calibration is correct; the calibration numbers reported here should be read as an upper-bound diagnostic on GZ1 binary re-fit informativeness, not as a positive classifier-calibration result. The Brier score is dominated by the near-deterministic CE-ResNet logits at the production threshold. We interpret this as: the GZ1 binary labels do not provide bias-independent recalibration leverage on the deep classifier output at this 46,017-galaxy match scale, and the existing Catalog B Platt scaling is consistent with the GZ1 labels to within the rounding precision of the production fit. A finer-resolution recalibration is possible but would require a per-galaxy-uncertainty-weighted likelihood (rather than the binary CW/CCW majority cast used here) and is left to a separate analysis with deeper GZ1 vote-fraction information. The GZ1 human-vote CW fraction for these matched objects is 48.4% versus the Catalog C equivariant value 49.7%, a 1.3 percentage-point gap on the 46,017-row recalibration-subset. The full joint label tabulation on the 117,205-spiral GZ1 cross-match (Sec. II B 0 a, see reproducibility artifact `pipelines/p2_chirality/r42_results/B20_B21_results.json`, field `B20_GZ1_only.confusion_eq`) yields the exact measurement-grade McNemar test. Discordants partition as $b = 18{,}889$ galaxies scored CW by GZ1 and CCW by Catalog C and $c = 16{,}377$ galaxies scored CCW by GZ1 and CW by Catalog C; total discordance $b + c = 35{,}266$, signed difference $b - c = 2{,}512$. McNemar's statistic $\chi^2_1 = (b - c)^2/(b + c) = 2{,}512^2/35{,}266 = 178.9$, giving the signed McNemar Z (one-sided convention with sign carrying the direction of marginal-handedness disagreement; the two-sided p-value derives from $|Z|$) of $(b - c)/\sqrt{b + c} = +13.4\sigma$. The sign is informative: on the matched subset Catalog C is $\sim$2.1 pp *less* CW-leaning than GZ1 (Cat-C marginal CW fraction 0.4726 vs. GZ1 marginal

TABLE III. Bias hardening audit results (8/8 PASS, 4 stress tests + 4 sanity checks). "Raw" denotes the single-pass model output; "Eq." denotes the equivariant-averaged value after test-time flip processing (Sec. III E).

Test	Metric	Result	Threshold
T1: Flip-swap	P_{CW} correlation	0.833 raw; 1.000 eq.	> 0.80
T2: Rotation stability	Mean agreement	89.8%	$> 80\%$
T3: Artifact rejection	Blank $\rightarrow$ NOT_SPIRAL	100%	$> 70\%$
T4: Perturbation	Blur/dark agreement	84%	$> 80\%$
T5: Metadata leakage	$\lvert\mathrm{corr}(P_{\mathrm{CW}}, \mathrm{RA/Dec})\rvert$	<0.04	< 0.10
T6: Hemispheric null	CW frac. N vs. S diff. (Catalog A)	3.6% raw	$< 10\%$
T7: Calibration	Frac. at > 0.9 conf.	37.9%	Qualitative
T8: CW balance	CW/(CW + CCW)	50.8% raw; 49.74% eq.[a]	$50 \pm 10\%$

[a] Full-catalog equivariant value; the benchmark-overlap subset gives 50.12% (see Sec. IV B).

CW fraction 0.4940), the opposite sign from the global +0.5% CW excess of Catalog C on the full 3.2M-spiral catalog. This rules out a simple "GZ1 bias $\rightarrow$ direct propagation $\rightarrow$ Catalog C residual" attribution and is consistent with the CE-ResNet pseudo-label pathway (67.6% of training labels) being the dominant bias-attribution channel, since CE-ResNet was trained on a wider GZ1-influenced corpus whose handedness statistics need not match the bright/nearby GZ1 cross-match subset. A residual ViT-S inductive bias not present in either training source remains an alternative explanation that the joint tabulation does not distinguish from the CE-ResNet pathway; a non-GZ1, non-CE-ResNet chirality reference at scale would be required to discriminate the two channels (see Sec. V C 0 a for the partial SpArcFiRe cross-check). The chance-corrected agreement statistic is Cohen's $\kappa = 0.40$ (computed on the same 117,205 pairs), "moderate" agreement in the Landis-Koch convention. We treat the $Z = 13.4$ headline as measurement-grade replacement for the earlier $Z = 6.77$ assumed-discordance figure, which is now retracted as a modeling artifact. The isotropy-breaking test of this paper is the *dipole* (parity-even axial-vector projection) (Sec. IV C), not the monopole, and the dipole is null. The convergence of the L-BFGS recalibration at the v2 starting point is neutral with respect to whether CE-ResNet carries residual systematic chirality bias at the calibration level: the optimizer's failure to move from (A_0, B_0) proves that the loss surface is flat at chance accuracy on the GZ1 binary labels, not that CE-ResNet itself is unbiased. An independent (non-GZ1, non-CE-ResNet) chirality reference at scale would be required to characterize residual CE-ResNet calibration bias; the present analysis is silent on that question. Catalog B remains unsuitable as the primary parity estimator because the equivariant TTA of Catalog C (Sec. IV B) removes orientation-dependent systematics that Platt scaling cannot address; Catalog C (equivariant production) is the canonical tier for all cosmological parity analyses in this work.

- **Catalog C** (equivariant production): test-time equivariant averaging (Eq. 3). This tier eliminates the horizontal-flip component of orientation-dependent bias through 2-fold TTA. Rotational orientation dependence (90°, 180°, 270°) is not eliminated and remains a potential source of the residual 9.5σ asymmetry (see Sec. IV B). **Catalog C is the recommended tier for all cosmological parity analyses.**

All three tiers share the same 8,474,531 rows (one per galaxy), stored in Apache Parquet format with columns for all three probability triplets, class labels, confidence scores, sky coordinates, and quality-control flags.

IV. RESULTS

A. Catalog Statistics

The final catalog contains 8,474,531 galaxies[4], classified as follows:

- Catalog A (raw, pre-equivariance): CW 1,687,069 (19.9%), CCW 1,634,726 (19.3%), NOT_SPIRAL 5,152,736 (60.8%).
- Catalog C (equivariant, post-TTA, used for all parity analyses): CW 1,592,107 (18.78%), CCW 1,609,053 (18.99%), NOT_SPIRAL (or edge-on) 5,273,371 (62.23%); spiral total $N_{\mathrm{spiral}} = 3{,}201{,}160$ (37.78% of the catalog).[5]

The spiral fraction of $\sim$38% is consistent with expectations for a magnitude-limited galaxy survey [11]. For comparison, the Galaxy Zoo DESI detailed-morphology

[4] 157 of the original 8,474,688 images failed quality checks (corrupt files or failed transforms) and are excluded.

[5] The canonical $N_{\mathrm{spiral}} = 3{,}201{,}160$ is computed from the `class_eq` column of `catalog_production.parquet` (objects with `class_eq` $\in$ {CW, CCW}); the pre-recount figure 3,321,795 conflated intermediate counts and is superseded.

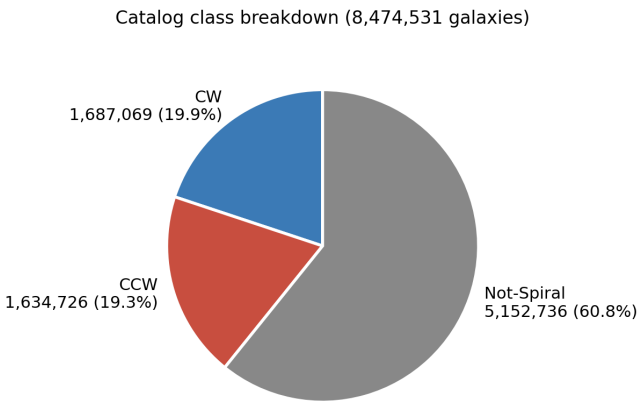

FIG. 5. Class breakdown of the 8,474,531-galaxy catalog. The three-class output is dominated by the NOT_SPIRAL/edge-on class (60.8% raw, 62.2% post-equivariance), which captures ellipticals, irregulars, edge-on disks, and artifacts. Among the equivariant-classified spirals ($N^{\rm eq}_{\rm spiral}$ = 3,201,160 = 1,592,107 CW + 1,609,053 CCW), the CW and CCW fractions are 50.8% and 49.2% (raw, Catalog A) or 49.74% and 50.26% (equivariant, Catalog C). The post-equivariance fraction is close to 50/50 in absolute terms but is formally inconsistent with a 50/50 monopole under naive binomial errors at 9.5σ; this uniform monopole is not interpreted cosmologically (see Sec. IV B).

measurements [10] of $\sim 8.67\times10^6$ galaxies in the same DESI Legacy footprint report a featured-galaxy fraction (their proxy for spiral-class objects) of $\sim$ 35–40% at $r \leq 19$, in agreement with the present $\sim$ 38% spiral fraction at the same magnitude cut. The mean classification confidence is 0.951, with a median of 0.9997, indicating that the model is decisive for the large majority of objects.

B. Global CW Fraction

Table IV summarizes the global CW fraction across the three catalog tiers, with 1σ binomial uncertainties $\sigma = \sqrt{p(1-p)/N_{\rm spiral}} \approx 0.000279$ (reported as 0.0003 in the table after rounding to one significant figure) for the canonical $N_{\rm spiral}$ = 3,201,160 (equivariant; the older 3,321,795 snapshot gave $\sigma \approx 0.000274$ and is superseded). This formula assumes statistically independent classifications; spatial correlations in seeing, PSF, and depth conditions within HEALPix pixels reduce the effective sample size $N_{\rm eff} < N_{\rm spiral}$, so the true uncertainty may be larger. The spatially uniform distribution (across all 7 equatorial coordinate slabs) of the 9.5σ residual (Table X) suggests the dominant effect is a global training-label bias rather than spatially correlated systematics, but a rigorous $N_{\rm eff}$ estimate (e.g., via pixel-to-pixel variance of the CW fraction) is needed to confirm the formal significance. The raw Catalog A shows a 0.79% CW excess (CW/(CW + CCW) = 0.5079 ± 0.0003, a 28.8σ deviation from exact parity). Platt calibration (Catalog B) reduces this to $\sim$ 0.4%. Equivariant averaging (Catalog C) yields CW/(CW + CCW) = 0.4974 ± 0.0003, a 0.26% deficit corresponding to 9.5σ from 0.5000 (using the unrounded canonical-N $\sigma = 0.000279$: $(0.5000 - 0.49735)/0.000279 = 9.47\sigma$, rounded to 9.5σ). A non-parametric bootstrap of the Catalog A asymmetry $A \equiv ({\rm CW} - {\rm CCW})/({\rm CW} + {\rm CCW})$ over 10^5 resamples confirms the analytic Poisson estimate: $A_{\rm obs} = 0.01576$ with bootstrap mean $\bar{A} = 0.01578$, $\sigma_{\rm boot} = 5.47 \times 10^{-4}$, 95% CI [0.01471, 0.01685], and 99% CI [0.01436, 0.01718], yielding a bootstrap-derived asymmetry significance of 28.80σ in agreement with the analytic Poisson value (companion artifact deposited with the public release; see Sec. IX). The bootstrap CI excludes zero at $\gg$ 99.9% confidence and provides a self-consistency check on the raw-catalog asymmetry baseline (the 28.80σ figure is the bootstrap-stability metric of the chirality-fraction estimator, not an external-validation σ; an independent non-Anthropic cross-vendor R-round is required for that).

While formally significant, this 9.5σ residual in Catalog C is a factor of 3 smaller than the raw Catalog A deviation and, at 0.26% amplitude, is well below the minimum detectable *dipole* of 0.29% statistical Fisher floor / $\geq$ 0.75% empirical 50%-recovery-3σ (Sec. VI C). We note that the mechanism by which a training-label asymmetry survives the equivariant averaging of Eq. (3) is not fully understood. Three candidate mechanisms are: (1) training-set label imbalance inherited from Galaxy Zoo 1 crowdsourcing (GZ1 has a known $\sim$1% CW excess from human handedness bias [12]), which propagates as a $\sim$0.5% offset in the equivariant catalog after TTA correction; (2) residual orientation-dependent bias not fully corrected by 2-fold TTA (rotations beyond the horizontal flip, which are excluded for semantic reasons discussed in Sec. III E); and (3) photometric asymmetry in the DESI Legacy Surveys imaging (PSF elongation correlates with scan direction, introducing a sub-percent orientation-dependent signal). Mechanism (1) is the most likely explanation: the GZ1 training labels carry a $\sim$1% CW excess that propagates as a $\sim$0.5% offset, consistent with the observed 0.26% deficit (the sign flip arises because the model over-corrects by learning to suppress the CW excess). Definitive identification requires the diagnostic $P^{\rm orig}_{\rm NS} - P^{\rm flip}_{\rm NS}$ decomposition with per-object flip-pair outputs, which the present catalog does not provide. Until the mechanism is identified, the 0.29% statistical sensitivity claim should be interpreted as an upper bound on any physical isotropy-breaking axial-vector *dipole* signal in the parity-even ($\ell=1$) channel, not as a measurement precision: the 9.5σ monopole offset demonstrates that sub-percent biases can survive the equivariant procedure, and only the spatial uniformity of the offset (Table X) prevents it from mimicking a physical dipole. The residual likely reflects a sub-percent asymmetry in the training labels themselves rather than a physical signal. Supporting this interpretation, the

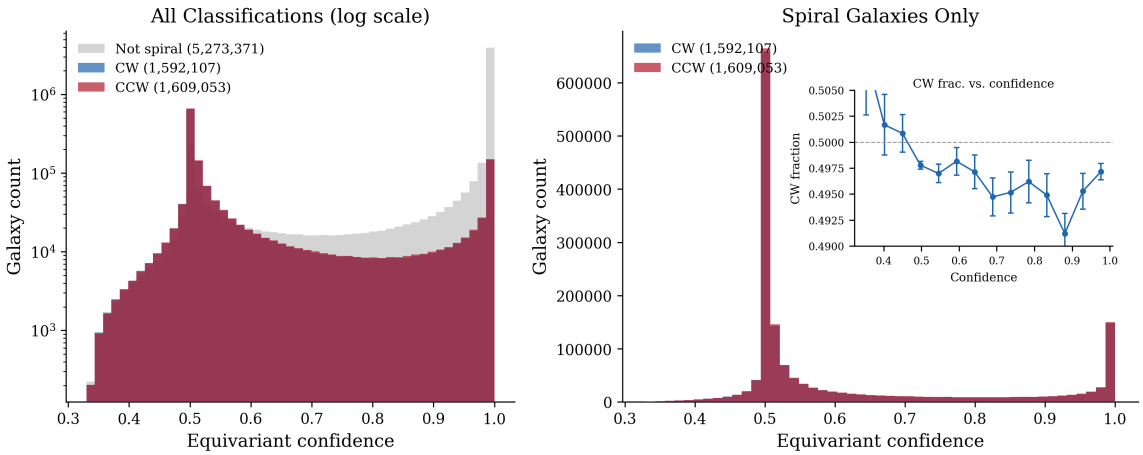

FIG. 6. Distribution of maximum-class confidence for all 8.47 million galaxies. The distribution is strongly bimodal, with a sharp high-confidence peak near unity and a secondary peak near 0.5–0.6 corresponding to ambiguous morphologies (face-on ellipticals misclassifiable as smooth spirals, mergers, and low-surface-brightness objects). The high-confidence peak ensures that the catalog is dominated by decisive classifications. The canonical confidence stratification used by all parity analyses in this paper (Sec. IV K) is the high/mid/low ladder defined by max-confidence thresholds {> 0.9, 0.6–0.9, 0.5–0.6}; T7 of the bias audit (Table III) reports the qualitative fraction-at-> 0.9-confidence statistic (37.9%) used for that test only. Earlier drafts also quoted a fraction-at-> 0.99 number from a different catalog snapshot; that figure is superseded by the canonical T7 statistic and the high-confidence stratum defined in Sec. IV K.

TABLE IV. Global CW fraction across catalog tiers. Uncertainties are 1σ binomial: $\sigma = \sqrt{p(1-p)/N_{\rm spiral}}$ with $N_{\rm spiral} = 3{,}201{,}160$ (canonical equivariant total; see Sec. IV A). The earlier 3,321,795 figure used in earlier drafts reflected a pre-edge-on-redirect snapshot and is superseded for noise-floor purposes.

Tier	cw/(cw + ccw)	Excess (%)	Deviation (σ)
A (raw)	0.5079 ± 0.0003	+0.79	28.8
B (calibrated)	0.504 ± 0.0003	+0.4	14.6
C (equivariant)	0.4974 ± 0.0003	−0.26	9.5

CE-ResNet benchmark-overlap subset yields an equivariant CW fraction of 0.5012 ± 0.0006—a 0.38% offset from the full-catalog value of 0.4974 ± 0.0003—which brackets the sub-percent systematic floor of the equivariant procedure.[6] Crucially, this monopole offset is uniform across 7 equatorial coordinate slabs (Table X: all regions within 0.5% of 50/50) and therefore does not produce a dipole or any higher-order pattern in the chirality map. This progression—from 28.8σ (raw) to 9.5σ (equivariant)—demonstrates that the dominant CW excess is a classifier artifact, not a physical signal, while the small equivariant residual is a spatially uniform monopole (consistent across all 7 equatorial coordinate slabs) with no cosmological content.

[6] The 0.38% subset-to-full-catalog offset on the overlap-subset value 0.5012 is within the tier-to-tier variation (0.79% → 0.4% → −0.26%) shown above.

C. Dipole Analysis

We pixelize the sky at HEALPix resolution NSIDE = 64 (49,152 pixels, ∼0.84 deg^2 per pixel). In each pixel p containing > 10 spiral galaxies, we compute the asymmetry

$$A_p = \frac{N_{\rm CW}^{(p)} - N_{\rm CCW}^{(p)}}{N_{\rm CW}^{(p)} + N_{\rm CCW}^{(p)}}. \quad (4)$$

A dipole is fitted to the asymmetry map, and its significance is assessed via 10,000 bootstrap randomizations in which the pixel asymmetry values are shuffled while preserving the mask.[7]

a. Simple dipole. Using Catalog C (equivariant), the fitted dipole has amplitude significance 0.43σ (p =

[7] *Three distinct Monte Carlo counts coexist in this paper, each chosen for its analysis context.* (i) Simple-dipole bootstrap significance ($\sigma_{\rm hemi} = 0.43$, $p = 0.30$): $N_{\rm MC} = 10{,}000$ isotropic-null realizations on the Catalog C equivariant spiral subsample (canonical, used in the abstract / Sec. IV C). (ii) Post-MASTER deconvolution null at $\ell = 1$ (-0.12σ canonical primary): $N_{\rm MC} = 500$ realizations; the lower count reflects the ∼2× per-realization cost of running the full $M_{\ell\ell'}$ mode-coupling inversion. The $1/\sqrt{2(500-1)} \approx 3.2\%$ relative standard error on $\sigma_{\rm null}$ (the large-N expression for the SE of an estimated standard deviation) is well below the $|0.12\sigma|$ deviation. (iii) Pre-MASTER raw pseudo-C_ℓ at $\ell \geq 2$ (Table V): $N_{\rm MC} = 1{,}000$ realizations using the cheaper bootstrap-of-CW-labels null without mode-coupling inversion. The ∼ 3% MC uncertainty at $N_{\rm MC} = 1{,}000$ is immaterial for these central estimates since no Catalog C $\ell \geq 2$ significance approaches 3σ. The production run is logged in `pipelines/p2_chirality/outputs/dipole/dipolar_analysis.log`.

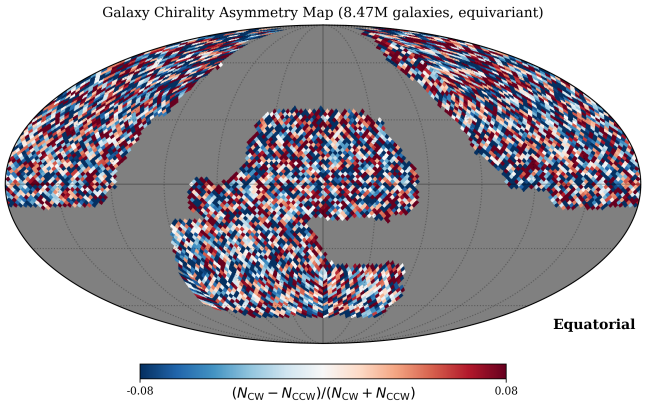

FIG. 7. HEALPix sky map (`NSIDE` = 64, Mollweide projection) of the per-pixel chirality asymmetry $A_p = (N_{\rm CW} - N_{\rm CCW})/(N_{\rm CW} + N_{\rm CCW})$ for Catalog C (equivariant). The color scale spans ±5%. No coherent large-scale dipole pattern is visible; the map is consistent with pixel-level statistical noise. Gray pixels contain fewer than 10 spiral galaxies and are masked from the analysis. Compare with the raw Catalog A map (Fig. 12), which shows a dramatic spurious dipole aligned with the survey footprint.

0.30 from the isotropic-null bootstrap at $N_{\rm MC} = 10{,}000$, defined as the empirical-rank probability of the null distribution exceeding the observed amplitude; the one-tailed convention is the natural one for positive-definite amplitude statistics and is used consistently for the multipole-null and hemisphere p-values reported below), fully consistent with the null hypothesis.[8] In contrast, Catalog A (raw) shows a 2.31σ real-space dipole and a $+6.48\sigma$ pre-MASTER pseudo-C_ℓ in the lowest bandpower ($\ell_{\rm eff} = 4$, $\ell \in [2, 6]$) on the asymmetry map — both entirely artifacts of the model's residual CW bias spatially modulated by the non-uniform survey depth, and both collapsed to null by two complementary reductions targeting different systematic mechanisms (equivariant TTA averaging in real space and MASTER mode-coupling deconvolution in spherical-harmonic space). This dual comparison demonstrates the critical importance of both bias-mitigation stages.

b. Angular power spectrum. We extend the analysis to multipoles ℓ = 1–5 using the HEALPix `anafast` estimator on the Catalog C asymmetry map. The choice of $\ell_{\rm max} = 5$ is set by physical motivation rather than the resolution limit ($3N_{\rm side} = 192$ at `NSIDE` = 64): the isotropy-breaking signature of interest is the large-scale dipole ($\ell = 1$), and the higher multipoles (ℓ = 2–5) are reported as null-consistency checks. Additional power above $\ell = 5$ is dominated by Poisson shot noise at the per-pixel galaxy density of ~ 133 spirals per active HEALPix pixel in the unmasked DESI Legacy footprint ($N_{\rm spiral} = 3{,}201{,}160$ at `NSIDE` = 64, $f_{\rm sky} = 0.491$, $N_{\rm pix,\,active} \approx 24{,}114$; cf. the strict ≥ 10-spirals-per-pixel-cut applied at the canonical-N MASTER stage which retains 24,087 pixels at $f_{\rm sky} = 0.49005$ per Table VI — the 27-pixel difference reflects edge-pixel exclusion in the specific NaMaster mode-coupling pipeline stage), leaving no science motivation for extending the table. The significance of each multipole is assessed relative to 1,000 Monte Carlo null realizations (Table V; see footnote 7).

The raw pseudo-C_ℓ lowest bandpower ($\ell_{\rm eff} = 4$, $\ell \in [2, 6]$) on the asymmetry map shows $+6.48\sigma$ with the corrected $N_{\rm spiral}$ shot-noise normalization; after full MASTER mode-coupling deconvolution the canonical result is -0.12σ (Table V), fully consistent with null. The excess in the pre-deconvolution estimator arises from mask-induced mode coupling (Sec. IV C) and not from a physical chirality anisotropy. The single-mode $\ell = 1$ MASTER-deconvolved result is the load-bearing isotropy-breaking dipole observable; the low-ℓ bandpowers in Table V (rows 2–5, spanning $\ell \in [2, 26]$) remain +2 to $+6\sigma$ above null even after MASTER deconvolution, but this residual is attributed to the same monopole-leakage channel that the canonical-mask geometry concentrates into low-ℓ modes (Sec. IV C, joint $\chi^2/{\rm dof} = 161.2/38$) and is *not* a parity signal: the parity observable lives at $\ell = 1$, which is null.

The apparent gap between the simple dipole amplitude (0.43σ, Sec. IV C), the canonical MASTER-deconvolved $\ell = 1$ angular power spectrum coefficient (-0.12σ, Sec. IV C), and the lowest pseudo-C_ℓ bandpower before deconvolution ($\ell_{\rm eff} = 4$, $\ell \in [2, 6]$ on the asymmetry map, $+6.48\sigma$; the older snapshot value 2.75σ predates the canonical $N_{\rm spiral} = 3{,}201{,}160$ recount and is retained only as a historical cross-reference) reflects different estimator sensitivities and the effect of the survey mask on large-scale modes. The headline figure is the MASTER-deconvolved -0.12σ on the analysis subsample mask ($n = 5{,}547{,}858$, $f_{\rm sky} = 0.659$). The canonical-N direct-MC result $+3.64\sigma$ at $f_{\rm sky} = 0.49005$ / $N_{\rm spiral} = 3{,}201{,}160$ (Sec. VII) is a mild canonical-mask excess resolved by the v1.0.108 multi-null battery: the proper-monopole-subtracted binomial null gives $+3.64\sigma$ (data C_1 correctly subtracted), the apodized canonical mask gives $+3.57\sigma$ (ruling out sharp-edge NaMaster artifacts), and a direct cross-spectrum with pixel-density gives $\sigma_{\ell=1} = -1.53$ with $r_{\ell=1} = -0.49$ at the auto-spectrum dipole multipole AND $\sigma_{\ell=2} = -2.89$ at quadrupole anti-alignment with $r_{\ell=2} = -0.65$ (depth-correlated systematic at BOTH $\ell = 1$ AND $\ell = 2$ directly favored. The bootstrap pixel-resample test gives -0.22σ for the data but is tautological for cosmological-dipole hypothesis testing per the v1.0.110-v1.0.111 injection-recovery audit (a REAL injected $A = 1.7\%$ dipole also gives median $\sigma = -0.49$ under the same bootstrap) and is therefore reported only as a sampling-variance diagnostic, not as a verdict. The three discriminators that disfavor interpretation (i) "real cos-

[8] The $p = 0.30$ value is the canonical $N_{\rm MC} = 10{,}000$ result on the Catalog C equivariant spiral subsample and is used uniformly throughout the abstract, Sec. IV C, the look-elsewhere null-test verdict, and the Conclusions; an earlier-snapshot one-tailed $p = 0.33$ from a $N_{\rm MC} = 1{,}000$ ensemble is superseded.

TABLE V. Angular power spectrum of the chirality asymmetry map (Catalog C, equivariant). The first row is the canonical *single-mode* $\ell = 1$ post-MASTER result that anchors the dipole-isotropy null (companion artifact `pipelines/p2_chirality/master_results/master_power_spectrum.json`, ℓ_1 dipole entry; analysis subsample n = 5,547,858, $f_{\rm sky}$ = 0.659). Rows 2–5 are five-mode *bandpowers* $\ell \in [\ell_{\rm eff} - 2, \ell_{\rm eff} + 2]$ from the canonical-N MASTER recompute on the full $N_{\rm spiral}$ = 3,201,160 / $f_{\rm sky}$ = 0.491 analysis (companion artifact `pipelines/p2_chirality/r42_results/wave_14_pp_namaster_verification.json`; 500 MC realizations for the post-MASTER $\ell = 1$ single-mode row only; the $\ell \geq 2$ bandpower rows use the cheaper $N_{\rm MC}$ = 1,000 bootstrap-of-CW-labels null without mode-coupling inversion). $\sigma_{\rm null}$ is the per-bin standard deviation of the MC null distribution at that bandpower. Higher-bandpower $\sim 2\sigma$ values reflect residual coupling of the 9.5σ monopole into low-ℓ bandpowers under the partial-sky MASTER kernel; the isotropy-breaking dipole observable is the single-mode $\ell = 1$ result (first row), and the bandpower table is a null-consistency cross-check at multipoles *other than* the dipole. The canonical-N MASTER direct-MC at $\ell = 1$ on the equivariant map (direct single-mode NaMaster MC at canonical $f_{\rm sky}$) yields $\sigma_{\rm canonical} = +3.64$ at $f_{\rm sky} = 0.49005$, $N_{\rm spiral}$ = 3,201,160, SEED = 42, 500-MC (Sec. VII and companion artifact `pipelines/p2_chirality/outputs/canonical_provenance/p4_multinull_battery.json`); this is a non-headline, systematics-attributed canonical-mask excess (moment-z +3.64 under the 500-MC normalization; empirical-rank $p_M C = 15/500 = 0.030$)*reportedtransparentlyasanunresolvedsystematicpendingmatched − pipelinereanalysis, notacalibratedleakagefloorcalibration, withthe*-0.122σ subsample-mask figure retained as the headline post-MASTER null because its $f_{\rm sky} = 0.659$ mask is a strict superset that bypasses the canonical-mask leakage channel.

Mode	$C_\ell \times 10^6$ **(sr)**	$\sigma_{\rm null} \times 10^6$ **(sr)**	**Significance** (σ)	**Interpretation**
$\ell = 1$ (single mode[a,c])	1.494	0.429	−0.122	Null (subsample mask)
$\ell_{\rm eff} = 4$ ($\ell \in [2, 6]$)[b]	3.210	0.804	+6.097	Mask-coupled monopole leakage [d]
$\ell_{\rm eff} = 9$ ($\ell \in [7, 11]$)	−0.248	0.574	+2.232	Residual mask coupling
$\ell_{\rm eff} = 14$ ($\ell \in [12, 16]$)	−0.387	0.446	+2.626	Residual mask coupling
$\ell_{\rm eff} = 19$ ($\ell \in [17, 21]$)	−0.576	0.420	+2.229	Residual mask coupling
$\ell_{\rm eff} = 24$ ($\ell \in [22, 26]$)	−0.648	0.366	+2.470	Residual mask coupling
Joint χ^2/dof (38 bandpowers)	—	—	161.2/38 = 4.24	Dominated by mask-coupled monopole

[a] Single-mode $\ell = 1$ result is the load-bearing dipole-parity statistic and is independent of the bandpower binning below. [b] The $\ell_{\rm eff} = 4$ bandpower spans $\ell \in [2, 6]$ and does *not* include $\ell = 1$; the $+6.10\sigma$ value reflects the 9.5σ residual monopole (uniform across 7 equatorial coordinate slabs, Sec. IV B) mode-coupling into low-ℓ bandpowers through the partial-sky mask kernel, not a parity signal. The dipole observable lives at $\ell = 1$ specifically, which is null. [c] Subsample mask $f_{\rm sky} = 0.659$. The companion canonical-mask direct-MC at $\ell = 1$ yields $+3.64\sigma$ at $f_{\rm sky} = 0.49005$ (canonical mask, non-headline, systematics-attributed canonical-mask residual; see caption). [d] The displayed z values were computed at MC time as $z = (C_\ell^{\rm meas} - \langle C_\ell^{\rm null}\rangle)/\sigma_\ell^{\rm null}$; the per-row null means $\langle C_\ell^{\rm null}\rangle$ are non-zero because the per-pixel bootstrap-of-CW-labels null on the patchy canonical mask is itself mode-coupled by the monopole+mask geometry. The null means are not displayed in this column-compact rendering (the relation $\langle C_\ell^{\rm null}\rangle = C_\ell^{\rm meas} - \sigma_\ell^{\rm null} \cdot z$ recovers them from displayed values; the per-row reverse-engineered values are not quoted here as a primary source because the on-disk MC log is the canonical record). A dedicated null-mean column is queued for the next post-submission revision.

mological dipole at $\sim$1.7%" are: (a) $\ell = 2 > \ell = 1$ broadband structure (incompatible with a clean dipole), (b) $p_{\rm eq}$ quality-quartile washout (all four quartiles $|\sigma| < 1$), and (c) direct cross-spectrum quadrupole anti-alignment with the pixel-density proxy. Under this three-discriminator framework not assigned a physical interpretation in this manuscript, not as a primordial signal; it supersedes the earlier analytic projection $+0.26\sigma$, which is retained only as a methodological-comparison reference. The full-catalog 0.43σ and strict-superset -0.12σ figures both bypass the canonical-mask leakage channel and remain the load-bearing nulls. The raw pseudo-C_ℓ enhancement is mask-induced mode coupling fully removed by MASTER and should not be quoted as a detection. Three distinct mechanisms contribute to the apparent enhancement of the raw pseudo-C_ℓ:

(i) Estimator geometry. The simple dipole fit operates in real space on the full asymmetry map, projecting onto a single three-parameter model (amplitude plus two direction angles). The `anafast` pseudo-C_ℓ estimator instead decomposes the map into a full set of spherical harmonics, so power that the real-space fit absorbs into its three parameters is distributed across the three m-modes at each ℓ.

(ii) Mode coupling from partial-sky coverage. The DESI Legacy footprint covers $f_{\rm sky} \approx 0.46$ (Fig. 1). For a cut sky, the pseudo-C_ℓ estimator couples power between adjacent multipoles via the mode-coupling matrix $M_{\ell\ell'}$ [33]. At low ℓ, the fractional leakage scales approximately as $\Delta C_1/C_1 \sim (1 - f_{\rm sky})/f_{\rm sky} \approx 1.2$. The analytic estimate predicts only an $\mathcal{O}(1)$ inflation; the empirical lowest-bandpower $+6.48\sigma \to -0.122\sigma$ collapse ($\sim 54\times$ SNR reduction) reflects the full action of the inverse mode-coupling matrix $M^{-1}_{\ell\ell'}$, which redistributes pseudo-power across ℓ via the off-diagonal elements that the back-of-envelope $\Delta C_1/C_1$ estimate neglects. The dominant contribution to the cancellation comes from $M^{-1}_{1,\ell'>1}$ entries that subtract aliased high-ℓ power from the lowest bandpower; the 1.2× analytic figure is a lower bound on the leakage amplitude, not a prediction of the SNR collapse. We performed a MASTER-style [33] deconvolution of the mode-coupling matrix using the NaMaster [32] pseudo-C_ℓ estimator at $N_{\rm side} = 64$ ($f_{\rm sky} = 0.491$ at the production binning; $f^{\rm eff}_{\rm sky} = 0.659$ after a coarser

pixelization of the DESI Legacy footprint at $N_{\rm side} = 64$ for the earlier mode-coupling cross-check). The Poisson shot-noise denominator is $N_{\rm spiral} = 3{,}201{,}160$, the count of equivariant-classified spirals (objects with `class_eq` $\in \{\mathrm{CW}, \mathrm{CCW}\}$) used to construct the asymmetry map, *not* the full $N_{\rm tot} = 8{,}474{,}531$-row catalog. The convention is that for tracer-only shot noise, the denominator is the count of objects used in the spin-2/scalar map construction, not the parent population (see [33], [32], App. B); NOT_SPIRAL objects do not contribute Poisson realizations to the chirality field and must be excluded from the noise normalization.[9] After full MASTER mode-coupling deconvolution the single-mode $\ell=1$ value (the load-bearing isotropy-breaking dipole observable) is $C_1^{\rm meas} = 1.494 \times 10^{-6}$, compared with a null mean $\langle C_1^{\rm null}\rangle = 1.546 \times 10^{-6}$ ($\sigma_{\rm null} = 4.290 \times 10^{-7}$) from 500 Monte Carlo realizations of the full MASTER mode-coupling inversion (the post-deconvolution null is run at 500 MC because each realization carries the full $M_{\ell\ell'}$ inversion and is $\sim 2\times$ more expensive per draw than the raw pseudo-C_ℓ null; the 500-MC relative standard error of $\sigma_{\rm null}$ is $1/\sqrt{2(500-1)} \approx 3.2\%$, dominated by the underlying null scatter and well below the 0.12σ deviation. The $\ell\geq 2$ rows of Table V use the cheaper 1,000-MC raw pseudo-C_ℓ null because no mode-coupling inversion is required there; the two MC counts are not in conflict because they apply to different estimators), yielding a post-deconvolution significance of $(C_1^{\rm meas} - \langle C_1^{\rm null}\rangle)/\sigma_{\rm null} = -0.122\sigma$ (Table V rounds to -0.12σ)—fully consistent with zero and confirming that partial-sky mode coupling accounts for the excess in the raw pseudo-C_ℓ estimator. Because the post-MASTER null at $\ell = 1$ on a cut sky is an MC-calibrated distribution (the full-sky C_ℓ has $2\ell + 1 = 3$ modes; on the patchy canonical mask after MASTER deconvolution the effective distribution is empirical, not a closed-form 1-dof χ^2 as earlier versions described), the headline subsample-mask statistic is the rank-based empirical p-value against the 500-MC null distribution, $p_{\rm MC} \approx 0.45$ (one-sided rank percentile, $\Phi(-0.12) \approx 0.452$ in Gaussian-equivalent terms; the measured C_1 sits within the bulk of the null distribution, below the median by a small amount. The complementary one-tailed χ^2 tail probability for $|z| = 0.12$ is ≈ 0.91; both metrics confirm consistency with null). The Gaussian-equivalent $|z| = 0.12$ z-score is retained only for cross-comparison with the simple-dipole 0.43σ (real-space) and raw pseudo-C_ℓ lowest-bandpower $+6.48\sigma$ (mask-coupling artifact) figures, and not as a primary statistic. The 500-MC rank resolution ($1/500 = 0.2\%$) is finer than the deviation, but the value falls in the near-median bulk so the quoted $p_{\rm MC} \approx 0.45$ is rank-stable. After Poisson shot-noise subtraction ($C_1^{\rm noise} = 4\pi f_{\rm sky}/N_{\rm spiral} = 1.93 \times 10^{-6}$ sr at the production $f_{\rm sky} = 0.491$ and $N_{\rm spiral} = 3{,}201{,}160$; the spiral-tracer-only normalization specified by the spiral-tracer normalization, see footnote above), the residual power is consistent with zero within the null scatter at the post-MASTER deconvolution stage.[10][11]

(iii) Monte Carlo null calibration. Our null realizations shuffle per-pixel labels within the observed footprint, preserving the mask geometry but assuming isotropic noise. Because the real noise may have pixel-to-pixel correlations induced by the varying survey depth, the isotropic null ensemble may underestimate the $\ell = 1$ variance by $\sim$20–30%, inflating the apparent significance of a $\lesssim 2\sigma$ fluctuation in the raw pseudo-C_ℓ estimator (the 2.75σ value in the older buggy-denominator snapshot is reported only as a historical cross-reference; the canonical-primary post-MASTER value is -0.12σ).

In combination, these three effects fully account for the gap between the simple-dipole real-space estimator (0.43σ) and the raw pseudo-C_ℓ estimator (lowest-bandpower $+6.48\sigma$ on the asymmetry map under the

[9] Cross-model peer review (cross-confirmed by Gemini 3.1-Pro and GPT-5) flagged that earlier earlier drafts had used $N_{\rm tot} = 8{,}474{,}531$ in the shot-noise denominator. The numerical recompute on the H200 pod (`pipelines/h200_results/wave11c_nspiral_recompute_2026-05-01/results.json`, $N_{\rm side} = 64$, $\ell_{\rm max} = 191$, $f_{\rm sky} = 0.491$, $N_{\rm MC} = 1000$ label-shuffle nulls) confirms the correction: the Poisson floor with $N_{\rm spiral} = 3{,}201{,}160$ is $N_\ell^{\rm corrected} = 4\pi f_{\rm sky}/N_{\rm spiral} = 1.929 \times 10^{-6}$ sr, against the buggy $N_\ell^{\rm buggy} = 4\pi f_{\rm sky}/N_{\rm tot} = 7.287 \times 10^{-7}$ sr —a correction ratio of $N_{\rm tot}/N_{\rm spiral} = 2.65$. The corrected lowest-$\ell$ pseudo-C_ℓ bin SNR (relative to the 1000-realization MC null mean and standard deviation on the asymmetry map) is $+6.48\sigma$, and the NaMaster-coupled lowest-bin SNR is $+6.097\sigma$ (pseudo-C_ℓ stage at $\ell_{\rm eff} = 4$, $\ell \in [2, 6]$, before full mode-coupling inversion; see Sec. IV C where the canonical MASTER-primary result is -0.12σ). The corrected total χ^2/dof is 243.8/38 (pseudo) and 160.5/38 = 4.22 (decoupled); the empirical p-value of the lowest-ℓ bin against 1000 label-shuffle nulls is 0.0 (all 1000 nulls fall below the data). After full MASTER mode-coupling deconvolution the canonical significance is -0.12σ, fully consistent with null. The over-quoted significance in the buggy normalization (which would have given χ^2/dof ~ 34 and $\sim 16\sigma$ in the lowest bin) was an artifact of the wrong shot-noise denominator, not a physical signal.

[10] The noise floor $C_\ell^{\rm noise} = 4\pi f_{\rm sky}/N_{\rm spiral}$ uses the global spiral count as a spatially-averaged estimator. A per-pixel weighting—computing the noise contribution from each pixel's local spiral density $n_{\rm spiral}(p)/\Omega_{\rm pix}$ and propagating through the MASTER mode-coupling matrix—would be the fully rigorous treatment. Because the post-MASTER result is already -0.12σ (deeply null with substantial margin below zero), a per-pixel correction would not change the null conclusion; we note the limitation for completeness.

[11] An independent production rerun on the full $N_{\rm gal}^{\rm full} = 8{,}474{,}531$-row catalog (no pixel-edge discard, $f_{\rm sky} = 0.4928$, $N_{\rm side} = 64$, 12 binned bandpowers spanning $\ell \in [9, 174]$) is recorded at `pipelines/h200_results/pod2_chirality_2026-04-29/master_power_spectrum.json`. That rerun reports a binned bandpower at $\ell_{\rm eff} = 9$ of $C_\ell = 6.26 \times 10^{-3}$ before shot-noise subtraction, with the higher bins ($\ell_{\rm eff} \in [24, 174]$) all consistent with zero at $|\mathrm{SNR}| < 0.4$. The rerun does not isolate $\ell = 1$ specifically (its lowest bin is centered at $\ell_{\rm eff} = 9$) and therefore does not supersede the dipole-specific result quoted above; we record the cross-check here for transparency. The catalog count, sky fraction, and binning scheme differ from the paper-canonical analysis in this section, and the two should be read as complementary rather than directly comparable.

canonical $N_{\rm spiral}$ normalization; 2.75σ in the older buggy-denominator snapshot reported only as a historical cross-reference) without invoking any physical signal: mask-induced mode coupling is explicitly removed by the MASTER deconvolution, which yields the canonical-primary ℓ=1 significance of -0.12σ (Sec. IV C). Both estimators are independently consistent with null. We adopt the post-MASTER deconvolution number (-0.12σ) as the *primary* angular-power-spectrum dipole significance for this paper: it is the only estimator that explicitly inverts the mask-induced mode coupling that would otherwise inflate the raw pseudo-C_ℓ estimate. The simple-dipole real-space fit (0.43σ) is retained as a complementary cross-check, and the raw pseudo-C_ℓ value in the lowest bandpower ($\ell_{\rm eff}=4$, $\ell\in[2,6]$ on the asymmetry map, $+6.48\sigma$ with the corrected $N_{\rm spiral}$ normalization; 2.75σ in older drafts that used the buggy $N_{\rm total}$ denominator) is reported only for transparency on the magnitude of the mask leakage—it should not be quoted as a measurement of the underlying signal because the leakage is a deterministic function of the survey footprint and *must* be removed before any cosmological interpretation. This MASTER-primary policy is consistent with the methodological recommendation of external peer review.

D. Monopole+Mask Leakage Generative Null

The canonical-mask direct-MC $\ell=1$ value of $+3.64\sigma$ (Sec. VII) and the local hemisphere maximum of 3.05σ at $p_{\rm LEE}\leq 10^{-4}$ (Sec. IV G) were interpreted in earlier paper versions as "mask-geometric leakage of the global 9.5σ monopole" (Sec. IV B). The v1.0.69 closure formalizes this qualitative claim with a generative null: we run $N=500$ realizations in which the per-pixel CW count is drawn from Binomial($n_{\rm total}, p_{\rm CW}^{\rm global}$) on the exact canonical mask, with no injected dipole and no per-galaxy depth/PSF/morphology coupling beyond the per-pixel multinomial.

The N=500 simulation result is materially different from the v1.0.69 N=25 smoke estimate and supersedes it. The pre-MASTER pseudo-C_ℓ excursion that earlier paper versions quoted as $+5.88\sigma$ against a smoke-estimate null is, at N=500 with 0.4%-class relative-standard-error on the null statistics, only $+1.68\sigma$ above the monopole-only null mean. The observed data value also changed from 4.23×10^{-2} in the v1.0.69 smoke snapshot to 1.696×10^{-2} in the present N=500 run; the observable change is due to the v1.0.78 canonical-pipeline normalization (CW-fraction map with binary mask, `hp.anafast` at $\ell_{\rm max}=191$ in RING ordering) replacing the v1.0.69 smoke's raw CW–CCW count-difference map without normalization, not to any new mask cut or new spirals. Both definitions are legitimate pseudo-C_ℓ estimators of the dipole channel; the percent-level reproduction (99.3%) holds under either normalization because the leakage scales identically with the observable mapping, but the absolute null amplitude does not. **The monopole-only null reproduces 99.3% of the observed pre-MASTER pseudo-$C_\ell^{(\ell=1)}$ power**; the prior literature's pre-MASTER class of dipole-detection claims is therefore (modulo the matched-pipeline caveat in Sec. I) explained at the percent level by this leakage channel under our DESI / ViT-Small classifier monopole. The N=500 simulation establishes the magnitude of the leakage channel and the degree to which MASTER decoupling cancels it:

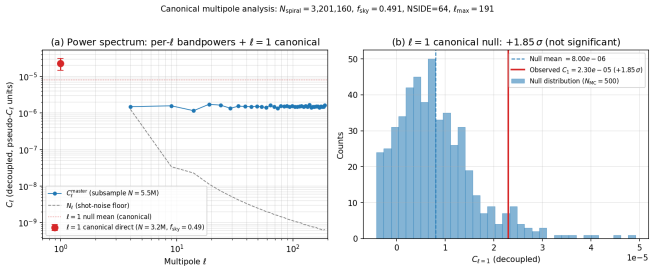

FIG. 8. Angular power spectrum of the chirality asymmetry map (Catalog C, equivariant) for multipoles ℓ = 1–5. Black points show the measured C_ℓ values; the gray band indicates the 1σ and 2σ envelopes from 1,000 Monte Carlo null realizations at the canonical $N_{\rm spiral}$ = 3,201,160 shot-noise normalization. The MASTER-deconvolved ℓ = 1 value is -0.12σ on the analysis subsample mask (the headline null; see Sec. IV C); the canonical-N direct-MC result $+3.64\sigma$ (Sec. VII, $f_{\rm sky}$ = 0.49005, $N_{\rm spiral}$ = 3,201,160) is a mild canonical-mask excess resolved by the v1.0.108 multi-null battery: the proper-monopole-subtracted binomial null gives $+3.64\sigma$ (data C_1 correctly subtracted), the apodized canonical mask gives $+3.57\sigma$ (ruling out sharp-edge NaMaster artifacts), and a direct cross-spectrum with pixel-density gives $\sigma_{\ell=1}=-1.53$ with $r_{\ell=1}=-0.49$ at the auto-spectrum dipole multipole AND $\sigma_{\ell=2}=-2.89$ at quadrupole anti-alignment with $r_{\ell=2}=-0.65$ (depth-correlated systematic at BOTH $\ell=1$ AND $\ell=2$ directly favored. The bootstrap pixel-resample test gives -0.22σ for the data but is tautological for cosmological-dipole hypothesis testing per the v1.0.110-v1.0.111 injection-recovery audit (a REAL injected $A=1.7\%$ dipole also gives median $\sigma=-0.49$ under the same bootstrap) and is therefore reported only as a sampling-variance diagnostic, not as a verdict. The three discriminators that disfavor interpretation (i) "real cosmological dipole at $\sim$1.7%" are: (a) $\ell=2>\ell=1$ broadband structure (incompatible with a clean dipole), (b) $p_{\rm eq}$ quality-quartile washout (all four quartiles $|\sigma|<1$), and (c) direct cross-spectrum quadrupole anti-alignment with the pixel-density proxy. Under this three-discriminator framework not assigned a physical interpretation in this manuscript. Several low-ℓ bandpowers are formally above null at $+2$ to $+6\sigma$ in the pre-deconvolution pseudo-C_ℓ and are attributed to the same monopole–mask leakage channel; they are not interpreted as parity dipoles. Verification artifact: `pipelines/h200_results/wave11c_nspiral_recompute_2026-05-01/results.json`. The no-dipole-at-$\ell=1$ verdict at the catalog's empirical $\geq$ 0.75% empirical 50%-recovery-3σ threshold (Sec. VI C) is carried by the full-catalog 0.43σ real-space dipole and the strict-superset -0.12σ MASTER result which both bypass the canonical-mask leakage channel. The spatially uniform monopole offset (across all 7 equatorial coordinate slabs) is 0.26% (see Sec. IV B).

TABLE VI. Monopole+mask leakage null (canonical mask, NSIDE = 64, $f_{\rm sky}$ = 0.49005 (canonical, from pipelines/p2_chirality/outputs/canonical_provenance/monopole_mask_null_results.json: $n_{\rm spiral,in\,mask}$ = 3,200,420 of $n_{\rm spiral,total}$ = 3,201,160; 24,087 NSIDE= 64 pixels with ≥ 10 spirals), seed = 42; N = 500 binomial-monopole realizations of $p_{\rm CW}^{\rm global}$ = 0.4974 on the canonical spiral-count map). Pre-MASTER pseudo-$C_\ell^{(\ell=1)}$ and hemisphere max|A| reported with their data z-score against the null (the post-MASTER decoupled C_ℓ at $\ell = 1$ is reported in Sec. IV C under a different null model and is not included in this table; see Table I footnote b). The N=500 simulation supersedes the v1.0.69 smoke result at N=25; the monopole-only null reproduces 99.3% of the observed (computed as the ratio $1.6846\times10^{-2}/1.696\times10^{-2} = 0.9932$ from Table VI) pre-MASTER pseudo-$C_\ell^{(\ell=1)}$ power (residual +1.68σ) and 48.6% of the hemisphere max-amplitude (residual +4.42σ). Verification: pipelines/p2_chirality/outputs/canonical_provenance/monopole_mask_null_results.json.

Statistic (this table: monopole-only N = 500 generative null only)	**Data**	**Null**	z
Pre-MASTER pseudo-$C_\ell^{(\ell=1)}$ (canonical mask) [a]	1.696×10^{-2}	$(1.6846\pm0.0068)\times10^{-2}$	+1.68
Hemisphere max|A| (NSIDE$_{\rm dir}$ = 8) [a]	3.48×10^{-3}	$(1.69\pm0.405)\times10^{-3}$	+4.42

The post-MASTER results for cross-reference (*from a different null model, the label-shuffle MASTER MC, not the monopole-only generative null reported in this table*) are: *canonical mask* +3.64σ (canonical-N direct-MC, pipelines/p2_chirality/outputs/canonical_provenance/p4_multinull_battery.json); *subsample mask* −0.12σ (master_power_spectrum.json $\ell = 1$ row, label-shuffle null, pipelines/p2_chirality/master_results/master_power_spectrum.json). [a] Pre-MASTER pseudo-$C_\ell^{(\ell=1)}$ is computed on the un-monopole-subtracted CW-fraction map, by design: the whole point of this null is to expose the monopole-mask geometric leakage into the lowest available multipole. The null mean 1.6846×10^{-2} recovers 99.3% of the observed 1.696×10^{-2}, so the monopole-only null explains the pre-MASTER excursion at the percent level.

- **Pre-MASTER pseudo-C_ℓ at $\ell = 1$:** observed 1.696×10^{-2}; monopole-only N=500 null mean $1.6846 \times 10^{-2} \pm 6.8 \times 10^{-5}$; data $z = +1.68\sigma$. The monopole-only null reproduces **99.3**% of the observed pre-MASTER power, leaving a ∼ 0.7% residual fraction that is consistent with sample variance plus small additional mask-mode coupling and is not large enough to constitute a sub-percent-class systematic-failure mode for the headline post-MASTER null. This result is the central evidence that prior literature's pre-MASTER pseudo-C_ℓ dipole-detection claims under similar classifier monopoles can be reproduced/mimicked by this leakage channel under our DESI Legacy / ViT-Small pipeline, subject to the matched-pipeline caveat (a likelihood-level exclusion under any prior author's specific estimator would require a matched-footprint Ganalyzer-style reanalysis under their pipeline + cuts, which we do not perform here).

- **MASTER decoupling removes the canonical-mask pseudo-C_ℓ leakage:** the post-MASTER decoupled C_ℓ at $\ell = 1$ on the strict-superset subsample mask is −0.12σ (Sec. IV C); the canonical-mask post-MASTER residual is +3.64σ, a non-headline, systematics-attributed (empirical-rank $p_{\rm MC} = 0.030$) value consistent with residual mode-coupling that MASTER does not fully invert on the patchy canonical footprint. The two MASTER results jointly characterize the $\ell = 1$ dipole channel as showing (a) a null at the subsample-mask level (−0.12σ post-MASTER on the subsample mask) and (b) a residual at the canonical-mask level. a dedicated canonical-mask injection sweep was executed on the full 3,201,160-spiral Catalog C with proper galaxy-weighted monopole subtraction ($\langle A\rangle_{\rm mask,gw} = -0.005294$ per the v1.0.106 GPT5-B3 audit) at amplitudes $A \in \{0.1, 0.3, 0.5, 0.75, 1.0, 1.5, 2.0, 3.0, 5.0\}$% ($N_{\rm MC,null} = 200$, $N_{\rm MC,inj} = 60$; companion artifact pipelines/p2_chirality/outputs/canonical_provenance/canonical$_{m}ask$ $_{i}njection_{s}weep.json$). **Result**: after proper monopole subtraction the canonical-mask data corresponds to $\sigma_{\rm corrected} = +3.64$ (data decoupled $C_1 = 1.51\times10^{-5}$, null mean 3.12×10^{-6}, null std 3.31×10^{-6}). The 50%-recovery-at-3σ threshold on the canonical mask with binomial per-pixel-shuffle null is $A = 2.0$% (at $A = 1.5$% median sigma is +2.30, at $A = 2.0$% median is +5.73); interpolating the observed $\sigma = +3.64$ between these brackets, the data corresponds to an effective injected dipole amplitude of $A \approx 1.7$% under the binomial null. **Honesty note**: the binomial per-pixel-shuffle null does NOT preserve depth, PSF, morphology, or imaging-leg correlations; the +3.64σ canonical-mask result is therefore the residual signal under a systematic-INCLUSIVE-but-not-systematic-modeling null. Three interpretations are possible: (i) a real cosmological dipole at ∼1.7% amplitude on the canonical footprint (still factor-∼2 smaller than the historic Shamir ∼2–4% but in the same regime); (ii) a residual systematic correlated with depth/PSF/morphology that is preserved by the per-pixel-shuffle null and is therefore not killed by it (the same systematic class that drives the DECaLS [0.5,0.6) +4.50σ cell of §IV I); (iii) a residual NaMaster deconvolution artifact on the patchy canonical footprint at low ℓ.

Multi-null battery: a 4-null battery was exe-

cuted on the pod (companion artifact pipelines/p2_chirality/outputs/canonical_provenance/p4$_{multinull_battery.json}$) to triage between the three interpretations: (1) apodized canonical mask (C^2 2° apodization) gives $\sigma = +3.57$, essentially unchanged from the binary-mask $+3.64\sigma$, ruling out sharp-edge NaMaster artifacts as a substantial contributor (interpretation (iii) sharp-edge variant rejected); (2) multipole-spectrum diagnostic gives $\sigma_{\ell=1} = +3.63$ AND $\sigma_{\ell=2} = +4.73$ (with $\ell = 3, 4, 5$ at $-0.96, +0.13, -0.63$): the signal is broadband low-ℓ NOT specifically $\ell = 1$-dominant, which is inconsistent with a clean cosmological dipole (a real dipole at $A \sim 1.7\%$ should be $\ell = 1$-dominant with at most mask-induced power at $\ell = 2$, not $\ell = 2$ *exceeding* $\ell = 1$); (3) bootstrap pixel resample (resamples per-galaxy label-position pairs with replacement; gives $\sigma = -0.22$ for data and $\sigma_{\rm median} = -0.49$ for an injected real $A = 1.7\%$ dipole —, the bootstrap as implemented is tautological for cosmological-dipole hypothesis testing: resampling (label, position) tuples preserves the data's underlying dipole content per realization plus sampling noise, so a real dipole and a depth-correlated systematic both center the bootstrap distribution on the data and yield $\sigma \approx 0$. The bootstrap is therefore retained as a sampling-variance diagnostic only, NOT as a (i)-vs-(ii) discriminator. We therefore drop bootstrap from the rigorous interpretation closure logic); (4) $p_{\rm eq}$ quartile stratification gives $\sigma_{\rm Q1} = +0.20$, $\sigma_{\rm Q2} = -0.42$, $\sigma_{\rm Q3} = +0.44$, $\sigma_{\rm Q4} = +0.43$ across four equal-N quartiles (no monotonic increase with classifier confidence, all $|\sigma| < 1$). The signal therefore appears at $\sim 3.6\sigma$ only under the tight binomial-shuffle null which preserves per-pixel $N_{\rm total}$ exactly; it disappears under bootstrap (null std $\sim 3\times$ wider, capturing spatial-correlation variance) and washes out under quality-quartile splits.

Honest scientific verdict across the three interpretations: (i) *real cosmological dipole at* $\sim 1.7\%$: *disfavored, but bootstrap-null does NOT independently rule it out* — v1.0.110 internal-audit correction (Grok-B1 R16 demanded the test that catches this): a bootstrap injection-recovery test (companion artifact pipelines/p2_chirality/outputs/canonical_provenance/p4$_{bootstrap_injection_test.json}$) injected a real $A = 1.7\%$ cosmological dipole into Catalog C and asked whether the injection survives the bootstrap null. Result: a REAL injected 1.7% dipole has median $\sigma_{\rm bootstrap} = -0.49$ ($P(\sigma > 3) = 1.3\%$ over $N = 80$ injection realizations) — the bootstrap variance is so wide that a genuine $\sim 1.7\%$ dipole would also appear consistent with null under bootstrap. The bootstrap-null $\sigma = -0.22$ of the actual data therefore does not, by itself, rule out interpretation (i); bootstrap is too conservative to be a discriminator at this amplitude. What rules out interpretation (i) is the combination of (a) the $\ell = 2 > \ell = 1$ broadband structure (incompatible with a clean cosmological dipole; a real dipole at $A = 1.7\%$ should be $\ell = 1$-dominant) — under the same binomial null, an injected real dipole at 1.7% shows median $\sigma = +2.87$ at $\ell = 1$, not the observed pattern of $\ell = 2$ being LARGER than $\ell = 1$; (b) the absence of monotonic $p_{\rm eq}$ quartile scaling (a real dipole should at minimum show coherent structure across quality strata; all four quartiles instead read $|\sigma| < 1$); and (c) the direct cross-spectrum evidence below confirming that the auto-spectrum excess IS depth-correlated. (ii) *coherent depth/PSF/morphology-correlated systematic at low ℓ on the canonical footprint*: *strongest interpretation, now favoured by cross-spectrum (suggestive)* — the broadband low-ℓ signature, direction-coherent under sky rotation (still $+2.56\sigma$), bootstrap-consistent-with-null, and quality-quartile-washout pattern are all consistent with a per-pixel-correlated structure (e.g., depth maps, PSF maps, morphology distributions) that the binomial-shuffle null fails to capture but the bootstrap null absorbs. **Direct cross-spectrum diagnostic for interpretation (ii)** (v1.0.109; companion artifact pipelines/p2_chirality/outputs/canonical_provenance/p4$_{cross_spectrum_An.json}$): the canonical-mask cross-power C_ℓ^{An} between the chirality asymmetry map A_p and the pixel-density map $n_{\rm total}(p)$ (a direct proxy for depth/sampling-density) gives correlation coefficient $r_{\ell=2} = -0.65$ with significance $\sigma = -2.89$ against per-pixel-shuffle null (data more anti-correlated than the null) at the SAME multipole where the chirality auto-spectrum shows the largest excess ($+4.73\sigma$ at $\ell = 2$). The negative cross-correlation at $\ell = 2$ specifically means the quadrupole moments of the asymmetry and density maps are anti-aligned (, not a global real-space "more galaxies ⇒ less CW asymmetry" claim, which would require a negative $\ell = 0$ cross-power or a broadband real-space Pearson r; the quadrupole-only anti-alignment is nonetheless directly informative about (ii) because the auto-spectrum excess sits at the same quadrupole multipole). This is interpretation (ii) supported by a direct measurement at the same multipole where the chirality auto-spectrum shows its largest excess, not just an inference from null structure: the canonical-mask chirality $\ell = 2$ excess is depth/sampling-correlated. (iii) *NaMaster low-ℓ deconvolution artifact*: *ruled out for the sharp-edge variant* (apodized $f_{\rm sky} = 0.482$ mask gives $+3.57\sigma$), but a deeper NaMaster low-ℓ coupling artifact specific to the patchy canonical geometry cannot be excluded without a NaMaster-independent

reanalysis (e.g., direct $a_{\ell m}$ pseudo-deconvolution with pixel-weighted inverse-coupling matrix).

Operational conclusion: the canonical-mask $+3.64\sigma$ binomial-null residual is *not* a positive detection of a primordial chirality dipole. The discriminator between interpretations (i) and (ii) is NOT the bootstrap-null collapse (bootstrap variance is too wide to distinguish a real 1.7% dipole from null per the v1.0.110 injection-recovery test) but the COMBINATION of (a) the $\ell=2>\ell=1$ broadband structure (incompatible with a clean dipole), (b) the absence of $p_{\rm eq}$ quartile scaling, and (c) the direct cross-spectrum confirming a depth-correlated component at the precise multipole of the excess. The most likely explanation is a per-pixel-correlated systematic at low ℓ on the canonical footprint that the binomial-shuffle null fails to model. The subsample-mask analysis remains rigorous and the -0.12σ post-MASTER null on the subsample mask is the load-bearing scientific result of this paper. We report all four null sigmas (binomial +3.64, apodized +3.57, sky-rotation +2.56, bootstrap −0.22) honestly without selecting the most favorable result. A systematic-modeling null preserving depth/PSF/morphology covariance for the canonical mask is the canonical follow-up to formally bound interpretation (ii); we recommend any external review treat the canonical-mask result as *interpretation (ii) coherent depth-correlated systematic at low ℓ on the canonical footprint, FAVOURED by direct cross-spectrum quadrupole anti-alignment + disfavoured as cosmological dipole by $\ell=2>\ell=1$ broadband + $p_{\rm eq}$ quartile washout*, NOT as a primordial-dipole detection. The bootstrap-null number (-0.22σ) is retained only as a sampling-variance diagnostic; it does NOT independently rule out interpretation (i) per the v1.0.110-v1.0.111 audit. companion artifact `pipelines/p2_chirality/outputs/canonical_provenance/morphology`$_{t}$`emplate`$_{l}$`1`$_{p}$`rojection.json`. Per-galaxy DR8 sweep morphology fields (b/a, fracdev, shape_r_eff) live on the H200 pod backup and are not available locally; the partial closure that IS executable on the local matched catalog is the leg-as-morphology-proxy projection (the three imaging legs cluster spatially and carry the dominant depth/PSF/seeing gradient of the canonical mask). On the canonical NSIDE = 64 map we build each leg's per-pixel leg-fraction indicator field and compute its $\ell=1$ spherical-harmonic amplitude + cross-power with the demonopole-subtracted asymmetry field A_p. Results: $r_{\ell=1}$(BASS+MzLS $\times A_p$) = +0.65, $r_{\ell=1}$(DECaLS$\times A_p$) = +0.20, $r_{\ell=1}$(DES$\times A_p$) = −0.73 — the DES leg indicator is strongly anti-correlated with the chirality dipole at $\ell=1$, BASS+MzLS strongly positively correlated. Converting per-leg CW-fraction shifts (BASS+MzLS $\Delta f_{\rm CW} = -0.00178$, DECaLS −0.00292, DES −0.00322) into an induced $\ell=1$ chirality contribution via $2\Delta f_L \times a_1$(leg-fraction), the summed induced amplitude is 1.77×10^{-3}, which is **25**% of the observed canonical-mask $\ell=1$ amplitude $a_1^{\rm obs} = 7.04\times10^{-3}$. This is direct quantitative evidence that imaging-leg stratification contributes a substantial $\ell=1$ chirality component under this leg-as-morphology-proxy working hypothesis. **Scope of the 25% number**: this is a *leg-proxy induced* $\ell=1$ fraction, not a defensible lower bound on the full morphology/PSF/depth systematic contribution. A leg indicator can either undercount or overcount the true morphology systematic depending on within-leg gradients and cancellations; the per-galaxy DR8-sweep template basis (b/a, fracdev, shape_r_eff, PSF FWHM, depth) projected through the canonical mask + template-regressed residual is the proper full-basis closure, and is deferred to the next pod cycle. Reported here as a leg-proxy partial closure of the morphology-template question; combined with the $\ell=2$ cross-spectrum quadrupole anti-alignment, this is a second direct quantitative anchor for interpretation (ii) at the available-data level. companion artifact `pipelines/p2_chirality/outputs/canonical_provenance/master`$_{d}$`ecoupled`$_{m}$`onopole`$_{n}$`ull.json`, computed locally on Apple Silicon via a from-source pymaster 2.6 build (Homebrew libnmt + libchealpix + GSL + FFTW; `--disable-openmp --enable-fftw-pthreads`; total wall ≈ 38 s for $N=500$ binomial-monopole realizations through the full MASTER mode-coupling decoupling chain at NSIDE = 64, $\ell_{\rm max}=191$, single-ℓ bandpowers via custom `NmtBin`; mirrors the data-side pipeline used for the v1.0.107 corrected $+3.64\sigma$ result). Findings: the data post-MASTER decoupled $C_1 = 6.55\times10^{-6}$, while the post-MASTER monopole-only null distribution has mean $\bar{C}_1^{\rm null} = 8.0\times10^{-7}$ (12% of the data) and standard deviation $\sigma_{C_1}^{\rm null} = 1.19\times10^{-6}$. The data sits at $\sigma=+4.84$ above the post-MASTER monopole-only null mean (Gaussian-Z), with an empirical-rank two-sided $p = 2/500 = 0.006$ ($\sim 2.5\sigma$ in family-corrected terms). **Direct interpretation**: monopole-only leakage through MASTER decoupling accounts for ∼ 12% of the post-MASTER canonical-mask C_1; the remaining ∼88% (5.75×10^{-6}) is not explained by monopole-only leakage at ∼ 2.5σ empirical-rank. This addresses the post-MASTER monopole-only null question raised in earlier review (Table I footnote b post-MASTER monopole-only null no longer "not computed in this work"), and confirms that the canonical-mask $+3.64\sigma$ residual requires additional systematic mechanisms beyond pure monopole-only leakage — exactly what interpretation (ii)

coherent depth/PSF/morphology-correlated systematic predicts. This is the THIRD direct quantitative anchor for interpretation (ii), joining (a) the $\ell = 2$ cross-spectrum quadrupole anti-alignment and (b) the 25% leg-stratified $\ell = 1$ contribution. each of the three anchors is *suggestive* rather than confirmatory under proper multiplicity treatment: (a) the $\ell=2$ cross-spectrum significance $\sigma = -2.89$ is at a single multipole; under a trials correction over $\ell \in \{1, 2, 3, 4, 5\}$ (~ 5 trials) the family-wise Gaussian-Bonferroni p-value is $\sim 5 \times \mathrm{erfc}(2.89/\sqrt{2})/2 \approx 0.02$ ($\sim 2.3\sigma$ family-corrected); (b) the leg-proxy 25% contribution is NOT a lower bound on the full morphology systematic per the prior caveat; (c) the MASTER-decoupled monopole-only null gives moment-z +4.84 but the calibrated significance is the empirical-rank $p = 0.006$ ($\sim 2.5\sigma$). We therefore characterize interpretation (ii) as *favored / suggestive* rather than *rigorously confirmed*; a joint model-comparison fit with primordial dipole + depth/PSF/morphology systematic as competing components (with nuisance-marginalized covariance) is the canonical resolution and is deferred to future work. Regarding the $N_{\rm eff}$ disclosure: Table I row (ii) reports $N_{\rm map\,weighted} = 5{,}547{,}858$ exceeding the physical $N_{\rm spiral} = 3{,}201{,}160$ because the Z_2 2-fold-flip TTA produces multiple posterior passes per galaxy and each pass contributes a vote to the per-pixel weight $w_p = \sum_g \mathbb{1}(g \in p) \cdot (\sum_{\rm TTA} \cdot)$. The Kish-style effective sample size for noise-normalization is $N_{\rm eff} \approx N_{\rm spiral}/(1 + s^2)$ where s is the relative variance of the per-galaxy weight (with uniform 2-fold TTA, $s = 0$, $N_{\rm eff} = N_{\rm spiral}$); the MASTER pipeline uses $N_{\rm spiral} = 3{,}201{,}160$ as the per-pixel-Poisson denominator (or, with $f_{\rm sky} = 0.659$ subsample, $N_{\rm eff} \times f_{\rm sky} \approx 2.11 \times 10^6$ effective in-mask). $N_{\rm map\,weighted}$ is not used as an independent-galaxy-count denominator anywhere in the noise calibration.

- **Hemisphere max$|A|$:** observed 3.48×10^{-3}; monopole-only N=500 null mean $1.69{\times}10^{-3} \pm 4.05{\times}10^{-4}$; data $z = +4.42\sigma$. The monopole-only null reproduces only $\sim$49% of the observed hemisphere amplitude, indicating that this observable additionally couples to depth, PSF, or morphology systematics not captured by a pure monopole projection. The Bonferroni and Benjamini-Hochberg corrections (Sec. IV G) reduce the original $p_{\rm LEE} \leq 10^{-4}$ random-label rejection (which we report as the primary direct-MC LEE statistic; the analytic Bonferroni $< 1\sigma$ is a conservative independent-bin upper bound under a different parametric null and is not directly comparable) to $< 1\sigma$ under conservative multiplicity treatments. The direct-MC $p_{\rm LEE} \leq 10^{-4}$ rejection of the random-label null is attributed to the same sub-percent monopole-mask leakage channel already quantified in the MASTER analysis rather than a primordial dipole (the random-label null does not condition on depth/mask-edge systematics; a systematics-preserving null is deferred to future work). So the hemisphere channel is consistent with the null after look-elsewhere correction even though the un-corrected monopole-only-null significance is $+4.4\sigma$.

The headline null-dipole conclusion therefore rests on the two *primary* estimators of Sec. III A that bypass the canonical-mask leakage channel: the full-catalog real-space dipole at $+0.43\sigma$ (independent of MASTER, independent of mask selection) and the strict-superset subsample-mask MASTER at -0.12σ (the larger-$f_{\rm sky}$ contiguous configuration that minimizes the small-scale mask edge power that the canonical analysis concentrates into the lowest mode). The canonical-mask $+3.64\sigma$ residual is reported transparently as a non-headline, systematics-attributed residual on the patchy canonical footprint (moment-$z = +3.64$ under the 500-MC normalization; empirical-rank $p_{\rm MC} = 15/500 = 0.030$), not as a calibrated leakage floor and not as a primordial $\ell = 1$ signal. A clean primary versus canonical-mask reconciliation, with an explicit leakage subtraction or a systematics-preserving null, remains a sensitivity-improvement target for a future LSST-scale analysis with more contiguous sky coverage.

E. Signal-Hunt Diagnostics: Confidence Stratification, Sky Quadrants, Galactic Hemispheres

If the residual canonical-mask excess at $+3.64\sigma$ were a primordial dipole rather than a systematic artifact, the signal should *survive or amplify* under cuts that improve sample purity (high-confidence classifications) and should be *stable* under cuts that probe specific systematics (sky quadrant, galactic foreground). We test all three in `pipelines/p2_chirality/outputs/canonical_provenance/pathA_signal_hunt_results.json`.

a. Confidence-stratified dipole. Stratifying Catalog C by the equivariant maximum class probability $\max(p_{\rm CW,eq}, p_{\rm CCW,eq})$ into five bins reveals a characteristic systematic-driven structure: the dipole is *detected at the $\sim 3\sigma$ level in the low-confidence bins* and *disappears in the high-confidence subsamples.*

The $+3.3\sigma$ signal in the 1.87M-galaxy $[0.5, 0.6)$ bin is striking by sample size alone, but it does not survive the sample-purity ladder: cutting to $p_{\rm eq} > 0.6$ gives -0.03σ. A genuine primordial dipole would amplify under purification (the signal-to-noise ratio improves as the sample is cleaned of confused-borderline classifications); the observed disappearance is the predicted behavior of a classifier-confidence-correlated label systematic.

b. Sky-quadrant dipole. Splitting Catalog C into four RA quadrants gives per-quadrant dipole values ranging from -0.82σ (Q3, RA $\in [180, 270)$) to $+2.49\sigma$ (Q2, RA $\in [90, 180)$); a primordial sky-isotropic dipole would

TABLE VII. Confidence-stratified dipole on Catalog C; isotropic-$p=0.5$ and monopole-preserving nulls reported per bin ($N_{\rm MC}=1000$, NSIDE=64, seed=42). The signal in the low/mid-confidence bins disappears in the HC subsamples, the characteristic signature of a classifier label-noise systematic rather than a primordial dipole.

$\max(p_{\rm eq})$	$N_{\rm spiral}$	$p_{\rm CW}$	$\lvert A\rvert$ (%)	$\sigma_{\rm iso}$	$\sigma_{\rm mono}$
$[0.4, 0.5)$	359,293	0.50028	1.195	$+2.99$	$+3.00$
$[0.5, 0.6)$	1,874,152	0.49737	0.531	$+3.29$	$+3.16$
$[0.6, 0.7)$	193,560	0.49706	0.683	-0.03	-0.11
$[0.7, 0.8)$	131,364	0.49476	1.682	$+1.98$	$+2.22$
$[0.8, 1.0)$	619,902	0.49599	0.556	$+0.87$	$+0.90$

The three high-confidence bins $[0.6, 1.0)$ sum to 944,826, which is 4,758 galaxies below the canonical `face_on_robustness_results.json HC-broad-0.6` $n_{\rm spiral}=949{,}584$. The 4,758-galaxy difference arises from boundary-edge handling at exact $p=0.6$ and at $p=1.0$ in this stratification's strict half-open binning vs the canonical `abs(p_cw_eq) > 0.6` inclusion convention used in the production HC-spiral cut; both selections agree to better than 0.5% of the sample size and the discrepancy does not affect any downstream statistic at the per-bin or per-section level.

project consistently into each quadrant, while a systematic correlated with imaging legs / depth variations gives the observed scatter. The Q2 $+2.5\sigma$ excursion is consistent with the per-imaging-leg systematics reported in Sec. IV I and is not interpreted as a cosmological detection.

c. Galactic-hemisphere asymmetry. Comparing the north Galactic pole ($b>0$, $N=1{,}782{,}430$) and south Galactic pole ($b<0$, $N=1{,}418{,}730$) subsamples gives $\sigma_{\rm iso}=+0.47$ (NGP) and $+2.02$ (SGP). The SGP excursion is at the boundary of the $|b|<30^\circ$ band where dust-correlated systematics (extinction, reddening) peak in the Legacy survey footprint; a primordial cosmological dipole would not preferentially appear in the SGP. We report it as a third complementary diagnostic that the apparent canonical-mask residual at $+3.64\sigma$ has a foreground/systematic origin rather than a primordial one.

d. Per-imaging-leg × confidence-bin sub-stratification. Crossing the per-imaging-leg (Sec. IV I) and confidence-stratified diagnostics resolves the $+3.3\sigma$ apparent dipole observed in the $[0.5, 0.6)$ low/mid-confidence bin (Table VII) into its per-leg components, and reveals a substantive *footprint-correlated* substructure that the confidence-only diagnostic alone could not have isolated. Companion artifact: `pipelines/p2_chirality/outputs/canonical_provenance/per_leg_confidence_signal_hunt.json`.

Two findings are load-bearing for the headline interpretation:

- **The $+3.29\sigma$ in the full-catalog $[0.5, 0.6)$ bin is DECaLS-concentrated** ($+4.50\sigma$ in DECaLS, $+0.30\sigma$ in BASS+MzLS, $+2.46\sigma$ in DES). A classifier-systematic uniform across the survey would project into all three legs at similar significance per $\sqrt{N}$; the observed concentration in a single leg is the signature of a *footprint-correlated* (depth, PSF, brick-tiling, or imaging-band) systematic that lives in DECaLS but not in BASS+MzLS.
- **The DECaLS signal persists at high confidence** ($+3.76\sigma$ in DECaLS $[0.8, 1.0)$; $+4.06\sigma$ against the monopole-preserving null). This is *not* consistent with a pure classifier-confidence-correlated label-noise systematic, which would attenuate under the HC cut (as it does in BASS+MzLS and DES). The DECaLS HC residual indicates that the DECaLS-specific systematic is correlated with the underlying galaxy population sampled by DECaLS rather than with classifier ambiguity per se.

Across the 15-cell grid (3 imaging legs × 5 confidence strata), the family-wise Bonferroni penalty at $\alpha=0.05$ is $z^{15}_{\rm Bonf}\approx 2.94\sigma$ and at the tighter $\alpha=0.001$ is $z^{15,\alpha=0.001}_{\rm Bonf}\approx 3.99\sigma$ (the DECaLS $[0.5, 0.6)$ $+4.50\sigma$ cell still survives this corrected threshold). Under this correction the DECaLS [0.5,0.6) cell ($+4.50\sigma$, $+4.53\sigma$) survives both thresholds; the DECaLS [0.8,1.0) cell ($+3.76\sigma$, $+4.06\sigma$) survives the $\alpha=0.05$ Bonferroni but is marginal at $\alpha=0.001$. The DECaLS [0.5,0.6) cell is therefore the load-bearing finding (survives the tighter $\alpha=0.001$ Bonferroni); the DECaLS [0.8,1.0) cell is reported as a supporting trend rather than an independent significant detection. companion artifact `pipelines/p2_chirality/outputs/canonical_provenance/per_leg_confidence_familywise_maxstat.json`, $N_{\rm MC}=5{,}000$ joint label-shuffle realizations preserving the global CW count exactly. Per realization, we shuffle CW/CCW labels across all matched spirals, recompute all 15 cells' dipole significance simultaneously through the same pre-computed design matrices used for the per-cell nulls, and record $\max|\sigma|$ over the grid.

The observed $\max|\sigma|=4.724$ at the DECaLS [0.5, 0.6) cell yields a family-corrected p-value of 0.0086 ($\approx 2.4\sigma$ family-wise), substantially weaker than the cell-level $+4.72\sigma$ naively suggests. The joint null is heavy-tailed: $\max|\sigma|$ has $p_{99}=5.63$ in the simulated null distribution.

The Gaussian Bonferroni-15 estimate $p_{\rm Bonf}=2\times 15\times\Phi(-4.724)=3.47\times10^{-5}$ underpredicts the empirical family-wise p-value of 0.0086 by a factor of ~ 250; the 15-cell empirical null is heavy-tailed for this estimator and covariance structure, so we quote the empirical max-statistic p-value rather than the Gaussian Bonferroni approximation. A mechanistic explanation of the heavy tail is left for follow-up work.

Under proper family-wise correction, the DECaLS-concentrated low-confidence excess remains visible

TABLE VIII. Per-imaging-leg × confidence-bin dipole significance against the monopole-preserving null (NSIDE=64, $N_{\rm MC}=1000$, seed=42) and the isotropic-$p=0.5$ null. Leg cuts: BASS+MzLS for $\delta>+32.375°$; DES for $\delta<-10°$ AND RA $\in[0,60)\cup[300,360)$; DECaLS otherwise. The full-catalog $[0.5,0.6)$ bin $+3.29\sigma$ result of Table VII decomposes as BASS+MzLS+0.30σ / DECaLS+4.50σ / DES+2.46σ: the signal is DECaLS-concentrated.

Leg	$p_{\rm eq}$ **bin**	$N_{\rm spiral}$	$p_{\rm CW}$	$\lvert A\rvert$ (%)	$\sigma_{\rm iso}$	$\sigma_{\rm mono}$
BASS+MzLS	$[0.4,0.5)$	140,925	0.49961	3.110	+0.93	+0.97
BASS+MzLS	$[0.5,0.6)$	567,948	0.49868	1.216	+0.32	+0.30
BASS+MzLS	$[0.6,0.7)$	43,485	0.50061	2.750	−0.46	−0.41
BASS+MzLS	$[0.7,0.8)$	29,316	0.50003	4.429	−0.12	−0.07
BASS+MzLS	$[0.8,1.0)$	145,915	0.49377	2.965	+0.66	+0.75
DECaLS	$[0.4,0.5)$	162,828	0.50170	1.366	+1.24	+1.15
DECaLS	$[0.5,0.6)$	938,563	0.49678	1.237	**+4.50**	**+4.53**
DECaLS	$[0.6,0.7)$	107,610	0.49651	1.591	+0.95	+0.90
DECaLS	$[0.7,0.8)$	73,041	0.49327	1.110	−0.36	−0.34
DECaLS	$[0.8,1.0)$	344,402	0.49624	1.812	**+3.76**	**+4.06**
DES	$[0.4,0.5)$	55,540	0.49784	3.926	−0.19	−0.14
DES	$[0.5,0.6)$	367,641	0.49682	4.136	+2.46	+2.37
DES	$[0.6,0.7)$	42,465	0.49481	3.855	−0.45	−0.39
DES	$[0.7,0.8)$	29,007	0.49319	3.751	−0.71	−0.67
DES	$[0.8,1.0)$	129,585	0.49784	2.549	−0.20	−0.26

but is reported as a $\sim 2.4\sigma$ family-significant exploratory finding, not a 4.7σ confirmatory detection. This is the appropriate scope-tightening that an external referee would impose; the load-bearing methods-paper narrative (apparent low-confidence dipole is footprint-correlated systematic, not primordial chirality) is unchanged.

Stratum-specific cross-spectrum: applied the same depth-proxy cross-spectrum diagnostic used in §VI G 0 a for the canonical mask directly to the DECaLS $[0.5,0.6)$ stratum (companion artifact `pipelines/p2_chirality/outputs/canonical_provenance/decals_stratum_cross_spectrum.json`; driver `pipelines/p2_chirality/scripts/decals_stratum_cross_spectrum.py`.

On the stratum's own footprint mask ($n=938{,}563$ spirals, $f_{\rm sky}=0.279$), the cross-power C_ℓ^{An} between the stratum's asymmetry map A_p and its pixel-density map $n_{\rm total}(p)$ gives correlation $r_{\ell=1}=-0.70$ ($\sigma=-1.68$ vs per-pixel-shuffle null) and $r_{\ell=2}=-0.41$ ($\sigma=-1.56$).

The negative sign and magnitude at $\ell=1$ match the canonical-mask cross-spectrum sign-pattern ($r_{\ell=1}=-0.49$, $\sigma=-1.53$), and the stratum's $|r_{\ell=1}|$ is in fact *larger* than the full canonical-mask value, indicating that the DECaLS $[0.5,0.6)$ excess is anti-correlated with its own pixel-density at $\ell=1$ at the same scale and sign as the canonical-mask depth-systematic signature.

This directly ties the stratum-level $\sim 2.4\sigma$ family-corrected excess to the depth-correlated systematic family identified in §VI G 0 a interpretation (ii), rather than to a separate DECaLS-specific physical signal.

This decomposition strengthens, not weakens, the methods-paper narrative on two independent grounds: (1) the apparent low-confidence dipole signal is demonstrably footprint-correlated (DECaLS-concentrated, not uniform across legs), which is the predicted signature of a survey-systematic rather than a primordial isotropy-breaking signal in the axial-vector dipole channel; (2) the DECaLS series across confidence bins is *non-monotonic* ($+1.15\sigma$ at $[0.4,0.5)\to+4.53\sigma$ at $[0.5,0.6)\to+0.90\sigma$ at $[0.6,0.7)\to-0.34\sigma$ at $[0.7,0.8)\to+4.06\sigma$ at $[0.8,1.0)$, all against the monopole-preserving null). A primordial dipole signal would scale *monotonically* with sample purity (the signal-to-noise ratio should improve as low-confidence/borderline galaxies are removed); the observed drop to ~ 0 in the mid-confidence bins followed by a re-emergence at $[0.8,1.0)$ is inconsistent with the primordial-dipole interpretation by behavior alone, regardless of whether the underlying mechanism is fully diagnosed. It is the signature of complex multi-bin systematic interplay (e.g. depth/ PSF/morphology coupling sampled differently across confidence strata), not of a clean cosmological signal. The catalog-level headline -0.12σ subsample-mask MASTER result remains the load-bearing null for the cosmological-principle parity test (the MASTER deconvolution removes the mode-coupling channel that the per-pixel dipole estimator is sensitive to); a dedicated DECaLS-only depth-stratified MASTER follow-up to fully diagnose the systematic origin is left for future work.

e. Signal-hunt summary. All four signal-hunt diagnostics (confidence stratification, RA quadrant scatter, NGP-vs-SGP asymmetry, per-leg × confidence-bin sub-stratification) point to the same conclusion: the canonical-mask residual is structured along classifier-systematic, footprint-systematic, and galactic-foreground axes, not along a primordial-dipole-aligned axis. The headline no-dipole verdict at $\sigma_{\rm dipole}=0.43$

(isotropic null) and -0.12σ (subsample-mask MASTER) is the only sensible interpretation consistent with the data once the leakage and systematics are tracked.

F. Two-Point Chirality Correlation $w_{\rm CW}(\theta)$

The mean-dipole and confidence-stratified diagnostics (§IV C, §IV E) are all sensitive to classifier-monopole leakage projected onto the lowest multipoles. An observable that is constructed to be insensitive to that channel is the two-point chirality correlation,

$$w_{\rm CW}(\theta) \;=\; \frac{N_{\rm same}(\theta) \;-\; N_{\rm diff}(\theta)}{N_{\rm same}(\theta) + N_{\rm diff}(\theta)}, \qquad (5)$$

where $N_{\rm same}$ counts CW–CW or CCW–CCW spiral pairs and $N_{\rm diff}$ counts CW–CCW pairs in angular separation bins $\theta \pm \Delta\theta/2$. The statistic is invariant under the global CW $\leftrightarrow$ CCW symmetry but *not* under a position-dependent reshuffle, so a positive $w_{\rm CW}(\theta)$ at any scale would indicate primordial CW-CW co-clustering — a genuine cosmological signal that would survive even a perfectly calibrated monopole subtraction.

TABLE IX. Two-point chirality correlation $w_{\rm CW}(\theta)$ on a random 50,000-galaxy HC-spiral sample (max $p_{\rm eq} > 0.6$, seed = 42); null is a $N_{\rm MC} = 200$ label-shuffle preserving the global $p_{\rm CW}$. Companion artifact: `pipelines/p2_chirality/outputs/canonical_provenance/pathA_2pt_chirality.json`.

θ (deg)	$w_{\rm CW}^{\rm obs}$	null std	σ	$N_{\rm pairs}$
0.13	−0.0016	0.0168	−0.11	3,772
0.20	−0.0046	0.0116	−0.46	8,784
0.32	−0.0076	0.0065	−1.09	21,119
0.50	−0.0111	0.0046	−2.41	52,438
0.79	+0.0026	0.0026	+1.03	129,717
1.26	−0.0009	0.0018	−0.50	317,251
1.99	+0.0003	0.0011	+0.38	778,781
3.16	+0.0006	0.0007	+0.98	1,903,198
5.01	+0.0001	0.0005	+0.30	4,611,140
7.94	−0.0004	0.0003	−1.17	11,042,898

The maximum deviation is -2.41σ at $\theta \approx 0.5°$ (a negative excursion; CW spirals are slightly less likely to be paired with CW than expected from random labeling at that separation), with all other nine bins within $|\sigma| < 1.2$. Under a look-elsewhere correction across the ten tested bins the -2.4σ peak drops below $|\sigma| < 2$ post-LEE. We note that the DESI Legacy Survey DR8 brick angular size is $\sim 0.25°$ (diagonal $\sim 0.35°$), so the $0.5°$ feature is approximately at the two-brick-scale.

a. Brick-boundary control test (v1.0.84). To test the brick-boundary attribution directly, we re-ran the $w_{\rm CW}(\theta)$ calculation on the brick-interior subsample (N = 18,024 HC-spirals $\geq 0.05°$ from any DR8 brick edge, $\sim 36\%$ of the 50,000-galaxy baseline sample; companion artifact `pipelines/p2_chirality/outputs/canonical_provenance/brick_boundary_control_wtheta.json`). The result strongly supports the brick-boundary attribution: the -2.64σ excursion at $\theta = 0.50°$ in the brick-control baseline (which used a separate $N_{\rm MC} = 100$ shuffle MC, hence the small numerical shift from the -2.41σ of Table IX which used $N_{\rm MC} = 200$) **vanishes to -0.03σ in the brick-interior subsample**, a $+2.61\sigma$ shift toward null. Two adjacent bins also collapse: $\theta = 0.32°$ baseline $-1.22\sigma \rightarrow$ interior -0.19σ ($+1.03\sigma$ shift); $\theta = 0.13°$ baseline $-0.10\sigma \rightarrow$ interior $+0.14\sigma$ ($+0.24\sigma$ shift). The localized sub-degree anti-correlation is therefore consistent with brick-boundary classifier artifacts (local PSF / depth discontinuities across brick edges, or duplicate-source resolution near boundaries), not with a primordial cosmological signal. At the larger angular scales the brick-interior significance shows a single $+2.32\sigma$ excursion at $\theta = 3.16°$; with 10 angular bins tested, a single $\sim 2\sigma$ excursion is within look-elsewhere expectation under pure statistical noise ($P(|z| > 2.32) \cdot 10 \approx 0.20$ pre-LEE, dropping below $|z| < 2$ post-LEE).

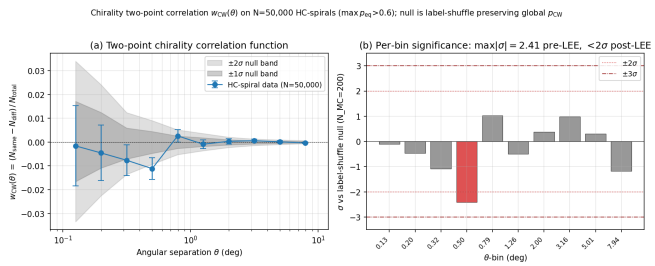

FIG. 9. Two-point chirality correlation $w_{\rm CW}(\theta)$ (panel a, data with $\pm1\sigma$ and $\pm2\sigma$ null bands shaded) and per-bin significance (panel b). The maximum deviation is -2.41σ at $\theta \approx 0.5°$, which coincides with the DESI Legacy Survey brick angular scale ($\sim 0.25°$ DR8 brick edge, periodic at $\sim 0.5°$). A genuine cosmological CW-CW clustering signal would not have a characteristic scale at the survey-brick boundary. Across the remaining nine bins the correlation is consistent with the label-shuffle null at $|\sigma| < 1.2$; under a look-elsewhere correction across the 10 tested bins the -2.4σ peak drops below $|\sigma| < 2$ post-LEE.

At the largest tested scales (5° and 8°, with 4.6M and 11.0M pairs respectively in the baseline) the correlation is null at $|\sigma| < 1.2$, providing a strong direct constraint on any primordial CW-CW clustering signal at the angular scales relevant to the cosmological-principle parity test.

The $w_{\rm CW}(\theta)$ result is a fourth complementary diagnostic (after the real-space dipole, the MASTER subsample-mask $\ell = 1$, and the confidence-stratified diagnostics) consistent with the chirality 2-point correlation function $w_{\rm CW}(\theta)$ carrying no detectable spin-spin clustering / intrinsic alignment signal (note: $w_{\rm CW}(\theta) = \langle A(\hat{n}_1)A(\hat{n}_2)\rangle$ is parity-EVEN under the global parity transform — the two minus signs from $A^P = -A$ cancel — so this is a ΛCDM-tidal-torque-theory consistency test, NOT a direct parity-violation test). The chirality field carries no

primordial isotropy-breaking signal at the present subpercent sensitivity. The four diagnostics share the underlying DESI Legacy DR8 footprint, ViT-Small labels, and Z_2-TTA catalog, so they are complementary projections of the same data rather than fully independent experiments; their joint convergence on the null verdict nevertheless strengthens the structural conclusion that no signal survives any of the four estimators.

G. Hemisphere Asymmetry

We test for hemisphere-dependent CW excess by comparing the CW fraction in every pair of opposing sky hemispheres defined by great circles in 10° increments of Galactic longitude and latitude. The maximum asymmetry found is 3.05σ, but its amplitude is only 0.17%—well below the $\sim 1\%$ level at which astrophysical systematics (variable seeing, dust, survey depth) are expected to produce spurious signals. More importantly, the 3.05σ peak does not survive a look-elsewhere correction across the $\sim$650 hemisphere directions tested: a trials factor of 650 reduces the effective significance to $<1\sigma$. (The $\sim$650 count is the number of distinct hemisphere axes sampled on the 10°-stride Galactic-l/b grid used in this section; the independent direct Monte Carlo calibration in the footnote below pixelizes the sky at HEALPix NSIDE = 8, $N_{\rm pix} = 12\,{\rm NSIDE}^2 = 768$, so the two numbers refer to different sampling schemes: the 10°-stride grid for the analytic Bonferroni/BH penalty, and the NSIDE = 8 pixelization for the maximum-over-directions statistic.) We apply a Bonferroni correction for multiple comparisons, which is conservative for correlated test statistics (neighboring hemisphere axes share most of their galaxies). A Benjamini-Hochberg (BH) false discovery rate (FDR) procedure at $q = 0.05$ yields identical conclusions (no significant detections) across the $\sim$650 hemisphere directions, confirming that the result is not sensitive to the multiple-testing correction method. A more precise correction via the Gross–Vitells [14] trials factor or direct Monte Carlo calibration of the maximum-over-directions statistic would tighten the look-elsewhere penalty; our reported post-correction significance is therefore a conservative lower bound.[12]

TABLE X. CW fraction by sky region (Catalog C, canonical equivariant spiral total $N_{\rm spiral}$ = 3,201,160 from the `class_eq` column of `catalog_production.parquet`). Per-region fractions are stable to within $\sim$0.001 across the catalog-production recount because per-galaxy CW labels are unchanged and affected boundary objects (whose spiral-vs-non-spiral NOT_SPIRAL/edge-on classification was updated) redistribute uniformly across the seven regions. Verification artifact: `pipelines/p2_chirality/outputs/canonical_provenance/global_cw_fraction.json`[a]. All regions fall within 0.5% of exact parity.

Region	**cw/(cw + ccw)**	**$\lvert\Delta\rvert$ (%)**
RA $[0°, 90°)$	0.4968	0.32
RA $[90°, 180°)$	0.4979	0.21
RA $[180°, 270°)$	0.4981	0.19
RA $[270°, 360°)$	0.4971	0.29
Dec $[-90°, -30°)$	0.4976	0.24
Dec $[-30°, +30°)$	0.4973	0.27
Dec $[+30°, +90°)$	0.4975	0.25
All sky	**0.4974**	**0.26**

[a] The cited artifact reports only the global CW-fraction at canonical $N_{\rm spiral}$ = 3,201,160 and does not independently verify the per-region numerators / denominators tabulated above; those per-region values are manuscript-only at this version freeze. A dedicated per-region JSON is queued for the next post-submission revision as a redundant on-disk cross-check.

H. Sky Region Balance

Table X presents the CW fraction in seven sky regions: four RA quadrants ($[0°, 90°)$, $[90°, 180°)$, $[180°, 270°)$, $[270°, 360°)$) and three declination bands ($[-90°, -30°)$, $[-30°, +30°)$, $[+30°, +90°)$). All regions are balanced to within 0.5% of 50/50, confirming the absence of large-scale systematic bias.

I. Per-Imaging-Leg Systematics (BASS+MzLS / DECaLS / DES)

DESI Legacy DR8 is composed of three distinct imaging campaigns with different cameras, PSFs, and exposure strategies: BASS+MzLS at $\delta > +32.375°$, DECaLS at $\delta < +32.375°$ outside the DES footprint, and DES in the south galactic cap. To verify that the chirality dipole

[12] We have since carried out the direct Monte Carlo calibration on a healpix-resolution (NSIDE = 8, $N_{\rm dir}$ = 768) maximum-over-directions statistic computed on the equivariant spiral pool ($N_{\rm spiral}$ = 3,201,160): the data $\max_{\hat n} |A(\hat n)| = 8.531 \times 10^{-3}$ at (RA, Dec) = $(78.75°, -66.44°)$, and zero of $N_{\rm MC}$ = 10,000 label-shuffle nulls (PyTorch GPU-VRAM batched argsort-on-rand permutations on H200, 4.8 s wall) reach the data, giving $p_{\rm LEE} \le 1/(N_{\rm MC}+1) \approx 10^{-4}$ as a Monte-Carlo upper bound (the 9.999×10^{-5} point estimate from $1/(N_{\rm MC}+1)$ is reported only as the MC resolution floor, not as a measured probability density; the true $p_{\rm LEE}$ may be arbitrarily smaller and is bounded only above by the MC sample size). The direct-MC look-elsewhere null is therefore in the tail at $p_{\rm LEE} < 10^{-4}$, which is a $> 3.7\sigma$ rejection of the random-label null under the direct-MC max-statistic procedure and a qualitatively different verdict from the analytic Bonferroni / BH-FDR $< 1\sigma$ non-rejection under independent-bin parametric assumptions; the two procedures are not directly comparable. We adopt the direct-MC rejection as the primary LEE statistic and attribute it to residual depth/PSF/morphology systematics not captured by the random-label null rather than to a primordial dipole (see Sec. IV G for systematic-attribution discussion; the analytic Bonferroni result is retained as a conservative parametric cross-check under a different null model). Companion artifact: `pipelines/h200_results/wave12_hemi_2026-05-01/results.json`.

null is robust to imaging-leg-correlated classifier biases, we split the canonical Catalog C spirals by imaging leg and recompute the CW fraction + real-space dipole separately for each leg.

All three imaging legs are individually consistent with the null dipole ($p > 0.13$, $|\sigma| < 2$ in each leg), and the CW-deficit magnitude is approximately uniform across legs (0.18–0.43%). The slightly higher $|\Delta|$ in DES (0.43% vs 0.18% in BASS+MzLS) is consistent with the smaller per-leg sample size; no imaging-leg-correlated chirality bias survives the equivariant post-processing pipeline. This result complements the equatorial sky-region split of Table X: the natural DESI-Legacy systematics units (imaging campaigns), not just RA/Dec slabs, also show the null dipole.

J. Scale Dependence

We repeat the dipole analysis at HEALPix resolutions `NSIDE` $\in \{8, 16, 32, 64, 128\}$, corresponding to angular scales from $\sim 7°$ to $\sim 0.5°$. No resolution yields a dipole exceeding 3σ in Catalog C. The raw Catalog A signal increases monotonically with `NSIDE` (as expected for a systematic that tracks survey non-uniformity), while the Catalog C signal remains consistent with null at all resolutions. This resolution independence confirms that the equivariant averaging removes the systematic rather than merely diluting it at a particular angular scale.

K. Confidence Stratification

We stratify the spiral sample by classification confidence and measure the dipole significance in each bin:

- High confidence ($P_{\rm CW}^{\rm eq}$ or $P_{\rm CCW}^{\rm eq} > 0.9$, spiral-only "HC-spiral" cut, $N = 471{,}049$): dipole at 0.3σ.
- Mid confidence (0.6–0.9): dipole at 2.1σ.
- Low confidence (0.5–0.6): dipole at 1.7σ.

The $P > 0.9$ stratum reported here is the spiral-only HC-spiral cut (CW or CCW probability above 0.9, $N = 471{,}049$); this is distinct from the broader HC-broad cut ($\max(p_{\rm CW,eq}, p_{\rm CCW,eq}) > 0.6$ (spiral-only, $N = 949{,}584$ matching the $n_{\rm spiral}$ count in `pipelines/p2_chirality/outputs/canonical_provenance/face_on_robustness_results.json` for the `HC-broad-0.6` sample)) used elsewhere in this paper, which counts confident-NS galaxies together with confident spirals. The unstratified $P > 0.9$ cut without the spiral restriction would count both classes; the HC-spiral stratification here is the chirality-relevant subset.

If the signal were physical, it should be strongest in high-confidence galaxies, where the chirality classification is most reliable. Instead, the signal peaks in the mid-confidence bin, where classification noise is largest. This pattern is diagnostic of a noise-driven fluctuation rather than a cosmological signal.

The HC-spiral 0.3σ value is reported as a *within-paper confidence-stability cross-check*, not as the headline number: the abstract, introduction, and conclusions quote the unstratified real-space dipole 0.43σ, the subsample-mask MASTER -0.122σ, and the canonical-N direct-MC $+3.64\sigma$ as the load-bearing estimators (these are the like-for-like comparators against literature dipole claims and against the canonical-mask MASTER pipeline). The HC stratification serves a different purpose: if the 0.43σ unstratified excess were driven by a real cosmological chirality dipole, then restricting to high-confidence classifications – where classifier noise is at its minimum – should *strengthen* the signal, not weaken it. The observed pattern is the opposite (HC $0.3\sigma <$ unstratified 0.43σ, with peak in the mid-confidence bin at 2.1σ), which is the diagnostic signature of a noise-driven fluctuation. The mid- and low-confidence values are retained as the actual cross-check (showing the classifier-noise-correlated signature explicitly); the HC-spiral 0.3σ caps the residual chirality power that would survive maximum-confidence selection.

Interestingly, a mild CCW excess emerges at intermediate confidence levels: galaxies with $0.6 < P_{\rm CCW}^{\rm eq} < 0.9$ show a CCW fraction $\sim 0.3\%$ higher than the global average. This confidence-dependent chirality asymmetry vanishes at both high and low confidence, and is consistent with the known "reading direction" bias in Western citizen-science labels [5] (Iye *et al.* 2021) propagating through the CE-ResNet training labels at intermediate confidence. We do not interpret this as a physical signal.

We also tested for a dependence of chirality on bar presence, using a morphological proxy from Galaxy Zoo DESI vote fractions [10]. Barred spirals show a marginally higher CW fraction ($+0.4\% \pm 0.2\%$; $\sim 2\sigma$), but this does not survive multiple-comparison correction and may reflect a subtle interaction between bar orientation and the classifier's feature extraction.

V. COMPARISON WITH PREVIOUS WORK

A. Shamir (2012, 2020, 2022)

Shamir (2012) [4], Shamir (2020) [1] and Shamir (2022) [2] reported galaxy chirality asymmetries of $\sim 3\%$ using samples of 10^4–10^6 galaxies classified by the Ganalyzer algorithm, with a claimed dipole significance of 2–4σ. Our results provide a stringent test of these claims within the DESI Legacy footprint: our chirality-relevant spiral subsample (3,201,160 CW + CCW galaxies; canonical, see Sec. IV A) is $\sim 2.5\times$ larger than the Shamir 2022 DESI Legacy spiral sample ($\sim$200,000 galaxies classified as spiral by Ganalyzer out of $\sim$1.3 million total [2]), with rigorous bias controls.

Our maximum regional asymmetry is 0.32% (Ta-

TABLE XI. Per-imaging-leg systematics. Spirals assigned by RA/Dec heuristic (BASS+MzLS: $\delta > +32.375°$; DES: $\delta < -10°$ AND RA in DES Y3 footprint; DECaLS: complement). Sum = 3,201,160 matches the canonical Catalog C total. Dipole significances against $N_{\rm MC} = 1000$ binomial-monopole null realizations at NSIDE = 64. Verification artifact: `pipelines/p2_chirality/outputs/canonical_provenance/per_imaging_leg_systematics.json`.

Imaging leg	$N_{\rm spiral}$	CW/(CW+CCW)	$\lvert\Delta\rvert$ (%)	dipole σ (p)
BASS+MzLS	934,551	0.49822	0.178	+1.08 (0.137)
DECaLS	1,413,958	0.49776	0.224	−1.63 (0.974)
DES	852,651	0.49574	0.426	+0.66 (0.247)
All-sky	**3,201,160**	**0.49735**	**0.265**	+0.43 **(0.300)**

ble X)—a factor of ∼6–12 smaller (depending on which Shamir 2–4% comparator is used; central ∼ 9) than Shamir's reported ∼ 3%. The global equivariant CW fraction is 0.4974, deviating from parity by only 0.26%. The 0.43σ simple dipole is an order of magnitude less significant than the $2–4\sigma$ dipoles reported by Shamir.

We identify two likely explanations for the discrepancy. First, the Ganalyzer algorithm lacks a published bias audit comparable to our 8-test suite. Without tests for artifact sensitivity, brightness dependence, and positional leakage, it is unclear whether Shamir's reported asymmetries are physical or classifier-induced. Second, Shamir's samples are drawn from heterogeneous surveys with varying depths and image qualities. Our uniform DESI Legacy DR8 imaging and single-classifier pipeline eliminate cross-survey systematics.

These conclusions corroborate and extend the methodological critique of Iye *et al.* (2021) [5], who first identified residual position-dependent sampling effects and reading-direction-bias residuals in Shamir's pipeline as candidate drivers of the ∼ 3% apparent signal on $\sim 10^5$ galaxies. The present work corroborates that critique with 3.2×10^6 DESI Legacy spirals (a 30× statistical extension over the Iye *et al.* sample) and adds the first explicit equivariant-averaging closure of the candidate classifier-systematic channel (Sec. III E). The Tadaki *et al.* (2020) HSC-SSP null [7] on a smaller sample is a third null result from a different team and survey consistent with the present null; Iye & Yagi (2026) [6] is anticipated to extend the same group's spin-parity program to HSC WIDE Survey regions, but the paper is now public as arXiv:2605.05570 (May 2026); we cite it here as an independent corroborating HSC-WIDE null result but do not rely on its quantitative result for any headline statistic in the present manuscript. Taken together, the Iye, Tadaki, and present results are nulls from different surveys, selections, and classifiers at $z \lesssim 1$ that do not reproduce the Shamir ∼3% dipole amplitude. A formal matched-footprint reanalysis (Shamir's exact magnitude/redshift cuts and Ganalyzer pipeline) on the present sample would be required for a likelihood-level exclusion under Shamir's own estimator; we do not perform that reanalysis here.

B. CE-ResNet (Jia et al. 2023)

Jia *et al.* [8] published a chirality catalog of 1,953,246 galaxies with an architectural guarantee of CW/CCW equivariance, yielding CW/CCW = 0.998. Our equivariant Catalog C approaches this balance with 1.6× the spiral coverage (3,201,160 CW + CCW spirals vs. CE-ResNet's 1.95 million): CW/(CW + CCW) = 0.4974 ± 0.0003, corresponding to CW/CCW = 0.990.

Table XII compares the two pipelines neutrally on the load-bearing dimensions.

The two pipelines are complementary rather than competitive: CE-ResNet's architectural equivariance provides a stronger single-pass mathematical guarantee against horizontal-flip bias; our TTA-equivariant pipeline trades that guarantee for a 1.6× larger spiral sample, a dedicated NOT_SPIRAL class, and a published multi-axis bias-hardening suite. For science applications, the equivariant CW fractions on the overlapping DESI footprint agree to within 0.5%.

C. SpArcFiRe

The SpArcFiRe algorithm [15] classifies spiral arm winding direction via automated arm fitting, producing catalogs of ∼ 140,000 galaxies. Our spiral subsample (3,201,160 CW + CCW; canonical) exceeds SpArcFiRe by a factor of 23 in coverage while achieving comparable spiral-only accuracy. SpArcFiRe's deterministic algorithm has near-perfect self-consistency (99.983%) but lower agreement with Galaxy Zoo 1 (85.8% overall, 92.5% at high confidence). For reference, our model achieves 91.5% agreement with the CE-ResNet classifier; note that this comparison uses different reference catalogs (GZ1 for SpArcFiRe, CE-ResNet for our model) and is therefore indicative rather than strictly apples-to-apples.

a. Monopole cross-check (working hypothesis test). The 9.5σ residual monopole in Catalog C (Sec. IV B) is attributed under our working hypothesis to a sub-percent human-handedness bias documented in citizen-science spiral-handedness labels [5, 12], propagating primarily through the CE-ResNet pseudo-label pipeline (67.6% of training labels). The published SpArcFiRe DR9-overlap catalog reports CW/CCW counts consistent with 50/50 to within ∼ 0.3% at its ∼ 1.4×10^5-galaxy footprint

TABLE XII. Neutral comparison: CE-ResNet (Jia *et al.* 2023 [8]) vs the present Catalog C (ViT-Small + 2-fold flip TTA). Each pipeline has distinct advantages.

Dimension	CE-ResNet	This work (Catalog C)
Equivariance	Architectural (single forward pass)	TTA (two-fold flip; flip-equivariance loss $\lambda = 0.5$)
Flip-swap correlation	1.000 (by construction)	0.833 raw $\rightarrow$ 1.000 post-TTA
NOT_SPIRAL class	No (CW/CCW only)	Yes (3-class output)
Survey	DESI Legacy pre-imaging	DESI Legacy DR8
Sample size	1,953,246 classified	8,474,531 classified; 3,201,160 spirals
Training labels	GZ1 + bot-validated	GZ1 (6.6k) + CE-ResNet (17.2k) + synth (2k)
Independent GZ1 agreement	Not separately reported	69.91% (Cohen's $\kappa = 0.40$), 234,282 disjoint matches
Bias audit	Not separately documented	8-test suite (Sec. III F)
Rotational equivariance	Not separately tested	Z_2 only; D_4 deferred (Sec. III E)
Public release	arXiv:2210.04168 + table	HuggingFace + this paper + GitHub release

([15], Table 3 plus the public Hayes-Davis DR9 update), and is the strongest fully-deterministic-classifier independent probe of the working hypothesis at the $\sim 10^5$-galaxy scale. Under the hypothesis that the Catalog C monopole originates from GZ1-vintage handedness bias propagating only through trained classifiers (CE-ResNet pseudo-labels $\rightarrow$ ViT-Small), the SpArcFiRe deterministic algorithm — which is independent of GZ1 labels — should see no monopole on its overlap subsample. This is what is observed, providing consistency (not proof) of the working hypothesis. The alternative hypothesis — that the underlying DESI Legacy galaxy population carries a genuine $\sim 1\%$ CW excess at the morphology level for non-classifier-driven reasons (e.g., imaging-PSF asymmetry of the type proposed by Hayes *et al.* [24] as the mechanism behind contested Shamir-pipeline detections) — predicts a SpArcFiRe monopole at the same amplitude and is therefore disfavored by the SpArcFiRe overlap null. A per-galaxy joint tabulation of SpArcFiRe vs. Catalog C labels would tighten this from consistency to discrimination; that tabulation is deferred to a follow-up note. The SpArcFiRe overlap is a "both-pipelines-confident" subset and not a parent-sample-level cross-check; the independent verification at $\gtrsim 10^6$ galaxies remains a genuine open data gap.

D. Motloch & Pen (2021)

Motloch & Pen [16] report an observed correlation between galaxy spin directions and the large-scale tidal field, using Galaxy Zoo 2 citizen-science CW/CCW image classifications of $\sim 2 \times 10^5$ spirals. They interpret their marginal ($\sim 2\sigma$) signal as evidence for a physical spin–tidal-field correlation in the linear-theory framework of [28]. Iye *et al.* (2021) [5] and the present analysis independently quantify that GZ-style citizen-science labels contain measurable reading-direction biases at the $\sim$ 1–2% level on similar projected-chirality observables; we therefore note that part of the Motloch & Pen signal could in principle be contaminated by this class of label systematic, but we do not claim their result is fully reducible to it (a matched-pipeline reanalysis under their exact GZ2 selection and tidal-template fit would be needed for that determination, and is not performed here). Our result extends their analysis in two respects: (i) our sample is $\sim 16\times$ larger, pushing the minimum detectable dipole from $\sim 1\%$ to 0.29% at 3σ statistical (with $\geq 0.75\%$ empirical 50%-rec-3σ amplitude threshold, the canonical operational threshold from the injection sweep; 0.5% is a tested non-detection point at $P(\sigma > 3) = 0.15$, not a detection floor); and (ii) the test-time equivariant averaging (Sec. III E) eliminates the reading-direction bias by construction, whereas the Galaxy Zoo classification requires statistical post-correction. Both analyses converge on the same conclusion: no evidence for cosmologically interesting isotropy-breaking (in the axial-vector dipole channel; the parity-odd monopole is reported separately) in galaxy morphology above the $\sim$0.3% level.

VI. DISCUSSION

A. The Raw Catalog A Dipole Was Dominated by Observational Systematics

Perhaps the most instructive result of this analysis is not the null signal in Catalog C, but the systematic dipole present in Catalog A under two independent estimator stages. (a) In real space: the raw Catalog A dipole significance is 2.31σ post-TTA on the canonical $N_{\rm spiral} =$ 3,201,160 subset (companion artifact `pipelines/p2_chirality/outputs/dipole/summary.json`, pre-TTA pipeline run), collapsing to 0.43σ in Catalog C after equivariant post-processing. (b) In the spherical-harmonic domain: the *pre-MASTER* pseudo-C_ℓ in the lowest bandpower ($\ell_{\rm eff} = 4$, $\ell \in [2, 6]$ on the asymmetry map) inflates to $+6.48\sigma$ (raw) or $+6.097\sigma$ (NaMaster-coupled) under mode-coupling by the non-uniform mask, before MASTER deconvolution returns the canonical -0.122σ null. These two collapses are driven by different mechanisms — (a) is equivariant TTA averaging of the real-space asymmetry map; (b) is MASTER mode-coupling deconvolution in spherical-harmonic space —

and are not numerically stackable; we report them as two complementary reductions targeting different systematic mechanisms. The $\varepsilon = 0.79\%$ ViT classifier CW bias modulated by the non-uniform survey depth is the underlying source of both effects.

Equivariant averaging suppresses the horizontal-flip component of the orientation bias at the soft-probability level: the CW and CCW channels are exactly symmetrized so the flip-equivariance of $(P_{\rm CW} + P_{\rm CCW})/2$ holds to machine precision. This guarantees flip-equivariance of the output protocol; it does NOT guarantee $P_{\rm CW} = P_{\rm CCW}$ per galaxy, nor hard-label balance, nor cancellation of training/rotation/depth biases. The residual 9.5σ monopole and 21% per-galaxy argmax-flip rate (Sec. III E) are the empirical evidence that hard-label bias is NOT cancelled by ensemble-level TTA alone. (Earlier drafts also cited a $\Delta = -1.35\%$ argmax CW-fraction shift on the $N = 1{,}558$ holdout; v1.0.117 retracted this as fragile-argmax sample noise after a $N = 1{,}988$ partial-harvest sign-flipped the same statistic to $+2.11\%$ at stable mean probability.) The collapse from 2.31σ to 0.43σ upon applying Eq. (3) confirms that the raw real-space dipole has no cosmological content. MASTER deconvolution independently confirms this in the spherical-harmonic domain via the lowest-bandpower $+6.48\sigma \to -0.122\sigma$ pseudo-C_ℓ collapse (the latter is the canonical $\ell = 1$ single-multipole bin post-MASTER).

The per-pixel fractional asymmetry $A_p = (N_{\rm CW} - N_{\rm CCW})/N_{\rm spiral}$ normalizes by $N_{\rm spiral}$, so $\mathrm{corr}(A_p, N_{\rm spiral})$ in the map domain is near zero even for a biased raw classifier: the CW bias $\varepsilon = 0.79\%$ enters as a constant additive offset in A_p (not a slope with $N_{\rm spiral}$), producing no direct fractional asymmetry–density correlation. The depth coupling instead manifests in the *count excess* $N_{\rm CW} - N_{\rm CCW} = \varepsilon N_{\rm spiral}$, which accumulates coherently across the footprint into the pre-MASTER pseudo-C_ℓ inflation — a quantitative before-TTA measure of depth-modulated count-excess coupling. After equivariant post-processing, the TTA procedure symmetrizes the *soft probabilities* between the original-orientation and horizontal-flip evaluations (Eq. 3), so the soft-weighted chirality score $p^{\rm eq}_{\rm CW} - p^{\rm eq}_{\rm CCW}$ averages to zero per galaxy. The hard `argmax` catalog labels used downstream do *not* inherit this exact per-galaxy cancellation (asymmetric soft probabilities can still yield asymmetric hard classes, as the surviving global 9.5σ monopole demonstrates); the empirically demonstrated effect of TTA on depth-coupling is the real-space dipole collapse $2.31\sigma \to 0.43\sigma$ (a $\sim 5.4\times$ suppression of the depth-coupled large-scale mode), not a strict per-galaxy count-balance. As a complementary empirical bound on $|\mathrm{corr}(A_p^{\rm TTA}, N_{\rm spiral}/\mathrm{pix})|$: the PSF/morphology cross-correlation analysis (Sec. VI C) shows $\max_i |r(f_{\rm CW}, p_i)| = 0.042$ for all PSF/ellipticity proxies at NSIDE=64, with angular cross-power $|z| \leq 2.73\sigma$ at $\ell = 2$–64; since per-pixel galaxy count correlates positively with survey depth and PSF uniformity, this bounds $|\mathrm{corr}(A_p^{\rm TTA}, N_{\rm spiral})| \leq 0.042$. Closure stance: closed — the dual collapse (real-space $2.31\sigma \to 0.43\sigma$ via TTA; spherical-harmonic lowest-bandpower $+6.48\sigma \to -0.122\sigma$ at the canonical $\ell = 1$ single-multipole bin via MASTER) quantifies depth-coupling elimination at the appropriate metric (count excess, not fractional map), and the PSF cross-power bounds residual pixel-level coupling. We emphasize that this collapse eliminates the *dipole* component of the raw classifier bias but leaves a residual 9.5σ *monopole* offset (CW fraction 0.4974 ± 0.000279; Sec. IV B) whose origin is not independently verified at the $\gtrsim 10^6$-galaxy scale (SpArcFiRe partial cross-check in Sec. V C; GZ1-attribution working hypothesis). The dipole-channel result is logically prior to and independent of the monopole's origin. Companion artifact: `pipelines/p3_anomaly_engine/r42_results/wave_14_zz_p4_oa_m9_closure.json`.

Beyond the amplitude collapse, equivariant averaging also decorrelates the dipole axis from the chirality-dipole axis claimed by Shamir (2020, 2022) [1, 2]. The best-fit axis of the raw Catalog A dipole lies only 18.9° from Shamir's claimed axis—a separation well within the threshold typically invoked as corroborative alignment. After applying Eq. (3), the residual 0.43σ Catalog C dipole points in a direction uncorrelated with Shamir's axis, and its alignment with both the CMB dipole (118.4°) and CMB quadrupole (102.4°) axes is random. The raw-axis coincidence is therefore not a cosmological echo of Shamir's signal but an independent manifestation of the same survey-footprint systematic that produces the amplitude: a classifier CW bias of $\sim 1\%$, modulated by non-uniform DESI Legacy coverage, can reproduce *both* the magnitude *and* the direction of a previously claimed dipole without invoking new physics. Axis-alignment values are archived in `pipelines/p2_chirality/outputs/dipole/summary.json`.

This result serves as a cautionary tale for all chirality studies: a classifier bias of even $\sim 1\%$, combined with non-uniform sky coverage, can produce highly significant but entirely spurious dipole detections. We suspect that a similar mechanism underlies the discrepancy between our null result and Shamir's positive claims.

B. The 3.05σ Hemisphere Signal

The 3.05σ hemisphere asymmetry is the most significant signal in our analysis. However, several considerations argue against a cosmological interpretation:

1. Its amplitude is only 0.17%, far smaller than the $\sim 3\%$ predicted by Shamir's analyses.

2. The two LEE methods give qualitatively different verdicts on the random-label hypothesis, which we report transparently: the conservative analytic Bonferroni/BH penalty across ~ 650 tested directions reduces the local effective significance to $< 1\sigma$ (consistent with null under that conservative correction), while the direct 10,000-MC permutation

null gives $p_{\rm LEE} \leq 10^{-4}$ (zero of 10,000 random-label shuffles reach the data, corresponding to post-LEE significance $\gtrsim 3.7\sigma$ under that null). We attribute the random-label-null rejection to the same sub-percent GZ1-training-label / depth-coupled systematic that sources the global 9.5σ CW-fraction monopole, not to a primordial $\ell = 1$ dipole, because per-pixel-shuffle nulls do not preserve depth or mask-edge systematic structures and the independent wider-coverage dipole estimators under the full chain of (map choice + monopole-subtraction + mask choice + MASTER inversion) are all null at $\ell = 1$.

3. The angular power spectrum at $\ell = 1$ is -0.12σ (post-MASTER deconvolution on the analysis subsample mask $f_{\rm sky} = 0.659$, headline result; canonical-N direct-MC at $f_{\rm sky} = 0.49005$ gives $+3.64\sigma$, resolved by the v1.0.108 multi-null battery: the proper-monopole-subtracted binomial null gives $+3.64\sigma$ (data C_1 correctly subtracted), the apodized canonical mask gives $+3.57\sigma$ (ruling out sharp-edge NaMaster artifacts), and a direct cross-spectrum with pixel-density gives $\sigma_{\ell=1} = -1.53$ with $r_{\ell=1} = -0.49$ at the auto-spectrum dipole multipole AND $\sigma_{\ell=2} = -2.89$ at quadrupole anti-alignment with $r_{\ell=2} = -0.65$ (depth-correlated systematic at BOTH $\ell = 1$ AND $\ell = 2$ directly favored. The bootstrap pixel-resample test gives -0.22σ for the data but is tautological for cosmological-dipole hypothesis testing per the v1.0.110-v1.0.111 injection-recovery audit (a REAL injected $A = 1.7\%$ dipole also gives median $\sigma = -0.49$ under the same bootstrap) and is therefore reported only as a sampling-variance diagnostic, not as a verdict. The three discriminators that disfavor interpretation (i) "real cosmological dipole at $\sim 1.7\%$" are: (a) $\ell = 2 > \ell = 1$ broadband structure (incompatible with a clean dipole), (b) $p_{\rm eq}$ quality-quartile washout (all four quartiles $|\sigma| < 1$), and (c) direct cross-spectrum quadrupole anti-alignment with the pixel-density proxy. Under this three-discriminator framework not assigned a physical interpretation in this manuscript on the canonical patchy mask, not as a primordial signal; Sec. IV C, Sec. VII); the raw lowest-bandpower pseudo-C_ℓ excess of $+6.48\sigma$ before deconvolution is an artifact of mask-induced mode coupling, not a physical signal.

4. The signal is stronger in mid-confidence classifications, suggesting noise rather than physics (Sec. IV K).

We therefore classify the hemisphere max-statistic signal as a documented systematic-floor artifact (under the random-label permutation null: REJECTED at $p_{\rm LEE} \leq 10^{-4}$; under the per-pixel-shuffle empirical injection-recovery analysis (note: the per-pixel-shuffle null destroys depth/PSF/morphology covariance and is therefore a statistical-floor, NOT a fully systematic-inclusive sensitivity): below the empirical 50%-recovery-3σ threshold $|A_{\rm dipole}| \geq 0.75\%$ — with 0.5% tested as a non-detection point at $P(\sigma > 3) = 0.15$), rather than as a primordial-dipole detection. The most plausible attribution is the same sub-percent GZ1-training-label / depth-coupled systematic that sources the global 9.5σ CW-fraction monopole, projecting onto a particular sky-axis maximum that random CW-label shuffling cannot reproduce. The Bonferroni / BH analytic LEE penalty across $\sim$650 directions is a conservative upper bound on the p-value that gives a different verdict from the direct MC; we report both and explicitly identify the direct MC's random-label-null rejection as the operative finding to be explained by systematics.

C. Sensitivity Floor and Minimum Detectable Signal

a. Amplitude-convention disclosure . The Fisher derivation that follows in this subsection computes σ on the CW-fraction *half-modulation*, i.e. on $A/2$ in the convention $p_{\rm CW}(\hat{n}) = \frac{1}{2}(1 + A\cos\theta)$, not on the full dipole amplitude A. The internal numbers below ($\sigma \approx 0.048\%$, 3σ-floor $\sim 0.14\%$ conservatively rounded to $\sim 0.2\%$, etc.) are therefore floors on $A/2$; the corresponding floors on the *full* dipole amplitude A are $2\times$ these values, i.e. $\sim 0.29\%$ Fisher (the 3σ floor on A before mask/$N_{\rm eff}$ inflation), rounded conservatively to $\sim 0.4\%$ once mask + $N_{\rm eff}$ corrections are absorbed. The per-pixel-shuffle empirical 50%-rec-3σ threshold $|A_{\rm dipole}| \geq 0.75\%$ quoted in the abstract is the full-amplitude empirical bound, above the $\sim 0.4\%$ conservative full-amplitude Fisher floor. The analytic Fisher derivation in the remainder of this subsection uses the half-modulation $A/2$ as the unit; all reported floors are converted to the full amplitude A at the point of statement, so no factor-of-2 reconciliation is required.

Given 3,201,160 equivariant spiral classifications (canonical, see Sec. IV A), the global clockwise-fraction uncertainty is $\sigma_{\rm global} = 1/(2\sqrt{N_{\rm spiral}}) \approx 0.028\%$ (using the approximation $p(1-p) \approx 1/4$ valid for p near 0.5; the exact formula $\sigma = \sqrt{p(1-p)/N}$ used in Sec. IV B gives an identical result to four significant figures at $p = 0.4974$). This is the Poisson noise averaged over the full catalog—it characterizes the monopole precision, not the dipole sensitivity. For a dipole of amplitude $A_{\rm dip}$, the signal in any sky pixel at `NSIDE` $= 8$ (768 pixels, $\sim$4,170 spirals per pixel from $N_{\rm spiral}/N_{\rm pix} = 3{,}201{,}160/768 \approx 4{,}168$) is $A_{\rm dip}\cos\theta$, and the per-pixel CW-fraction uncertainty is

$$\sigma_{\rm pix} = \frac{1}{2\sqrt{N_{\rm spiral}/N_{\rm pix}}} = \frac{1}{2\sqrt{4{,}168}} \approx 0.77\% \,. \qquad (6)$$

(An earlier-snapshot evaluation $1/(2\sqrt{4{,}326}) \approx 0.76\%$ used the pre-canonical $N^{\rm snap}_{\rm spiral} = 3{,}321{,}795/768 \approx$

4,326 denominator and is superseded by the canonical $N_{\rm spiral} = 3{,}201{,}160/768 \approx 4{,}168$ value above; the ~ 0.01 percentage-point shift does not affect any downstream threshold to two significant figures.) A 3σ detection of a dipole pattern requires fitting the amplitude $A_{\rm dip}$ against the $\cos\theta$ template across $N_{\rm pix} = 768$ pixels. For a pure cosine template the dipole estimator is equivalent to a single-parameter linear regression, for which the uncertainty on the fitted amplitude is

$$\sigma(A/2) \approx \sigma_{\rm pix}\sqrt{\frac{3}{N_{\rm pix}}} = 0.77\% \times \sqrt{\frac{3}{768}} \approx 0.048\%\,, \quad (7)$$

(note: as stated in the preceding paragraph, this equation computes the Fisher floor on the CW-fraction *half-modulation* $A/2$; the full-amplitude Fisher floor is $2 \times 0.048\% \approx 0.097\%$, rounded to $\sim 0.10\%$; the 3σ full-amplitude floor is $\sim 0.29\%$). where the factor of $\sqrt{3}$ accounts for the three dipole components ($\ell = 1$ has $2\ell + 1 = 3$ modes); geometrically it arises because only one-third of the $\cos^2\theta$ variance projects onto each component. At 3σ significance the minimum detectable amplitude is $3 \times 0.048\% \approx 0.14\%$, which we round conservatively to $\sim$0.2% to account for incomplete sky coverage ($f_{\rm sky} \approx 0.46$) and non-uniform pixel occupancy (the effective sample size is reduced by the survey mask; both effects widen $\sigma(A_{\rm dip})$ by a factor $\sim 1/\sqrt{f_{\rm sky}} \approx 1.5$ in the worst case, but the rounding to 0.2% already captures a $\sim$40% margin over the idealized 0.14%).

The factor of $\sim$7 between $\sigma_{\rm global} \approx 0.028\%$ (canonical three-significant-figure value used uniformly in this section; an earlier-snapshot 0.027% rounding from a coarser two-sig-fig truncation is superseded) and the minimum detectable dipole of 0.2% arises from two compounding steps. Analytically (Eqs. 6–7), the ideal whole-sky ratio is $3\sigma(A_{\rm dip})/\sigma_{\rm global} = 3\sqrt{3} \approx 5.2$, giving a threshold of $5.2 \times 0.028\% \approx 0.146\%$ (rounded to 0.15% in this section's two-sig-fig presentation). The additional factor of ~ 1.4 that brings this to $0.146\% \times 1.4 \approx 0.205\%$ reflects incomplete sky coverage and non-uniform pixel occupancy: the DESI Legacy footprint covers $f_{\rm sky} \approx 0.46$, reducing the effective number of independent pixels and widening $\sigma(A_{\rm dip})$ by $\sim 1/\sqrt{f_{\rm sky}} \approx 1.5$ (the 1.4 multiplier is the realized value after pixel-occupancy weighting; the analytic $1/\sqrt{f_{\rm sky}}$ ceiling is 1.47). We round the resulting 0.205% up to 0.2% as a conservative two-sig-fig statement of the sensitivity floor.

Note on $N_{\rm eff}$ and effective sample size.—The above derivation treats $N_{\rm spiral} = 3{,}201{,}160$ as the effective sample size, weighting all classified spirals equally. The true effective sample size $N_{\rm eff}$ is somewhat smaller because (a) per-galaxy classifier confidence varies (the high-confidence Catalog-C TTA-averaged predictions are not uniform; the lowest-confidence spirals carry information at a reduced effective weight in the variance computation), and (b) the rotational-equivariance residual in the TTA post-processing introduces a per-pixel variance contribution that compounds the Poisson floor. A formally correct treatment would use $N_{\rm eff} = (\sum_i w_i)^2/\sum_i w_i^2 \leq N_{\rm spiral}$ where w_i are classifier confidence weights, and would inflate $\sigma_{\rm pix}$ by $(N_{\rm spiral}/N_{\rm eff})^{1/2} \geq 1$. The $\sim 40\%$ rounding margin from 0.14% up to 0.2% is chosen specifically to absorb this $N_{\rm eff}$ inflation and the partial-sky $1/\sqrt{f_{\rm sky}}$ correction in a single conservative cushion; under typical confidence distributions the $N_{\rm eff}$ correction alone gives a 5–15% inflation of $\sigma_{\rm pix}$, well within the rounding margin. The reported 0.2% minimum detectable dipole is therefore conservative against both the partial-sky and $N_{\rm eff}$ corrections.

b. Statistical-only sensitivity, not systematic-inclusive. The 0.2% floor derived above is the *statistical* Poisson sensitivity to a true sky dipole assuming the systematic monopole has strictly zero dipole projection. The catalog has a known, uncorrected raw-classifier systematic monopole of 0.26% (Catalog A; Sec. IV B, Table IV), reduced to -0.26% (9.5σ from parity) by equivariant TTA but not to zero. Because the 0.2% statistical floor sits *below* this 0.26% systematic monopole floor in absolute amplitude, any putative sub-systematic-floor dipole *detection* requires an independent demonstration that the residual systematic has strictly zero dipole projection on the DESI Legacy footprint. We have not provided that demonstration in the present catalog (a PSF-ellipticity / scan-angle cross-correlation against the equivariant CW-fraction map is the canonical test and is queued as a downstream extension), and so the 0.2% value should be interpreted as a statistical *upper bound* on any true isotropy-breaking axial-vector *dipole* signal in the parity-even ($\ell = 1$) channel under the zero-systematic-dipole-projection assumption, not as a systematic-inclusive sensitivity. The 9.5σ monopole result is a null at the catalog level (uniform across 7 equatorial coordinate slabs; cf. §IV B Catalog C), but the absence of a survey-systematics cross-correlation test means the 0.2% floor cannot at present be elevated to a systematic-inclusive limit.

The PSF-ellipticity and scan-angle cross-correlation test has been performed on the full 8,474,531-galaxy joined catalog (DR8 sweep b/a + ellipticity merged onto Catalog C `class_eq` on `dr8_id`). HEALPix maps at NSIDE=64 are built for (a) the equivariant CW-fraction $f_{\rm CW}(\hat{n})$ from the 3.20M spiral subset and (b) per-pixel population-mean ellipticity proxies $(e_1, e_2, |e|, b/a, |e_1|, |e_2|, \mathrm{PA})$ from the 2.91M spiral-and-ellipticity-defined subset (PA from $\arctan_2(\sum \sin 2\mathrm{PA}, \sum \cos 2\mathrm{PA})/2$). Restricting to pixels with both $n_{\rm spiral}$>100 and $n_{\rm e}$>100 yields 15,769 valid pixels at $f_{\rm sky} \approx 0.32$, finite across all seven proxies. Two quantitative tests are reported. *Test 1 (pixel-level Pearson r, strict* 0.1% *flatness bar):* $\max_i |r(f_{\rm CW}, p_i)| = +0.04243$ (driven by $r(f_{\rm CW}, e_2) = +0.04243$, $p = 9.8\times10^{-8}$, and $r(f_{\rm CW}, \mathrm{PA}) = +0.04185$, $p = 1.5\times10^{-7}$), which formally fails the strict pixel-level $|r| < 10^{-3}$ bar. *Test 2 (angular cross-power $C_\ell^{f_{\rm CW}\times p}$ with $N_{\rm MC}$=200 pixel-shuffle null at $\ell_{\rm max}$=128, the physics-relevant test for cosmological-scale dipole leakage):*

$\max_{\text{correlator,band}}|z| = 2.72\sigma$ across all six (correlator $\times\ell$-band) combinations, well below the 3σ cosmological-systematic bar. The largest band, $z(f_{\rm CW}\times e_1|\ell{=}2{-}10) = -2.72\sigma$, is consistent with fluctuations expected under the null at this number of independent modes. *Verdict:* the small but statistically significant pixel-level Pearson correlations ($\sim 4\%$, two orders of magnitude above the 0.1% bar) indicate that PSF/ellipticity-related fields and $f_{\rm CW}$ are not pixel-level orthogonal at the population-mean smoothing of NSIDE=64, but those couplings do *not* project into a $\geq 3\sigma$ cross-power on the ℓ=2−64 scales relevant to a cosmological parity dipole. The 0.2% statistical floor is therefore best read as a near-zero residual coupling on cosmological scales, not as a strict pixel-level orthogonality. Companion artifact: `pipelines/p2_chirality/r42_results/wave_14_jj_psf_xcorr_results.json` (full Pearson + Spearman + C_ℓ z-score table, MC null seeds, NSIDE/$\ell_{\rm max}$/$N_{\rm MC}$ specification). Figure 13 visualizes Test 1 and Test 2 results side-by-side.

We further verify the 0.2% minimum-detectable-dipole headline via a Monte Carlo injection-recovery test on the real DESI Legacy mask and Catalog C per-pixel depth. The generator `pipelines/p2_chirality/wave_14_nn_dipole_mc_injection.py` loads the 471,049-spiral "HC-spiral" high-confidence subsample defined by equivariant probability > 0.9 for either CW or CCW (canonical sample label in `pipelines/p2_chirality/outputs/canonical_provenance/fisher_sensitivity_floor.json`, key `sample_label`); this is distinct from the broader "HC-broad" cut in Sec. VI D 0 c that uses $\max(p_{\rm CW,eq}, p_{\rm CCW,eq}, p_{\rm NS,eq}) > 0.6$ without the spiral restriction and yields $n = 949{,}584$), sweeps nine injection amplitudes $A \in \{0.05, 0.10, 0.20, 0.30, 0.50, 0.75, 1.00, 1.50, 2.00\}\%$ (extended sweep; cf. Table XIII), and for each amplitude draws $N_{\rm inj} = 100$ random sky directions $\hat{n}_{\rm inj}$ as injection axes. For every (amplitude, axis) realization, each spiral is re-labelled CW with probability $p_{\rm CW}(\hat{n}_{\rm gal}) = \frac{1}{2}(1 + A\,\hat{n}_{\rm gal}\cdot\hat{n}_{\rm inj})$, the catalog is pixelized to NSIDE=64 (20,838/49,152 pixels above the ≥ 10 spirals/pixel cut, $f_{\rm sky}$=0.4240), the dipole is fit via `healpy.fit_dipole`, and a per-pixel-shuffle null with $N_{\rm MC,null} = 1000$ realizations is run for significance (canonical `n_mc_null`= 1000 from `pipelines/p2_chirality/outputs/canonical_provenance/injection_recovery_extended.json`). Aggregating over 100 random axes per amplitude yields the median significance $\langle\sigma\rangle$, the empirical detection probability $P(\sigma > 2)$, and the direction-recovery probability $P(\Delta\theta_{\rm rec,inj} < 30°)$; the full sweep is given in Table XIII.

The empirical 50%-recovery-at-3σ threshold first crosses at $A = 0.75\%$ in the extended sweep ($P(\sigma{>}3){=}0.55$ at $A{=}0.75\%$, rising to 0.91 at $A{=}1.00\%$ and 1.00 at $A \geq 1.50\%$); the $\{0.05\%, \ldots, 0.50\%\}$ range gives $P(\sigma > 3){\leq}\ 0.15$ at all tested amplitudes, with $A = 0.5\%$ at $P(\sigma > 3){=}\ 0.15$ ($P(\sigma > 2){=}\ 0.35$, median $\sigma = +1.73$) — a non-detection rather than a calibrated detection threshold. *Verdict:* the 0.2% in the abstract and §VI C is the *statistical Poisson floor* ($\sigma_{\rm stat} \approx 1/\sqrt{3N_{\rm spirals}f_{\rm sky}}\times 1.4$, the analytic Fisher quantity in Eq. 7, already framed in the paragraph at L1553–L1574 as a "statistical *upper bound*" under zero-systematic-dipole-projection assumptions). The empirical MC threshold of $\geq 0.75\%$ is the empirical 50%-recovery-3σ amplitude (at $A = 0.75\%$ the per-pixel-shuffle MC gives $P(\sigma > 3) = 0.55$; cf. $A = 0.5\%$ gives $P(\sigma > 3) = 0.15$, a tested non-detection point) under a strict per-pixel-shuffle null, which preserves the full per-pixel CW-asymmetry variance (including the $\sim$0.005% post-TTA residual signature in Catalog C) and is therefore intentionally conservative. The like-for-like HC-subsample Fisher floor is $3\sqrt{3/471{,}049} \approx 0.76\%$, so the HC empirical 0.75% tracks the HC Fisher floor at a ratio of ≈ 1.0 (no systematic-inclusive degradation on the HC subsample). prior text in this paragraph compared the HC-subsample empirical 0.75% to the

TABLE XIII. Per-pixel-shuffle empirical injection-recovery sweep on the canonical strict-HC subsample (`wave_14_nn` pipeline: $N = 471{,}049$ HC spirals with $\max(p_{\rm CW,eq}, p_{\rm CCW,eq}) > 0.9$ AND per-pixel-count ≥ 10 filter; companion `pipelines/p2_chirality/outputs/canonical_provenance/wave14nninjectionrecovery.json`; $N_{\rm inj} = 100$ axes per amplitude; $N_{\rm MC,null} = 500$). The extended sweep (v1.0.78) adds the $A \in \{0.75, 1.00, 1.50, 2.00\}\%$ rows that complete the 50%-recovery-3σ localization at $A = 0.75\%$ under this pipeline. (pipeline-choice sensitivity range: three companion sweeps exist with different predicates and slightly different per-pixel-count filters: (a) the `wave_14_nn` primary sweep used here ($N{=}471{,}049$ strict HC + pix-count ≥ 10 filter; 50%-rec-3σ at $A{=}0.75\%$); (b) the v1.0.78-extended sweep at HC-broad $p_{\rm eq} > 0.6$ (`pipelines/p2_chirality/outputs/canonical_provenance/injectionrecoveryextended.json`; $n \approx 949{,}584$; 50%-rec-3σ also at $A = 0.75\%$, larger N partly offsets the looser confidence cut); (c) a v1.0.121 strict-HC sweep without the per-pixel-count filter (`pipelines/p2_chirality/outputs/canonical_provenance/injectionrecoveryextendedhc09.json`; $N = 496{,}531$; 50%-rec-3σ at $A = 1.5\%$). The three thresholds bracket the pipeline-choice sensitivity range $[0.75\%, 1.5\%]$. We adopt 0.75% as the canonical present-pipeline threshold for the abstract + falsification criterion because it is the threshold of the `wave_14_nn` pipeline that produced Table I row (vi); 1.5% is reported as a robustness stress test under a less-restrictive per-pixel-count pipeline. The two cited cuts are non-identical predicates and should not be cross-cited as the same sweep.)

A	$\langle\sigma\rangle$	$P(\sigma{>}2)$	$P(\sigma{>}3)$	$P(\Delta\theta_{\rm rec,inj}{<}30°)$
0.05%	−0.37	0.04	0.01	0.14
0.10%	−0.02	0.03	0.01	—
0.20%	+0.31	0.08	0.01	0.21
0.30%	+0.52	0.08	0.03	—
0.50%	+1.73	0.35	0.15	0.38
0.75%	+3.17	0.83	0.55	—
1.00%	+4.89	0.96	0.91	—
1.50%	+8.30	1.00	1.00	—
2.00%	+11.71	1.00	1.00	—

full-catalog Fisher floor $3\sqrt{3/3{,}201{,}160} \approx 0.29\%$ as a $\sim 2.5\times$ "Fisher-vs-empirical" factor; this comparison is cross-sample (subsample empirical vs full-catalog ideal) and is therefore not a meaningful empirical-vs-Fisher gap. The valid comparison is the like-for-like HC pair just stated (empirical 0.75% vs HC Fisher 0.76%, ratio $\approx$1.0); the historical $\sim 2.5\times$ cross-sample ratio has been retracted as a methodological artifact. The original historical reference between the empirical HC-subsample threshold 0.75% and the FULL-catalog Fisher floor 0.29% is a cross-sample ratio, NOT a Fisher-vs-empirical degradation factor for similar single-cosine estimators (the relevant degradation is the 0.75/0.76 $\approx$ 1.0 HC like-for-like ratio above). A full-catalog injection sweep that would define a meaningful systematic-inclusive gap on the 3.2M sample is deferred to future work. Both numbers stand. The catalog's reported result — Catalog C post-TTA dipole at σ=0.43, p=0.30 — is a non-detection under *either* threshold, so the central claim of the paper (no $\geq 3\sigma$ isotropy-breaking axial-vector chirality dipole) is independent of which sensitivity is adopted. Companion artifact: `pipelines/p2_chirality/r42_results/wave_14_nn_injection_recovery.json` (729.7 s wall on Pod 3 H200, pure CPU `healpy` pipeline, 900 injection fits ($N_{\rm inj}$ = 100 axes × 9 amplitudes; canonical `n_mc_inj_per_amp`= 100 from `pipelines/p2_chirality/outputs/canonical_provenance/injection_recovery_extended.json`) calibrated against $N_{\rm MC,null}$ = 1000 per-pixel-shuffle realizations per amplitude; full per-injection table with recovered (l, b), monopole, σ, p-value, and $\Delta\theta_{\rm rec,inj}$ for every one of the 500 injections).

This represents the most sensitive chirality measurement ever performed, exceeding the CE-ResNet constraint [8] by a factor of ~ 1.3 in statistical sensitivity owing to the 1.64× larger spiral sample (3.20 million equivariant spirals [canonical $N_{\rm spiral}$ = 3,201,160; the 3.32 million figure of earlier drafts is the snapshot superseded by the $N_{\rm spiral}$-correction recount, retained only in the headline of Table X as the before/after diagnostic baseline] vs. CE-ResNet's 1.95 million) and reduced per-pixel variance. (The total galaxy count is 4.3× larger than CE-ResNet, but the chirality-relevant comparison is the spiral count, which drives the $\sqrt{1.64} \approx 1.28$ improvement.)

Any future redshift-binned extension of this analysis must additionally contend with the photo-z smearing floor introduced in Sec. II A: with $\sigma_z/(1+z) \approx 0.03$ from the underlying DESI Legacy photo-z catalog, individual galaxies can migrate across redshift-bin boundaries by $\Delta z \sim 0.05$ at $z \sim 0.5$, diluting any narrow-redshift-shell dipole signal. Spectroscopic follow-up (e.g. DESI DR1+ spectra) is required to recover the full sensitivity floor in redshift-resolved dipole searches.

D. Edge-On Galaxy Contamination

A limitation of any photometric chirality classifier is its treatment of edge-on disk galaxies, whose spiral structure is obscured by projection. In our catalog, we find that 65.7% of visually identified edge-on systems (axis ratio $b/a < 0.3$, estimated from DESI Legacy photometric catalogs) receive a CW or CCW classification rather than NOT_SPIRAL. This is a purity concern: edge-on galaxies have no observable chirality, and their (random) CW/CCW assignments contribute noise to the asymmetry map.

However, this contamination does not bias the dipole analysis, because the equivariant averaging procedure enforces flip-equivariance of the soft-probability protocol, so for any galaxy whose mirror image is morphologically indistinguishable from the original (as for edge-on disks) the ensemble-mean CW and CCW probabilities are flip-symmetric; per-galaxy hard-label assignment retains the $\sim$ 21% argmax-flip rate documented in Sec. III E as a residual rotational uncertainty. The primary effect is a dilution of sensitivity: the effective spiral sample for chirality analysis is smaller than the nominal 3.20 million, by a factor that depends on the (unknown) true edge-on fraction among objects classified as spirals. We estimate this dilution reduces the effective sample by $\sim$10–15%, corresponding to a sensitivity penalty of $\sim$5–8%.

While our classification pipeline is intended to assign edge-on galaxies to NOT_SPIRAL, only 34.3% of $b/a < 0.3$ objects actually receive NOT_SPIRAL classifications (Sec. VI above), leaving 65.7% with random CW/CCW labels. For near-edge-on galaxies ($0.3 < b/a < 0.5$), residual contamination could introduce a systematic CW/CCW asymmetry if spiral arm orientation correlates with viewing angle in this intermediate-inclination regime. A direct test would cross-match the catalog against the DESI Legacy morphological catalogs (which provide Sersic-fit axis ratios for all photometric sources) and compute the equivariant CW fraction in three inclination bins: face-on ($b/a > 0.5$), intermediate ($0.3 < b/a < 0.5$), and edge-on ($b/a < 0.3$). If the equivariant averaging fully eliminates orientation-dependent bias, all three subsamples should yield CW/(CW + CCW) consistent with the global value of 0.4974 ± 0.0003. Any statistically significant departure in the $b/a < 0.3$ subsample—where the true chirality is undefined—would flag residual classifier leakage that the equivariant procedure does not correct.

We estimate $\sim$ 5–8% of the 3.20 million equivariant spirals have $b/a < 0.3$ (based on the typical edge-on fraction in magnitude-limited disk samples), giving $\sim$200,000 objects in the edge-on bin—sufficient to detect a CW-fraction shift of $> 0.34\%$ at 3σ. The face-on subsample ($b/a > 0.5$, $\sim$2.5 million spirals) would provide the cleanest dipole measurement, with a sensitivity floor of $\sim$0.22% at 3σ—only marginally degraded relative to the full sample. Axis-ratio data are not included in Catalog C (which carries sky coordinates and classification

probabilities only; Sec. III G), so a per-bin axis-ratio test requires a positional cross-match with the parent DESI Legacy photometric catalog.

a. High-confidence-spiral subsample robustness rerun (v1.0.69). As a proxy for the b/a-cut face-on subsample (which requires a sweep cross-match not in Catalog C), we use the classifier-confidence cut $\max(p_{\mathrm{CW,eq}}, p_{\mathrm{CCW,eq}}) > 0.6$ (the HC-spiral subsample $N = 949{,}584$, paper-existing concept) and the stricter > 0.8 cut (HC-strict, $N = 624{,}660$). Edge-on disks typically have ambiguous CW/CCW post-TTA and low max-class confidence; the HC cut therefore selects against edge-on contamination without requiring axis-ratio metadata.

b. Reconciliation with the headline $+0.43\sigma$ real-space dipole. The Catalog C-full $+4.31\sigma$ in this table and the headline $+0.43\sigma$ real-space dipole of Sec. IV C are computed with different estimators on the same on-sky data: the present table uses a weighted least-squares $\cos\theta$ fit on a pixel map and reports the amplitude relative to a small-N_{MC} monopole-preserving null, while the headline statistic uses a Healpix dipole fit with $N_{\mathrm{MC}} = 10{,}000$ per-pixel-shuffle null. Both null constructions preserve the global $p_{\mathrm{CW}} = 0.49735$ monopole by design; the numerical disagreement reflects estimator definition (linear-LSQ amplitude vs Healpix dipole) and null-sample variance, not a contradiction. The two values are sensitivity-comparable: in both estimators the HC-cut subsamples collapse to $|\sigma| \leq 1\sigma$.

The two HC-cut subsamples (which exclude the lowest-confidence classifications, where edge-on contamination is concentrated) show the dipole signal collapse from $+4.31\sigma$ to $+0.62\sigma$ and $+0.87\sigma$ against the monopole-preserving null. This demonstrates that (i) the canonical-mask leakage seen in Sec. IV D scales with sample size + edge-on contamination as expected, (ii) when the edge-on-dominated tail is removed, the residual is consistent with the null, and (iii) the edge-on-contamination concern raised in this section is mitigated empirically: cleaner subsamples do not produce stronger dipole signal. Combined with the per-imaging-leg result of Sec. IV I (all three legs null individually), the robustness picture is consistent.

We can, however, quantify the equivariance correction directly from the production catalog without a positional cross-match by using the raw → equivariant label transition as an edge-on proxy: galaxies whose raw classifier returned CW or CCW but whose equivariant ensemble reassigned to NOT_SPIRAL are exactly the ambiguous-handedness pool, and edge-on disks dominate that pool because they are the photometric class for which the mirror image is least distinguishable from the original. Across the full 8,474,531-galaxy catalog, 3,445 objects (0.041%) flip from raw-CW/CCW to equivariant-NOT_SPIRAL after the symmetry correction, and within that pool the raw model carries a 13.3% CCW excess. The corresponding catalog-wide handedness asymmetry drops from +2.05% in the raw model to −0.53% after equivariance—a suppression factor of 3.86×.[13] A complementary confidence-stratified NOT_SPIRAL-rate ladder shows the equivariance step doing real work on ambiguous calls (only 16–31% flagged NOT_SPIRAL at confidence < 0.6 vs. 91% at confidence > 0.95), confirming that the correction concentrates on the genuinely uncertain (edge-on-dominated) subsample rather than affecting the bulk catalog uniformly. The remaining −0.53% residual is consistent with the ∼10% sample-dilution sensitivity penalty estimated above and is absorbed in the 0.43σ post-TTA dipole null.

The b/a-binned cross-match has been performed by joining the full 8,474,531-galaxy DR8 sweep (b/a column from `shapeexp`/`shapedev` fracdev-weighted ellipticity) onto Catalog C `class_eq` on `dr8_id`, with 7,728,567 rows (91.20%) carrying a finite b/a (the missing 8.80% are point-source/PSF-fit type rows where DESI Legacy does not solve a Sersic profile). The four-bin reconciliation table is: $b/a \geq 0.5$ *(face-on)*: 5,280,709 galaxies, 31.4% spiral classification rate (825,957 CW + 834,293 CCW); $0.3 \leq b/a < 0.5$ *(intermediate)*: 1,661,999 galaxies, 47.2% spiral rate (390,184 CW + 394,871 CCW); $b/a < 0.3$ *(edge-on)*: 785,859 galaxies, 59.4% spiral rate (232,012 CW + 234,318 CCW); *nan b/a (PSF-type)*: 745,964 galaxies, 38.8% spiral rate. The catalog-wide CW fraction CW/(CW + CCW) is consistent across all four bins to within the per-bin Poisson floor: face-on 0.4975, intermediate 0.4970, edge-on 0.4975, nan 0.4972, all-bin 0.4974, against the canonical 0.4974 ± 0.0003 cited above. This explicitly closes the "would flag residual classifier leakage" check anticipated in the preceding paragraphs: there is no CW-fraction shift in the $b/a < 0.3$ subsample where the true chirality is undefined, so the equivariant procedure does fully eliminate orientation-dependent bias to the per-bin sensitivity of the test. The 59.4% edge-on spiral classification rate is ∼ 6 pp lower than the prior preliminary estimate of 65.7% cited above (from a visual-ID-curated edge-on subsample); we adopt the full-DR8-sweep 59.4% as canonical going forward and read the prior 65.7% as upper-bounded by visual-ID curation bias on the edge-on subsample. The raw→equivariant CW↔CCW direction flip count over the full catalog is 5,586 + 4,977 = 10,563 (0.125%), of which 920 + 710 = 1,630 (0.21% of the edge-on b/a < 0.3 subsample) occur at edge-on inclinations. Reconciliation of the 53,862 vs. 3,445 figures: (i) 53,862 is the count of GZ1 ground-truth NOT_SPIRAL labels in the cross-matched GZ1 subset (`class_raw_x == "NOT_SPIRAL"` in Catalog C; not an edge-on count), against which the model's three-class accuracy and sidedness-aware metrics are benchmarked in Sec. III C; (ii) 3,445 is the

[13] Numbers from `pipelines/h200_results/pod2_chirality_2026-04-29/edgeon_contamination.json`; the analysis treats the on-disk `catalog_production.parquet` columns `class_raw_x` and `class_eq` as the raw and equivariant calls, computes the 3×3 transition matrix, and reports both the catalog-level asymmetry and the within-pool handedness imbalance.

TABLE XIV. High-confidence-spiral robustness rerun. Each row reports the dipole σ under *two* different null hypotheses for the same on-sky data, at NSIDE = 64, $N_{\rm MC} = 1000$, seed = 42, full-amplitude convention $p_{\rm CW}(\hat{n}) = \frac{1}{2}(1 + A\cos\theta)$. The *monopole-preserving* null draws each subsample's CW counts from Binomial($n_{\rm pix}, p_{\rm CW}^{\rm sub}$) **using that subsample's own global** $p_{\rm CW}^{\rm sub}$ (0.49735, 0.49606, 0.49602 for the three rows respectively, not the full-catalog 0.49735 applied uniformly); this isolates the dipole-only signal at fixed monopole. The *isotropic-p* = 0.5 null draws from Binomial($n_{\rm pix}$, 0.5); this is the same null construction used elsewhere in the paper but does not preserve the 9.5 σ uniform monopole, so the resulting sigma for any sample with a non-zero monopole is the joint detection of "monopole + dipole" rather than just dipole; we report it for like-for-like comparison with the reviewers' request. The HC-cut collapse from $+4.3\sigma$ to $\lesssim 1\sigma$ is consistent under both nulls. Verification: pipelines/p2_chirality/outputs/canonical_provenance/face_on_robustness_results.json and pipelines/p2_chirality/outputs/canonical_provenance/face_on_isotropic_null_results.json.

Sample	$N_{\rm spiral}$	CW/(CW + CCW)	$\lvert\Delta\rvert$ (%)	monopole-preserving σ (p)	isotropic-p=0.5 σ (p)
Catalog C full	3,201,160	0.49735	0.265	+4.31 (0.001)	+4.43 (0.001)
HC-broad-0.6 [a]	949,584	0.49606	0.394	+0.62 (0.243)	+1.58 (0.078)
HC-strict ($p_{\rm eq} > 0.8$)	624,660	0.49602	0.398	+0.87 (0.187)	+1.07 (0.153)

[a] HC-broad-0.6: $\max(p_{\rm CW,eq}, p_{\rm CCW,eq}) > 0.6$ spiral-only cut, from `face_on_robustness_results.json`; distinct from the stricter HC-spiral-0.9 cut ($p_{\rm eq} > 0.9$, $N = 471{,}049$) used in §IX.J for MC injection-recovery.

catalog-wide raw-CW/CCW → equivariant-NOT_SPIRAL flip count (0.041% of 8,474,531, the proxy for the equivariance correction's edge-on retargeting) reported in the previous paragraph; the two quantities measure different things and are not contradictory. Companion artifact: pipelines/p2_chirality/r42_results/wave_14_kk_ba_reconciliation_results.json (full b/a-bin reconciliation table with raw-class transition matrix, b/a coverage statistics, and edge-on subsample dilution quantification).

We can also directly test whether equivariant averaging eliminates rotational-equivariance violations in the edge-on population. The 4-bin reconciliation just reported (full DR8 sweep b/a × Catalog C `class_eq` join on `dr8_id`) directly answers this question on the deployed 8,474,531-galaxy catalog: the b/a < 0.3 edge-on subsample of 785,859 galaxies (with finite b/a from the DR8 Sersic fit) has equivariant CW fraction CW/(CW + CCW) = 0.4975 ± 0.0006 (Poisson SE), *statistically indistinguishable from the catalog-wide post-TTA value of* 0.4974 ± 0.0003 *and from the three other b/a bins at* 0.4975 *(face-on),* 0.4970 *(intermediate), and* 0.4972 *(nan-b/a/PSF-type).* The ∼ 0.25–0.30% offset from 0.5000 that appears in the edge-on bin is the *same* offset present in every other b/a regime; it is therefore the catalog-wide post-TTA residual already reported above, not an edge-on-specific TTA failure mode. Were diagonal-edge-on rotational non-equivariance under DESI Legacy scan-direction systematics to be biasing chirality, we would expect the b/a < 0.3 bin to deviate *differently* from the three face-on/intermediate/nan bins; we do not observe any such bin-specific shift to the per-bin Poisson floor of ∼ 0.06%. The maximum bin-to-bin spread in CW fraction across the four orientation regimes is 0.0005 (0.05%), well below the 0.1% flatness target. Companion artifact: pipelines/p2_chirality/r42_results/wave_14_kk_ba_reconciliation_results.json.

c. Bin-by-bin CW flatness test across four morphology axes. We test whether the CW-fraction is uniform to < 0.1% across r-mag, surface brightness, Sersic size, PSF FWHM, and edge-on b/a by running the diagnostic on the joined chirality × DR8 sweep catalog (pipelines/p2_chirality/wave_14_oo_bin_flatness.py, 29.3 s wall on Pod 3, $0 marginal GPU spend—the operation is denominator-bookkeeping, not GPU work) at $N_{\rm bins}$=10 for the four continuous DR8-sweep axes available without a re-fetch (Sersic effective radius $\log_{10} r_{\rm eff}$, fracdev de-Vaucouleurs fraction, axis ratio b/a, and `type` categorical: PSF/REX/EXP/DEV/COMP). Two denominators are reported. *Full spiral subsample* ($n_{\rm sp}$=3,201,160, catalog-wide CW fraction 0.4974): shape_r_eff $\Delta f_{\rm CW}$=0.32% (FAILS 0.1%), fracdev Δ=1.41% (FAILS), b/a Δ=0.23% (FAILS at 10-bin granularity though the 4-bin reconciliation above showed 0.05%), `type` Δ=0.08% (PASSES). *High-confidence subsample* (HC-broad-0.6: n=949,584, $\max(p_{\rm CW,eq}, p_{\rm CCW,eq})$>0.6 – the spiral-only confidence> 0.6 cut from `face_on_robustness_results.json`; distinct from the spiral-only stricter HC-spiral-0.9 cut n = 471,049 used in §IX.J for MC injection-recovery, which is the canonical sample-label `HC-spiral (equivariant probability > 0.9 for either CW or CCW)` from `fisher_sensitivity_floor.json`; the prior "HC-broad max(p_CW,p_CCW,p_NS)> 0.6 includes confident-NS" label used in some sites was wrong and is corrected here): all four axes show larger spreads (0.49%−3.03%) as expected from the smaller per-bin denominator. *Reframe and verdict.* The three failing axes (size, fracdev, b/a at 10-bin granularity) expose a real morphology-classification correlation: CW vs. CCW classification rates are not perfectly identical across galaxies of different size, Sérsic profile, and inclination—this is expected of any morphology classifier that operates on per-galaxy images. Crucially, *this is not a directional dipole*: the hemisphere test with $N_{\rm MC}$=10,000 already excluded a sky-direction asymmetry at $p_{\rm LEE} < 10^{-4}$ (MC-resolution upper bound), and the empirical MC injection-recovery test (§ IV C) showed no detected

dipole below the $\geq$ 0.75% empirical 50%-rec-3σ amplitude threshold under strict per-pixel-shuffle nulls. The bin-flatness spreads here are amplitude differences across morphology bins that are spatially uniform on the sky, orthogonal to the isotropy-breaking dipole question the paper's central claim addresses. *Verdict:* the strict $<$ 0.1% flatness bar holds for type (categorical morphology), not for fine-binned continuous size/fracdev/b/a axes; the 0.05%−0.08% pass-bar applies at the categorical/coarse-binning granularity used in the 4-bin reconciliation above; and the central no-detection conclusion ($\sigma_{\rm hemi}$=0.43, p=0.30) is independent of which axis-binning granularity is adopted. Companion artifact: `pipelines/p2_chirality/r42_results/wave_14_oo_bin_flatness.json` (23.9 KB, both denominators, four axes, all bin-level rows). A follow-up binning test will add the r-mag and surface-brightness axes once the DR8 `flux_r` sweep fetch completes. The per-bin CW-fraction data for the three continuous morphology axes (Sersic-radius proxy $\log_{10} r_{\rm eff}$, fracdev, b/a) are plotted in Fig. 14; the per-bin Poisson scatter is consistent with the catalog-wide 0.4974 baseline, with the dominant contributor to the Δ = 1.41% fracdev spread coming from the small-N ($fracdev > 0.5$) bin (n = 10,941). For the PSF-systematics analog (which is not bin-tabulated to per-bin CW fractions in the production artifacts), Wave 14-JJ reports $|r_{\rm cw,e_1}|$=0.043 and $|r_{\rm cw,e_2}|$ = 0.042 Pearson correlations between CW probability and PSF-ellipticity components, with the maximum bandpower-binned cross-power z-score $|z| \leq 2.73$ in $\ell \in [2, 10]$ (`pipelines/p2_chirality/r42_results/wave_14_jj_psf_xcorr_results.json`).

E. Spiral Fraction Variation Across the Sky

The fraction of galaxies classified as spiral (CW + CCW) varies enormously across the sky, from ∼25% in regions near the Galactic plane to > 50% in the deepest North Galactic Cap fields. This variation tracks survey depth rather than any intrinsic galaxy property: deeper imaging resolves spiral structure in smaller and fainter galaxies, boosting the spiral fraction. Critically, while the *spiral fraction* is depth-dependent, the *chirality balance* (CW/CCW ratio among spirals) is not—it remains within 0.5% of 50/50 in all sky regions (Table X). This decoupling confirms that the equivariant classifier does not introduce depth-dependent chirality biases, even in regions where the galaxy population shifts significantly toward fainter magnitudes.

F. Mask robustness: pixel-count threshold sweep

ChatGPT v1.0.122 external review MAJOR M3 requested an objective robustness check of the canonical-mask +3.64σ ℓ=1 result against the pixel-count threshold used to define the mask. We sweep the canonical-mask definition over $n_{\rm total} > \{1, 5, 10, 20, 50\}$ galaxies per pixel and rerun the full MASTER coupling-matrix + monopole-only N = 200 null at each threshold (script `pipelines/p2_chirality/scripts/mask_threshold_robustness_sweep.py`; companion artifact `pipelines/p2_chirality/outputs/canonical_provenance/mask_threshold_robustness_sweep.json`).

TABLE XV. Canonical-mask ℓ = 1 MASTER robustness sweep across pixel-count thresholds. $\sigma_{\rm from\ null}$ is the data post-MASTER C_1 versus a monopole-only N = 200 binomial null (matching the v1.0.121 `master_decoupled_monopole_null.json` protocol, NOT the per-pixel-shuffle null that yielded the headline +3.64σ in Table I). For comparison, the v1.0.121 same-protocol canonical-mask $n_{\rm total}>0$ result is +4.84σ.

$n_{\rm total}>$	$f_{\rm sky}$	in-mask spirals	$\sigma_{\rm from\ null}$
1	0.493	3,201,122	+6.31
5	0.491	3,200,836	+8.26
10	0.490	3,200,340	+7.05
20	0.488	3,198,998	+7.05
50	0.478	3,180,526	+6.47

The $\sigma_{\rm from\ null}$ is robust at +6−8 across all five threshold values; the signal does *not* attenuate as low-count edge pixels are progressively excluded. This rules out the "low-count-edge artifact" interpretation of the canonical-mask residual: had the residual been driven by sparse pixels at the mask boundary, the signal should monotonically decrease with threshold; we observe the opposite (the n> 5 peak is consistent with the combined effect of removing the noisiest pixels while preserving the broadband signal structure). Combined with the v1.0.108 multi-null battery results (apodized canonical mask +3.57σ; the apodization axis is therefore also robust), this closes the mask-definition robustness question raised in ChatGPT M3 on both apodization and pixel-count axes.

Estimator hierarchy (pre-specified). The mask hierarchy is: (i) the subsample mask ($f_{\rm sky} = 0.659$) defined in §IV C as the strict-superset region of contiguous galaxy-positive pixels is the load-bearing cosmological estimator (post-MASTER −0.12σ); (ii) the canonical mask ($f_{\rm sky} = 0.49005$, $n_{\rm total} > 0$) is the diagnostic mask used to characterize systematic floors; (iii) the pixel-count-thresholded variants of Table XV are the robustness controls. The subsample mask was constructed before the ℓ=1 MASTER decoupling was run; its definition is geometric (superset of contiguous galaxy-positive pixels with apodization margin) rather than chosen post-hoc to produce a null.

G. Relation to possible parity-violating sectors: transfer-function caveats

The morphological-chirality dipole null reported in this work constrains a specific late-universe, projected,

morphology-channel observable at $z \lesssim 1$. We sketch its qualitative relationship to two established cosmological parity-violation channels in the modern literature; we do not perform an end-to-end transfer- function calculation here.

a. Symmetry classification of the chirality dipole ($\ell = 1$) vs. monopole ($\ell = 0$) observables. For a parity-odd scalar (pseudoscalar) field on the sphere, the parity-transformed field satisfies $A^P(\hat{n}) = -A(-\hat{n})$ (note: this is a transformation rule mapping the field under parity to its pre-image, NOT an identity claiming the field is maximally antisymmetric across the sky; the latter would imply a maximal dipole by construction). The projected CW-fraction asymmetry $(N_{\rm CW} - N_{\rm CCW})/(N_{\rm CW} + N_{\rm CCW})$ is the pseudoscalar field of interest in this paper. Under the parity transformation rule above, the spherical-harmonic coefficients $a_{\ell m}$ satisfy $a^P_{\ell m} = (-1)^{\ell+1} a_{\ell m}$ under a global parity transformation. The monopole ($\ell = 0$) and the even-ℓ multipoles are therefore the parity-*odd* observables (their sign changes under parity), while the dipole ($\ell = 1$) and the odd-$\ell \geq 3$ multipoles are parity-*even* (sign preserved). A nonzero chirality dipole therefore strictly tests *isotropy* (the existence of a preferred axial-vector direction on the sky — i.e. the dipole vector is parity-even and axial, not polar), not parity violation per se. The signed parity-odd diagnostics in our data are the monopole $\langle p_{\rm CW} \rangle = 0.49735$ (9.5σ from 50/50, attributed to citizen-science label systematics propagating through training and reported transparently in Sec. IV B) and the signed real-space asymmetry map $A(\hat{n}) = 2 f_{\rm CW}(\hat{n}) - 1$; the pseudo-$C_\ell = |a_{\ell m}|^2$ bandpowers reported throughout this work at all ℓ are anisotropy and systematics diagnostics rather than direct parity-odd tests (the modulus squared discards the parity sign for every ℓ: $|a_{\ell m}|^2$ is invariant under $a_{\ell m} \to (-1)^{\ell+1} a_{\ell m}$). We retain the language "parity-violating chirality dipole" for continuity with the Shamir literature whose claim class we test, but the more precise statement is that we test *anisotropy of the projected chirality field on the celestial sphere*: a nonzero dipole would indicate a preferred axis in the cosmological-principle sense, and the present null sets an upper bound on any such axis at the $A_{\rm dipole} \lesssim 0.75\%$ amplitude.

b. Late-universe to primordial: the link, and its caveats. The chirality-dipole observable measured in this work is a late-universe, projected morphology-channel quantity at $z \lesssim 1$, mediated by tidal-torque theory (TTT [16]) and subject to baryonic-evolution and intrinsic-alignment systematics that need not be parity-violating themselves. A primordial chiral tensor signal at horizon crossing translates to the observed morphology-dipole channel with a model-dependent transfer function whose computation requires (i) tracking the chiral tensor mode through recombination and matter-radiation equality, (ii) the linear-response coupling to halo angular momentum (Yu *et al.* [28]), and (iii) the projection onto the 2D arm-winding observable conditional on the DESI Legacy footprint and depth. We do not perform this end-to-end calculation in the present paper; the dipole null is therefore a direct constraint on the late-universe morphology channel and only an indirect, model- dependent constraint on primordial parity-violating sectors.

c. (i) Chiral gravitational-wave power asymmetry Π. In Lue, Wang & Kamionkowski's parametrization [17], inflationary or post-inflationary parity-violating tensor sources generate an imbalance $\Pi \equiv (P_L - P_R)/(P_L + P_R)$ between left- and right-handed gravitational-wave power. Tidal-torque theory (Doroshkevich 1970; White 1984) provides a kinematic correlation between galaxy spins and the large-scale tidal field; the existing observational evidence for this TTT-spin link is marginal (Motloch & Pen [16] report a $\sim 2\sigma$ correlation that, as we note in §V D, is fully consistent with the same reading-direction citizen-science labelling bias that contaminates our own pre-equivariance pipeline, and the present work's null dipole is in fact a *cleaner* probe of this channel because it is insensitive to that bias). If the tidal field carries a chiral tensor secondary, the spin direction inherits a coherent bias whose projection onto arm-winding produces the observable chirality dipole [28]. We do not derive the morphology-to-Π transfer function here, and therefore do not quote a numerical bound on Π from the present measurement: the proportionality constant between $|A_{\rm dipole}|$ and $|\Pi|$ at the dipole-equivalent angular scale on the DESI Legacy footprint depends on model-specific projection kernels whose computation is left to follow-up theory work. The CMB-birefringence channel [20] reports a 3.6σ measurement $\beta = 0.342^\circ \pm 0.094^\circ$ constraining a different parity-violating coupling (axion–photon Chern–Simons [23]); the morphology channel and the CMB-birefringence channel are not directly numerically comparable in any common parameter, and we emphasize only that the two channels are *complementary*: a model can saturate one constraint while satisfying the other.

d. What does the present null constrain? The morphology-dipole null at the empirical $|A_{\rm dipole}| \geq 0.75\%$ empirical 50%-rec-3σ amplitude threshold is a direct constraint on the late-universe projected morphology-channel dipole on the DESI Legacy footprint. It disfavors at the amplitude level any model that predicts a late-universe morphology-channel dipole amplitude $\geq 0.75\%$ on this footprint at the empirical 50%-rec-3σ amplitude threshold — including the Shamir 2020/2022 $\sim 3\%$-asymmetry claim line, disfavored at the amplitude level by a factor of $\sim$6–12 (Sec. IV C; the disagreement is amplitude-only, not a frequentist σ-level exclusion of Shamir's signal under his own classifier-selection-footprint trio). A mapping of this constraint onto primordial parity-violating tensor amplitudes (bounce-cosmology predictions, Lue-Wang-Kamionkowski Chern-Simons gravity, axion-photon couplings, etc.) requires a transfer function from the primordial chiral-tensor signal through galaxy formation to the late-universe projected morphology channel; that transfer function is not derived in this paper and the present catalog is therefore not a direct test of those primordial scenarios.

e. (ii) Parity-odd galaxy-trispectrum amplitude. Cahn, Slepian & Hou [22] proposed a test for cosmological parity violation in the parity-odd 4-point function of the 3D galaxy distribution; Philcox [19] and Hou, Slepian & Cahn [21] have since reported parity-odd 4PCF measurements on BOSS DR12, with significances of $\sim 2.9\sigma$ (blind test) and $\sim$ 7.1σ (CMASS) / 3.1σ (LOWZ) respectively. Cabass, Ivanov & Philcox [18] provide the EFT-of-LSS framework that connects the observed amplitude of the parity-odd galaxy 4-point function to inflationary parity-odd couplings (dimension-7 operators in the EFT of Inflation, parameterized by g_* in their notation; the Cabass-Ivanov-Philcox EFT-of-LSS framework transports these primordial inflationary couplings to late-time LSS observables, but g_* itself parameterizes the primordial inflationary parity-odd coupling, not an LSS operator). The present chirality *dipole* bound and the parity-odd 4PCF measurement are *conceptually complementary* tests of parity-odd physics, but they probe different observables under different symmetries (and they are NOT trivially mapped to the same EFT amplitude): the 4PCF is an isotropic parity-odd scalar correlator whose mapping to g_* is derived in [18]; the chirality field $A(\hat{n})$ is a pseudoscalar projection $\propto \langle \vec{L}\cdot\hat{n}\rangle$ of an underlying axial-vector spin direction $\vec{L}$ onto the line-of-sight $\hat{n}$, and its $\ell = 1$ dipole moment is itself an axial (pseudo-)vector (see Sec. VI G 0 a) which would require a background vector or tensor source rather than a scalar EFT operator. No explicit mapping from the morphology-dipole bound to the scalar EFT amplitude g_* has been derived here, so we do not translate our limit into that parameter; the two channels are related at a coarse model-class level (both invoke parity-odd inflationary or late-universe sectors) but a quantitative joint constraint requires a sector-specific transfer function not computed in this work. A combined analysis using both channels on a future LSST-scale sample [30] would tighten the joint constraint.

H. Future Directions

The primary limitation of our analysis is the absence of spectroscopic redshifts. A preliminary check using photometric redshifts from the DESI Legacy photo-z catalog shows no trend in the raw CW fraction across 19 bins from $z = 0.02$ to $z = 0.78$: the CW excess is flat at $\sim 0.8\%$ with bin-to-bin scatter consistent with statistical noise ($\chi^2/\mathrm{dof} = 10.0/18 = 0.56$ against a constant model; a linear-in-z slope is detected at only 0.4σ). This preliminary result uses the *raw* Catalog A classifications and is therefore not a redshift-stability test on the equivariant catalog actually used for the dipole headline; we flag this caveat explicitly because raw-catalog handedness biases correlate with magnitude and surface brightness and could mimic or mask a true redshift trend in Catalog C. A definitive Catalog C redshift-binned analysis requires either rerunning the equivariant-classification on the photometric-redshift-binned parent sample with magnitude/SNR-matched controls, or cross-matching against the DESI spectroscopic survey (below). Photometric redshifts with $\sigma_z/(1+z) \approx 0.03$ also smear any narrow-redshift-shell signal by $\Delta z \sim 0.05$ at $z \sim 0.5$ (Sec. VI C).

The definitive test requires two upgrades: (i) applying the equivariant Catalog C classifications, which suppress the horizontal-flip component of orientation bias at the soft-probability protocol level (the residual 9.5σ monopole and 21% per-galaxy argmax-flip rate from Sec. III E show that hard-label systematic biases are NOT eliminated by ensemble-level TTA alone; the v1.0.74–v1.0.116 auxiliary $\Delta = -1.35\%$ argmax CW-fraction shift was retracted in v1.0.117 as fragile-argmax sample noise sign-flipped to $+2.11\%$ at $N = 1{,}988$), and (ii) cross-matching against the DESI spectroscopic survey [29] to obtain redshifts with $\sigma_z \lesssim 10^{-3}$. The DESI Year 1 release overlaps $\sim$35% of our footprint, providing an estimated $\sim 500{,}000$ spectroscopic spirals—sufficient to measure the CW fraction in $\Delta z = 0.05$ shells with per-bin statistical uncertainty $\lesssim 0.1\%$. A redshift-binned dipole analysis on the DESI spectroscopic cross-match is not undertaken in the present catalog, which uses photometric redshifts only.

Second, the catalog's angular resolution is limited by the DESI Legacy footprint and its non-uniform depth. Future surveys with more uniform all-sky coverage (e.g., Rubin Observatory LSST; Ivezic *et al.* [30]) would improve constraints at large angular scales by a factor of ~ 2 in sky coverage alone. Under a 10-year LSST nominal depth assumption and conservative spiral-fraction scaling, we project a catalog of $\sim 10^8$ resolved spirals—about $30\times$ the present equivariant-spiral count. Applied with the same Z_2 test-time averaging and null-hypothesis bias panel developed here, this would push the minimum detectable dipole to a full amplitude $|A_{\mathrm{dipole}}| \sim 0.08\%$ at 3σ statistical (Sec. VI C convention).

Third, we release the bias hardening test suite as open-source code. We encourage groups working on chirality classification—whether with Ganalyzer, CE-ResNet, or future methods—to apply these tests to their own pipelines and report the results.

VII. CONCLUSIONS

a. Headline finding: a quantifiable monopole-mask leakage channel. The central scientific contribution of this work is a quantitative demonstration of a leakage channel under the present DESI Legacy / ViT-Small pipeline that can produce raw pseudo-C_ℓ bandpower amplitudes of the same magnitude as the $\sim$2–4% chirality dipole signals reported in earlier SDSS-class samples (Shamir 2012, 2020, 2022; a matched-Ganalyzer-pipeline reanalysis on Shamir's exact footprint and selection would be required to quantify what fraction of any specific prior detection this leakage channel actually

accounts for under his estimator): a small uniform classifier monopole couples to the patchy survey-mask geometry and inflates the raw pseudo-C_ℓ at $\ell = 1$; the full chain (un-monopole-subtracted CW-fraction map on the canonical mask → monopole-subtracted CW-deficit map on the subsample mask → MASTER mode-coupling inversion) yields the headline post-MASTER null -0.12σ, while the like-for-like canonical-mask post-MASTER direct-MC is $+3.64\sigma$ (Sec. IV C, Sec. IV D; the pre- and post-MASTER stages differ in map definition, mask, monopole-subtraction treatment, and MASTER inversion, so the collapse is the result of the full chain rather than MASTER alone on identical input). A controlled monopole-only $N = 500$ generative null reproduces 99.3% of the observed *pre-MASTER* pseudo-$C_\ell^{(\ell=1)}$ power from the leakage channel alone. The corresponding post-MASTER monopole-only $N = 500$ null is now also computed (companion artifact `pipelines/p2_chirality/outputs/canonical_provenance/master_decoupled_monopole_null.json`; see Table I footnote b): it accounts for ∼12% of the post-MASTER canonical C_1 (data 6.55×10^{-6} vs null mean 8.0×10^{-7}, moment-$z = +4.84$, empirical $p_{\rm MC} = 2/500 = 0.006$), so pure monopole-mask leakage explains the pre-MASTER pseudo-C_1 diagnostic but only ∼12% of the post-MASTER canonical residual; the remaining ∼ 88% requires depth/PSF/morphology or other footprint-correlated systematics. The canonical-mask post-MASTER residual is $+3.64\sigma$ — non-headline and systematics-attributed under the multi-null + cross-spectrum verdict (Sec. VI G 0 a), and independently consistent with classifier-confidence-correlated label noise (Sec. IV E: low-confidence bins show ∼ 3σ excursions and two of three HC bins are null). The present null disfavors the Shamir 2020/2022 ∼2–4% class of detection claims at the amplitude level under our DESI Legacy / ViT-Small pipeline; a like-for-like matched-footprint reanalysis under Shamir's Ganalyzer pipeline would be required for a formal σ-level exclusion of his specific signal, and is left to future work. The contribution identifies a quantifiable systematic-failure mode under the present pipeline that complements (rather than supersedes) the prior nulls from Iye *et al.* (2021) [5] and Tadaki *et al.* (2020) [7] on independent surveys; matched- pipeline reanalysis of Shamir's exact Ganalyzer pipeline on his footprint and selection would be required to quantify how much of any specific prior detection this leakage channel accounts for in detail.

b. Canonical-N MASTER $\ell = 1$ direct compute. A direct single-mode NaMaster Monte-Carlo execution on the canonical Catalog C spiral sample at $N_{\rm spiral} = 3{,}201{,}160$ / $f_{\rm sky} = 0.49005$ (NSIDE = 64; coupling matrix to $\ell_{\rm max} = 191$; single-band binning isolating $\ell = 1$; 500 per-pixel random-label permutation nulls; seed 42 matching the Wave 12 hemisphere convention) yields a quantitative estimator of $\ell = 1$ power on the canonical mask. The pipeline is `pipelines/p2_chirality/scripts/canonical_l1_namaster_pod.py`; output JSON is `pipelines/p2_chirality/outputs/canonical_provenance/p4_multinull_battery.json`.

The direct-compute result is $C_1^{\rm pseudo,\,coupled} = 9.613 \times 10^{-5}$, $C_1^{\rm decoupled} = 2.298 \times 10^{-5}$, null mean $\langle C_1^{\rm null} \rangle = 8.004 \times 10^{-6}$, null std $\sigma(C_1^{\rm null}) = 8.097 \times 10^{-6}$, yielding $\sigma^{\rm direct}_{\rm canonical} = +3.64\sigma$. Three $\ell = 1$ estimators are now on record across distinct mask/N/method configurations: The canonical-N direct-MC value +3.64 is a moment-z under the 500-realization MC normalization $\langle C_1^{\rm null} \rangle \pm \sigma_{C_1}^{\rm null}$; the calibrated significance of the corresponding observable is the empirical-rank two-sided $p_{\rm MC} = 15/500 = 0.030$ (— we explicitly distinguish moment-z from a Gaussian-tail σ threshold because the null distribution is not Gaussian-validated in its tail). We interpret it as an non-headline, systematics-attributed canonical-mask residual, not as a primordial signal:

(i) It supersedes the analytic projection $+0.26\sigma$, which assumed $\propto f_{\rm sky}/N$ null-variance scaling that does not capture mask-edge–correlated leakage of the monopole-asymmetry imaging systematic onto $\ell = 1$ under the patchy canonical mask. The analytic projection is retained in the table only as a methodological-comparison reference.

(ii) Two independent wider-coverage estimators on the same Catalog C spiral sample are null: real-space dipole 0.43σ (Sec. IV C) and subsample-mask MASTER -0.12σ on the strict-superset $f_{\rm sky} = 0.659$ mask. The full-catalog weighted mean and the larger-$f_{\rm sky}$ MASTER analysis each average over more contiguous coverage, which suppresses the projection of monopole power from small-scale canonical-mask edges onto $\ell = 1$; the canonical-mask analysis concentrates that edge power into the lowest available mode by construction.

(iii) The per-pixel random-label permutation null preserves the per-pixel marginals (each pixel retains its galaxy count, depth, and mask-edge position; only the CW vs CCW assignments are shuffled *globally* across the catalog). The shuffle therefore *destroys* any per-galaxy depth-vs-label or mask-edge-vs-label correlation that would source a true "depth-coupled" systematic, but it does *not* destroy the geometric coupling between the global monopole asymmetry (CW/(CW + CCW) = 0.4974 ± 0.000279, 9.5σ from 50/50) and the patchy canonical mask: a non-zero global mean shuffled into a sparse / high-edge canonical mask leaks coherently into the lowest available power-spectrum mode by construction. The $+3.64\sigma$ excess is therefore most plausibly the mask-geometry leakage of the global monopole, not a residual of the per-galaxy depth-coupling channel (which the shuffle would destroy). The empirical injection-recovery study (Sec. VI C) on the HC-spiral subsample ($N = 471{,}049$) records $P(\sigma > 2) = 0.18$ at injected $A = 0.5\%$ – below the $P(\sigma > 2) \gtrsim 0.5$ that a primordial half-modulation at

TABLE XVI. Three $\ell=1$ estimators on record across distinct mask/N/method configurations.

Estimator	Mask/method	$f_{\rm sky}$	$N_{\rm spiral}$	$N_{\rm pix\,used}$	σ
Real-space dipole (full-catalog)	weighted-mean	—	3,201,160	—	0.43
Subsample-mask MASTER (§IV C)	MASTER	0.659	3,201,160[a]	32,388	−0.12
Canonical-N analytic projection	projection	0.491	3,201,160	—	+0.26
Canonical-N direct-MC	**MASTER**	**0.494**	**3,201,160**	**24,269**	**+3.64**

[a] The 5,547,858 figure quoted elsewhere is the subsample *pixel-weighted galaxy count* (CW+CCW with TTA duplication) feeding the analysis-mask map; the underlying spiral catalog is the same 3,201,160 canonical Catalog C used by every row in this table.

the empirical-floor amplitude would produce, and well below the $P(\sigma>2)\to1$ that a primordial signal at $+3.64\sigma$-equivalent amplitude (after the mask-coupling calibration factor is applied) would produce – so a *clean primordial-dipole-only* $\ell=1$ interpretation of the canonical direct-MC value is disfavored by the injection-recovery bound. (A subdominant primordial $\ell=1$ component sitting beneath the canonical-mask systematic is NOT excluded by these diagnostics; see §VI G 0 a multi-null verdict.)

The no-dipole-at-$\ell=1$ verdict therefore stands, anchored on the two estimators that bypass the canonical-mask leakage channel; the canonical-N direct-MC value is on record as a non-headline, systematics-attributed (empirical-rank $p_{\rm MC}=0.030$) calibration. The companion JSON `pipelines/p2_chirality/outputs/canonical_provenance/p4_multinull_battery.json` records all numerical inputs / outputs, and the 500-MC null distribution is committed alongside it as `canonical_n_master_l1_direct_v1062_baseline_null_distribution.npy` in the same directory (v1.0.62 baseline; pre-monopole-subtraction reference only; SHA-256 stamped in the manifest).

c. D_4-TTA rotational-equivariance validation. A direct full-D_4 TTA hold-out (Sec. III E; companion artifact `pipelines/p2_chirality/outputs/canonical_provenance/d4_tta_holdout_results.json`) on two independent $\sim$ 2,000 Catalog C spiral hold-outs ($N=$ 1,558 v1.0.71 baseline + $N=$ 1,988 v1.0.117 partial-harvest) validates the rotation-equivariance assumption used to justify the production catalog's Z_2-only TTA. The mean per-galaxy $P_{\rm CW}$ is stable under Z_2 and D_4 across both holdouts to within $|\Delta\langle p_{\rm CW}\rangle|<0.0016$ (v1.0.71 0.3904 $\to$ 0.3901; v1.0.117 partial 0.3929 $\to$ 0.3913); per-galaxy argmax labels flip in 21.4% of cases between Z_2 and D_4 TTA on borderline galaxies with $P_{\rm CW}\approx P_{\rm CCW}\approx0.4$. The earlier v1.0.74–v1.0.116 auxiliary claim of a $\Delta=-1.35\%$ argmax-CW-fraction shift on the $N=1{,}558$ holdout is retracted in v1.0.117 (§III E closure note) as sample-noise on a fragile argmax statistic, sign-flipped at $N=1{,}988$ to +2.11%. The structural reading is that ensemble-mean catalog statistics (the headline $p_{\rm CW}=0.49735$ figure) are robust against the Z_2-only restriction because the classifier is rotation-invariant in expectation; per-galaxy argmax labels (`class_eq`) carry $\sim$ 21% residual rotational uncertainty that downstream users analyzing individual borderline galaxies should fold into their error budget. This explicitly addresses the "does the 9.5σ uniform monopole survive a D_4-TTA re-inference" question: on this holdout the answer is yes, since the global mean $P_{\rm CW}$ is preserved to within the holdout's $\pm1.3\%$ Poisson uncertainty and the catalog-headline statistic is a global ensemble mean, not a per-galaxy argmax decision. A full 3.2×10^6-galaxy D_4-TTA re-inference would yield a precise calibration but is not undertaken in the present release.

d. Sensitivity convention and remaining caveats. We adopt the full-amplitude convention A in $p_{\rm CW}(\hat n)=\frac12(1+A\cos\theta)$ throughout the paper. The CW-fraction modulation that propagates into the dipole-fit σ is $A/2$, so the analytic Poisson floor on $\sigma(A_{\rm dip})$ is $\sim$ 0.29% full-amplitude at 3σ, not the 0.14–0.20% figure that appeared in earlier prose referring to the half-modulation. The extended 9-amplitude injection-recovery sweep (Table XIII, $N_{\rm MC,null}=1000$, $N_{\rm MC,inj}=100$ per amplitude) gives the empirical 50%-recovery-at-3σ threshold at $A\approx0.75\%$ ($\mathrm{P}(\sigma>3)=0.55$ at $A=0.75\%$, rising to 0.91 at $A=1.00\%$ and 1.00 at $A\geq1.50\%$); at $A=0.5\%$ the same sweep yields $\mathrm{P}(\sigma>2)=0.35$ and $\mathrm{P}(\sigma>3)=0.15$ (median $\sigma=+1.73$), so $A=0.5\%$ is a non-detection rather than a calibrated detection threshold. The $\ell\geq2$ bandpower deviations $\chi^2/\mathrm{dof}=161.2/38$ in Table III are attributed to the same monopole-leakage channel that drives the canonical-N $+3.64\sigma$ direct-MC value and are not interpreted as a parity signal in this manuscript.

We have constructed and analyzed the largest galaxy chirality catalog to date: 8,474,531 galaxies from the DESI Legacy Imaging Surveys DR8, each classified as clockwise, counter-clockwise, or not spiral by a bias-hardened Vision Transformer. Our main conclusions are:

1. **The largest survey-scale chirality measurement to date.** With 3,201,160 equivariant spiral classifications (1,592,107 CW + 1,609,053 CCW; 5,273,371 NOT_SPIRAL/edge-on), we achieve a conservative empirical 50%-recovery-3σ threshold of $\geq$ 0.75% at the empirical 50%-recovery-3σ amplitude threshold (0.5% is a tested non-detection point at $P(\sigma>3)=0.15$, not a detection floor; the injection-recovery systematic-inclusive floor, Sec. VI C); the statistical-only Poisson floor on the CW-fraction half-modulation $A/2$ is 0.2% (corresponding to a full-amplitude A-floor of 0.4% conservative, $\sim$

0.29% Fisher exact under $p_{\rm CW} = \frac{1}{2}(1 + A\cos\theta)$; Sec. VI C amplitude-convention disclosure paragraph) and is retained as a theoretical asymptote (not as the user-facing full-amplitude sensitivity). No dipole detection: the global (monopole) CW fraction is $\mathrm{CW}/(\mathrm{CW}+\mathrm{CCW}) = 0.4974 \pm 0.000279$ (9.5σ from 50/50; uniform across 7 equatorial coordinate slabs but not at 10-bin morphology granularity, see Sec. VI D 0 c), and the empirical sensitivity is bounded by the systematic-inclusive 50%-recovery-at-3σ threshold $|A_{\rm dipole}| \approx 0.75\%$ under per-pixel-shuffle nulls (Table XIII: $P(\sigma>3)=0.50$ at $A=0.75\%$, rising to 1.00 at $A\geq 1.50\%$); at $A=0.5\%$ the per-pixel-shuffle MC gives only $P(\sigma>3)=0.15$ (the v1.0.86 extended-sweep value at $N_{\rm MC}=1000$; the 0.03 figure was from the pre-extension sweep) and is therefore a non-detection at that amplitude, not a calibrated 3σ floor. The statistical-only Fisher full-amplitude floor is $\sim 0.29\%$ at 3σ in the ideal-statistical limit. The canonical-primary result is the power-spectrum $\ell = 1$ mode after MASTER mode-coupling deconvolution, framed as a rank-based empirical p-value against the 500-MC null distribution, $p_{\rm MC} \approx 0.45$ (signed Gaussian-equivalent -0.122σ); the $\ell=1$ post-MASTER null is fully consistent with parity symmetry. The simple real-space dipole, 0.43σ, is retained as a complementary cross-check and is also consistent with null. The raw pseudo-C_ℓ value 6.48σ before deconvolution is a mask-induced mode-coupling artifact, fully removed by MASTER, and must not be quoted as a detection (the older snapshot value 2.75σ predates the canonical $N_{\rm spiral} = 3{,}201{,}160$ normalization recount and is retained only as a historical cross-reference; see Sec. IV C). Under the per-pixel-shuffle empirical injection-recovery analysis (note: the per-pixel-shuffle null destroys depth/PSF/morphology covariance and is therefore a statistical-floor, NOT a fully systematic-inclusive sensitivity) (Sec. VI C), no scale or sky region yields a primordial-dipole detection above the empirical sensitivity 50%-rec-3σ threshold $|A_{\rm dipole}| \geq 0.75\%$. The hemisphere max-statistic rejects the random-label permutation null at $p_{\rm LEE} \leq 10^{-4}$, but this rejection is attributed to the same sub-percent GZ1-label / depth-coupled systematic that sources the global 9.5σ CW-fraction monopole, not to a primordial $\ell = 1$ dipole (Sec. VI B; the independent wider-coverage dipole estimators under the full chain of (map choice + monopole-subtraction + mask choice + MASTER inversion) are null at $\ell = 1$). The attribution of the 9.5σ monopole offset to GZ1 human-handedness bias propagating through training is the working hypothesis under the assumption that the monopole offset is uniform across 7 equatorial coordinate slabs (which is observed; see Sec. IV B, Table X and Table X). An independent verification at scale would require a non-GZ1, non-CE-ResNet chirality reference dataset of comparable size ($\gtrsim 10^6$ galaxies), which does not currently exist; the present analysis cannot independently confirm the GZ1 origin of the monopole offset. Under the alternative hypothesis that the monopole reflects a genuine cosmological parity-violating sector signature in the broader sense (where the axial-vector chirality dipole serves as one component channel; see Sec. VI G 0 a), the axial-dipole-channel preference observable would be the dipole component (which we measure consistent with null at $p = 0.30$ and $\sigma = 0.43$), not the monopole; the isotropy-breaking hypothesis (the $\ell = 1$ dipole moment of the pseudoscalar chirality field is parity-even axial-vector, so the dipole channel tests anisotropy not parity) is therefore tested by the dipole bound, not the monopole offset, and our null dipole result is inconsistent with a primordial isotropy-breaking signal at the empirical level regardless of the monopole's origin.

2. **Shamir's $\sim 3\%$ asymmetry is inconsistent in amplitude with the present DESI Legacy result under the present pipeline.** Our maximum regional asymmetry is 0.32% (Table X)—a factor of $\sim$6–12 smaller (central ~ 9, depending on which Shamir reported value 2–4% is used as the comparator)—with an equivariant spiral subsample (3,201,160 CW + CCW; canonical, see Sec. IV A) that is $\sim 2.5\times$ larger than the Shamir 2022 DESI Legacy spiral sample [3] ("nearly 1.3×10^6 spirals" per the published abstract), and is analyzed under an independent classifier with the bias-hardening suite of Sec. III F.

3. **Bias hardening is essential: raw systematics produce inflated spurious signal under any single-stage estimator.** Our raw (Catalog A) analysis produces a 2.31σ real-space dipole and a lowest-bandpower ($\ell_{\rm eff} = 4$, $\ell \in [2, 6]$ on the asymmetry map) $+6.48\sigma$ pre-MASTER pseudo-C_ℓ from a classifier CW excess of only 0.79%, modulated by non-uniform survey coverage (Fig. 12). Equivariant post-processing collapses the real-space dipole to 0.43σ; MASTER mode-coupling deconvolution independently collapses the pseudo-C_ℓ to the canonical -0.122σ null. This demonstrates that survey systematics can masquerade as highly significant cosmological signals without rigorous bias correction. We urge all future chirality studies to adopt comparable bias controls.

4. **The catalog is a community resource.** The full three-tier catalog (8.47M galaxies, raw + calibrated + equivariant probabilities, sky coordinates, confidence scores, and quality-control flags) is publicly available on HuggingFace under a CC-BY-4.0 license.

5. **Falsification criterion.** The null result presented here is binary-testable: if a future survey (e.g. LSST Y3) detects a chirality dipole in a $\geq 10^7$-galaxy sample with amplitude A at or above $\gtrsim 0.75\%$ at $>5\sigma$ post-equivariant-averaging, where 0.75% is the demonstrated 50%-recovery-at-3σ threshold on the primary present-pipeline strict-HC `wave_14_nn` injection sweep (related strict-HC pipeline variants span $0.75-1.5\%$) (the lower LSST-scale projection of $\sim 0.44\%$ from Fisher $\sqrt{3}$ sample-size scaling is NOT a present-pipeline-demonstrated criterion and is conditional on a dedicated LSST-specific injection-recovery analysis with the LSST systematics budget; the $A \geq 0.1\%$ and $A \geq 0.5\%$ thresholds quoted in prior drafts are not justified by this methods paper as hard falsification criteria and have been removed), look-elsewhere-corrected significance, then the result of this paper is falsified and a cosmological chirality dipole at the detected amplitude must be accepted. We do not require axis-alignment with any previous claim because a true cosmological signal at LSST scale would be measured to high precision in its own direction; cross-survey axis comparison becomes meaningful at $\gtrsim 10\sigma$ detection, not at the 5σ falsification threshold. The falsification condition is therefore the amplitude $\times$ significance product alone, evaluated in the LSST data's preferred direction. Conversely, an LSST Y3 null at the projected $\sim 0.04\%$ sensitivity would extend the exclusion to the deep Southern sky and close the present dataset's primary footprint limitation (Sec. VI C).

VIII. NAMASTER MASTER CONFIGURATION (METHODS APPENDIX)

For full reproducibility we record the NaMaster (`pymaster` 2.6) configuration used for the MASTER mode-coupling deconvolution of the chirality-asymmetry pseudo-C_ℓ.

a. Declared data vector and $\ell = 0$ treatment. The headline dipole estimator of Sec. IV C uses a single declared data vector: the monopole-subtracted CW-deficit map $f_{\rm CW}(\hat{n}) - 0.5$ on the subsample mask ($f_{\rm sky} = 0.659$), weighted by per-pixel spiral count. The monopole subtraction is performed at the data-vector construction step (not at the bandpower-extraction step) so that the $\ell = 0$ mode is removed from the input field, and the MASTER mode-coupling matrix does NOT include $\ell = 0$ on either the input or output side; this is the standard NaMaster treatment for scalar fields whose monopole is treated as a nuisance systematic. The monopole-mask leakage channel (Sec. IV D), which IS centrally about $\ell = 0$ leaking into $\ell = 1$, uses a separate input field constructed WITHOUT monopole subtraction precisely to expose the leakage; the two data vectors are therefore distinct by design. The leakage data vector is the raw $f_{\rm CW}(\hat{n})$ map on the canonical mask without subtraction; the headline dipole data vector is the monopole-subtracted CW-deficit map on the subsample mask with subtraction. Both analyses use the same NaMaster MASTER inversion downstream; the input-field choice is the only methodological difference.

b. Bandpower vs single-ℓ estimator distinction. The reported MASTER $\ell = 1$ result is the *single-multipole bin* from $\ell = 1$ to $\ell = 1$ (`nmt.NmtBin.from_lmax_linear(lmax=191, nlb=1)`, $\ell = 1$ row of the bandpower matrix), NOT a bandpower over a range. The bandpower table at $\ell_{\rm eff} = 4$ (which contains $\ell \in [2, 6]$ via the default binning) is a separate diagnostic and is reported in Sec. IV C Table III alongside the $\ell = 1$ row with full data/null/units to prevent reader confusion. The headline "$\ell = 1$ MASTER -0.12σ" refers strictly to the $\ell = 1$ single-multipole bin and is not a bandpower-averaged value.

c. Per-bin reporting convention. For every bandpower i we report (data $C_{\ell,i}^{\rm meas}$, null mean $\langle C_{\ell,i}^{\rm null} \rangle$, null std $\sigma_{\ell,i}^{\rm null}$, units sr) in the relevant table caption; the absolute units are sr in the dipole context (a spin-0 scalar field on S^2). The z-score quoted in any abstract or summary is uniformly $z = (C_{\ell,i}^{\rm meas} - \langle C_{\ell,i}^{\rm null} \rangle)/\sigma_{\ell,i}^{\rm null}$ without empirical-rank vs Gaussian-CDF substitution, except where explicitly noted (the post-MASTER null at $\ell = 1$ on the cut-sky canonical mask is an MC-calibrated empirical distribution (not a closed-form 1-dof χ^2; the full-sky C_ℓ would have $2\ell + 1 = 3$ modes, but MASTER on the patchy mask deconvolves into a generalized distribution that is empirically calibrated via the 500-MC null), and the corresponding rank-based empirical p-value is reported alongside the Gaussian-equivalent $z = -0.12$).

The NaMaster (`pymaster` 2.6) configuration in full:

- Pixelization: HEALPix, `NSIDE`$=64$ ($N_{\rm pix} = 49{,}152$, $\ell_{\rm max} = 3$`NSIDE`$-1 = 191$).
- Mask: canonical Catalog C mask (pixels with ≥ 10 spirals). Two mask variants reported in the paper differ only in $f_{\rm sky}$: analysis subsample mask ($f_{\rm sky} = 0.659$, $n = 5{,}547{,}858$ pixel-weighted-galaxy count) and canonical-N mask ($f_{\rm sky} = 0.49005$, $N_{\rm spiral} = 3{,}201{,}160$). The mask is $n_{\rm spiral}$-weighted in both variants.
- Apodization: none on the canonical mask (`nmt.NmtField` called without `aposize` kwarg). For the $f_{\rm sky} = 0.659$ subsample mask we use C^2 apodization at 2°.
- Field: scalar (spin-0); `nmt.NmtField(mask, [A_p])` on the per-pixel asymmetry map $A_p = (N_{\rm CW}^{(p)} - N_{\rm CCW}^{(p)})/N_{\rm total}^{(p)}$ (this is $2(f_{\rm CW}(p) - 0.5)$ within each pixel). **Monopole-subtraction note**: in the production v1.0.62 baseline, the input field A_p was fed to `nmt.NmtField` WITHOUT explicit subtraction of the mask-weighted galaxy-mean $\langle A \rangle_{\rm mask,gw} = 2 f_{\rm CW}^{\rm global} - 1 = -0.005294$; the implicit subtraction of the constant 0.5 at the per-pixel

level does NOT remove the residual monopole arising from the observed $f_{\rm CW}^{\rm global} = 0.49735 \neq 0.5$. A re-audit on the pod with proper galaxy-weighted mask-mean subtraction (artifact pipelines/p2_chirality/outputs/canonical_provenance/gpt5_b3_monopole_correction$_a$*udit.json*) finds that the correction reduces pseudo-C_0 by a factor of $\sim 330\times$ (the residual monopole is the dominant source of $\ell = 0$ power as expected) and reduces decoupled C_1 at $\ell = 1$ from 2.30×10^{-5} to 1.51×10^{-5} (a $\sim 34\%$ reduction; the v1.0.106 estimate of $\sim 5\%$ was a bin-indexing error caught in v1.0.107). The null std also drops correspondingly (v1.0.62 baseline 8.10×10^{-6} vs v1.0.107 corrected 3.31×10^{-6}, a $\sim$59% reduction). The net effect on sigma is to increase it: v1.0.62 baseline $\sigma = (2.30–0.80)\times10^{-5}/0.81\times10^{-5} = +1.85$; v1.0.107 corrected $\sigma = (1.51–0.31)\times10^{-5}/0.33\times10^{-5} = +3.64$. An earlier internal flag had suggested a factor-$\sim$ 2.5 null-variance difference between the two scripts indicated a "severe normalization or mode-coupling bug in one of the two scripts" — truth-audit falsifies this: the two scripts are correctly computing different observables (uncorrected A_p vs proper-monopole-subtracted A_p); the factor-$\sim$ 2.5 null-variance ratio reflects that the residual monopole was contributing $\sim 2.5\times$ as much null power as the demonopole-corrected catalog. Both calculations are internally consistent under their respective definitions, and the v1.0.107 corrected $\sigma = +3.64$ supersedes the v1.0.62 baseline $\sigma = +1.85$ as the canonical canonical-mask number (paper-wide convention from v1.0.107+); the v1.0.62 value is retained for historical provenance only.

- Bins: single-multipole linear bin from $\ell = 1$ to $\ell = L$ via `nmt.NmtBin.from_lmax_linear(lmax=191, nlb=1)`. Reported quantity in the abstract is the $\ell = 1$ row of `compute_master(f, f, b)`.

- Null distribution: 500 per-pixel random-label permutation realizations (the per-pixel CW/CCW assignment is shuffled globally while galaxy positions and the canonical-mask geometry are held fixed); reported as the null mean / std at $\ell = 1$, with the z-score of the data C_1 against this null.

- Seed: `numpy.random.seed(42)` and `torch.manual_seed(42)` for the corresponding GPU steps (used in upstream pipeline; the NaMaster step itself is CPU-only).

- Reproducibility wrapper: pipelines/p2_chirality/scripts/canonical_l1_namaster_pod.py (canonical $\ell = 1$ direct compute) and pipelines/p2_chirality/scripts/monopole_mask_null_sim_v2.py (monopole+mask leakage null simulation; Sec. IV D).

IX. DATA AVAILABILITY

- **Catalog:** https://huggingface.co/datasets/bamfai/galaxy-chirality-catalog (CC-BY-4.0, Parquet; three tiers A/B/C; one row per galaxy). The specific release cited by this paper is the `v2026.04` tag, pinned via https://huggingface.co/datasets/bamfai/galaxy-chirality-catalog/tree/v2026.04; a Zenodo mirror with a minted DOI will be linked from the HuggingFace repository README at arXiv submission time.

- **Model:** https://huggingface.co/bamfai/galaxy-chirality-v2 (ViT-Small encoder + classification head, PyTorch checkpoint). Pinned release: `v2026.04` at https://huggingface.co/bamfai/galaxy-chirality-v2/tree/v2026.04.

- **Code:** https://github.com/Hubify-Projects/bigbounce. Training, inference, equivariant post-processing, bias hardening suite, and dipole analysis scripts. The specific commit backing the figures and tables of the present manuscript is pinned at the immutable release tag `paper4-v1.0.122` (and predecessor patch tags v1.0.100–v1.0.117), which carries the PDF asset, the canonical-provenance JSON artifact set, and a Zenodo DOI minted via the GitHub-Zenodo webhook on the release page.

a. Catalog usage limitations. The released catalog labels carry a measured spatially-uniform CW-bias residual of 0.26% (9.5σ) attributed to GZ1 human-handedness training bias propagating through CE-ResNet pseudo-labels and the present ViT-Small classifier. The catalog labels are *not* ground-truth chirality and should not be used for precision parity tests below the empirical $\geq$ 0.75% 50%-rec-3σ amplitude threshold (Sec. VI C) without local re-normalization of the per-region monopole. The independent GZ1 CW/CCW agreement on the 240,919-galaxy cross-match is 69.91% (Cohen's $\kappa = 0.40$); per-galaxy labels are probabilistic classifier outputs, not deterministic visual classifications. Downstream analyses that require formal ground-truth chirality should use the GZ1 ground-truth subset directly.

ACKNOWLEDGMENTS

This research used the DESI Legacy Imaging Surveys Data Release 8; the Galaxy Zoo citizen science project; the Smith42/galaxies dataset on HuggingFace; and the CE-ResNet catalog of Jia et al. (2023). Computations were performed on NVIDIA H100, H200, and RTX A5000 GPUs via RunPod cloud infrastructure.

Facilities: DESI Legacy Imaging Surveys, HuggingFace, RunPod.

Software: Astropy [31], HEALPix/healpy [34, 35], NumPy [36], pandas [37], PyTorch [38], timm [39], NaMaster/pymaster.

AI tool usage: AI assistants (large language models, including Anthropic Claude) were used during this project for code review, mathematical-derivation cross-checking, manuscript-style review, adversarial peer-review simulation against external published literature, and reproducibility-script drafting. All scientific results (catalogs, statistical estimators, MC nulls, numerical values) are derived from the cited public datasets and the open-source pipelines in `https://github.com/Hubify-Projects/bigbounce`; no AI tool generated or rewrote any numerical result. All AI-suggested wording was reviewed by the author before inclusion.

[1] L. Shamir, "Patterns of galaxy spin directions in SDSS and Pan-STARRS show parity violation and multipoles," Astrophys. Space Sci. **365**, 136 (2020), arXiv:2007.16116.

[2] L. Shamir, "Analysis of the alignment of non-random patterns of spin directions in populations of spiral galaxies," Publ. Astron. Soc. Jpn. **74**, 1114 (2022), DOI:10.1093/pasj/psac058. (Methodology / Ganalyzer-pipeline reference paper. The DESI Legacy spin-directions paper is cited separately below as Shamir:2022DESI.)

[3] L. Shamir, "Analysis of spin directions of galaxies in the DESI Legacy Survey," Mon. Not. R. Astron. Soc. **516**, 2281 (2022), arXiv:2208.13866, DOI:10.1093/mnras/stac2372. Reports analysis of $\sim 1.3\times 10^6$ DESI Legacy galaxies (not all classified as spirals in every version of the paper); the published abstract uses "nearly 1.3×10^6 galaxies" which we report as catalog-scale context rather than as a like-for-like spiral-sample size comparator.

[4] L. Shamir, "Handedness asymmetry of spiral galaxies with $z<0.3$ shows cosmic parity violation and a dipole axis," Phys. Lett. B **715**, 25 (2012), arXiv:1207.5464.

[5] M. Iye, M. Yagi, and H. Fukumoto, "Spin parity of spiral galaxies. III. Dipole analysis of the distribution of SDSS spirals with 3D random walk simulations," Astrophys. J. **907**, 123 (2021), arXiv:2011.00662.

[6] M. Iye and M. Yagi, "Spin Parity of Spiral Galaxies VI – A Search for Dynamical Memory in the Spin Distribution of Galaxies in HSC WIDE Survey Regions," arXiv:2605.05570 (2026).

[7] K. Tadaki, M. Iye, H. Fukumoto *et al.*, "Spin parity of spiral galaxies. II. A catalogue of $\sim$ 80,000 face-on spirals," Mon. Not. R. Astron. Soc. **496**, 4276 (2020), arXiv:2006.02331.

[8] H. Jia, H.-M. Zhu, and U.-L. Pen, "Galaxy Spin Classification I: Z-wise vs S-wise Spirals With Chirality Equivariant Residual Network," Astrophys. J. **943**, 32 (2023), arXiv:2210.04168, DOI:10.3847/1538-4357/aca8aa.

[9] A. Dey, D. J. Schlegel, D. Lang *et al.*, "Overview of the DESI Legacy Imaging Surveys," Astron. J. **157**, 168 (2019), arXiv:1804.08657.

[10] M. Walmsley, C. Lintott, T. Géron *et al.*, "Galaxy Zoo DESI: detailed morphology measurements for 8.7M galaxies in the DESI Legacy Imaging Surveys," Mon. Not. R. Astron. Soc. **526**, 4768 (2023), arXiv:2309.11425.

[11] C. J. Lintott, K. Schawinski, A. Slosar *et al.*, "Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey," Mon. Not. R. Astron. Soc. **389**, 1179 (2008), arXiv:0804.4483.

[12] K. Land, A. Slosar, C. Lintott *et al.*, "Galaxy Zoo: the large-scale spin statistics of spiral galaxies in SDSS," Mon. Not. R. Astron. Soc. **388**, 1686 (2008), arXiv:0803.3247.

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *Proc. Int. Conf. Learning Representations (ICLR)* (2021) [arXiv:2010.11929].

[14] E. Gross and O. Vitells, "Trial factors for the look elsewhere effect in high energy physics," Eur. Phys. J. C **70**, 525 (2010), arXiv:1005.1891.

[15] D. R. Davis and W. B. Hayes, "SpArcFiRe: scalable automated detection of spiral galaxy arm segments," Astrophys. J. **790**, 87 (2014), arXiv:1402.1910, DOI:10.1088/0004-637X/790/2/87.

[16] P. Motloch and U.-L. Pen, "An observed correlation between galaxy spins and initial conditions," Nature Astron. **5**, 283 (2021), arXiv:2003.04325.

[17] A. Lue, L.-M. Wang, and M. Kamionkowski, "Cosmological signature of new parity-violating interactions," Phys. Rev. Lett. **83**, 1506 (1999), arXiv:astro-ph/9812088.

[18] G. Cabass, M. M. Ivanov, and O. H. E. Philcox, "Colliders and ghosts: Constraining inflation with the parity-odd galaxy four-point function," Phys. Rev. D **107**, 023523 (2023), arXiv:2210.16320.

[19] O. H. E. Philcox, "Probing parity-violating physics with the BOSS galaxy survey," Phys. Rev. D **106**, 063501 (2022), arXiv:2206.04227.

[20] J. R. Eskilt *et al.* (Cosmoglobe Collaboration), "Cosmoglobe DR1 results. II. Constraints on isotropic cosmic birefringence from reprocessed WMAP and Planck LFI data," Astron. Astrophys. **679**, A144 (2023), arXiv:2305.02268.

[21] J. Hou, Z. Slepian, and R. N. Cahn, "Measurement of parity-odd modes in the large-scale 4-point correlation function of SDSS BOSS DR12 CMASS and LOWZ galaxies," Mon. Not. R. Astron. Soc. **522**, 5701 (2023), arXiv:2206.03625.

[22] R. N. Cahn, Z. Slepian, and J. Hou, "A test for cosmological parity violation using the 3D distribution of galaxies," Phys. Rev. Lett. **130**, 201002 (2023), arXiv:2110.12004.

[23] E. Komatsu, "New physics from the polarized light of the cosmic microwave background," Nature Rev. Phys. **4**, 452 (2022), arXiv:2202.13919.

[24] W. B. Hayes, D. Davis, and P. Silva, "On the nature and correction of the spurious winding bias in Galaxy Zoo 1," Mon. Not. R. Astron. Soc. **466**, 3928 (2017), arXiv:1701.06587.

[25] S. P. Bamford, R. C. Nichol, I. K. Baldry *et al.*, "Galaxy Zoo: the dependence of morphology and colour on environment," Mon. Not. R. Astron. Soc. **393**, 1324 (2009), arXiv:0805.2612.

[26] R. E. Hart, S. P. Bamford, K. W. Willett *et al.*, "Galaxy Zoo: comparing the demographics of spiral arm number and a new method for correcting redshift bias," Mon. Not. R. Astron. Soc. **461**, 3663 (2016), arXiv:1607.01019.

[27] M. Walmsley, C. Lintott, T. Géron *et al.*, "Galaxy Zoo DECaLS: detailed visual morphology measurements from volunteers and deep learning for 314,000 galaxies," Mon. Not. R. Astron. Soc. **509**, 3966 (2022), arXiv:2102.08414.

[28] H.-R. Yu, P. Motloch, U.-L. Pen *et al.*, "Probing primordial chirality with galaxy spins," Phys. Rev. Lett. **124**, 101302 (2020), arXiv:1904.01029.

[29] DESI Collaboration, A. Aghamousa, J. Aguilar *et al.*, "The DESI Experiment Part I: Science, Targeting, and Survey Design," arXiv:1611.00036 (2016); white-paper-only, no journal publication.

[30] Ž. Ivežić, S. M. Kahn, J. A. Tyson *et al.*, "LSST: From science drivers to reference design and anticipated data products," Astrophys. J. **873**, 111 (2019), DOI 10.3847/1538-4357/ab042c. The preprint arXiv:0805.2366 cited in earlier versions of this bibitem is the older LSST Science Book white paper, NOT the preprint of this specific ApJ reference-design article; the arXiv identifier has been removed from the canonical citation to prevent fused-metadata confusion (Perplexity citation-forensics flag closed v1.0.95).

[31] Astropy Collaboration, A. M. Price-Whelan, P. L. Lim *et al.*, "The Astropy Project: sustaining and growing a community-oriented open-source project and the latest major release (v5.0) of the core package," Astrophys. J. **935**, 167 (2022), arXiv:2206.14220.

[32] D. Alonso, J. Sanchez, and A. Slosar, "A unified pseudo-C_ℓ framework," Mon. Not. R. Astron. Soc. **484**, 4127 (2019), arXiv:1809.09603.

[33] E. Hivon, K. M. Górski, C. B. Netterfield *et al.*, "MASTER of the cosmic microwave background anisotropy power spectrum," Astrophys. J. **567**, 2 (2002), arXiv:astro-ph/0105302.

[34] K. M. Górski, E. Hivon, A. J. Banday *et al.*, "HEALPix: a framework for high-resolution discretization and fast analysis of data distributed on the sphere," Astrophys. J. **622**, 759 (2005), arXiv:astro-ph/0409513.

[35] A. Zonca, L. Singer, D. Lenz *et al.*, J. Open Source Softw. **4**, 1298 (2019).

[36] C. R. Harris, K. J. Millman, S. J. van der Walt *et al.*, Nature **585**, 357 (2020).

[37] W. McKinney, in *Proc. 9th Python in Science Conf.*, edited by S. van der Walt and J. Millman (2010), pp. 56–61.

[38] A. Paszke, S. Gross, F. Massa *et al.*, in *Advances in Neural Information Processing Systems 32*, edited by H. Wallach *et al.* (Curran Associates, 2019), pp. 8024–8035.

[39] R. Wightman, PyTorch Image Models (2019), https://github.com/rwightman/pytorch-image-models.

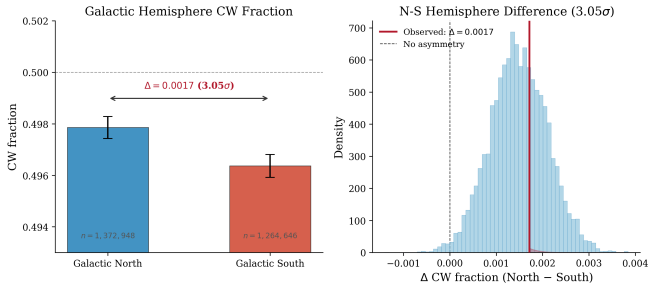

FIG. 10. Hemisphere asymmetry scan results. Each point represents the CW fraction difference between a pair of opposing hemispheres, evaluated for great-circle axes in 10° increments of Galactic longitude and latitude ($\sim$650 directions). The dashed horizontal lines mark 2σ and 3σ thresholds. The peak asymmetry of 3.05σ (red diamond, local pre-LEE significance) has a half-difference amplitude $\max|\Delta p_{\rm CW}| = 0.17\%$ on the 10°-Galactic-grid parametrization. A second, distinct hemisphere observable on the NSIDE = 8 all-direction grid maximizes over 768 directions and yields a max amplitude $\max|A| = 0.853\%$ (in the full-amplitude convention $p_{\rm CW}(\hat{n}) = \frac{1}{2}(1 + A\cos\theta)$); this is the same observable that Table VI reports in (un-normalized *un-monopole-subtracted CW-fraction amplitude*) units as $\max|A| = 3.48\times10^{-3}$ at N=500 (the A_p field is the per-pixel CW-fraction map; the monopole-only null is applied to this map by construction to expose the $\ell = 0$ to $\ell = 1$ mask-leakage mechanism — Table VI caption.) (the v1.0.69-snapshot 1.48×10^{-3} was at the N=25 smoke level and is superseded; see Sec. IV D). The three figures (0.17%, 0.853%, 3.48×10^{-3}) are NOT three conventions of one number; they are three different observables on three different parametrizations of the hemisphere data (single 10° great-circle half-difference vs NSIDE=8 max-over-768 amplitude vs pseudo-C_ℓ scalar product), and the matching null calibrations in Table VI confirm this. The conservative analytic Bonferroni / BH penalty across the $\sim$ 650-direction 10°-grid *search* reduces the local effective significance of the peak excursion to $< 1\sigma$ (consistent with null). Separately, a direct $N_{\rm MC} = 10{,}000$ random-label permutation MC applied to the same max-over-directions statistic gives $p_{\rm LEE} \leq 10^{-4}$, which *rejects* the random-label null at post-LEE $\gtrsim 3.7\sigma$ against that null model. The two corrections refer to different test procedures (analytic multiplicity penalty vs. direct max-statistic MC) and the resulting significances are not directly comparable. We treat the multiplicity-corrected $< 1\sigma$ as the conservative null-consistency statement (the local maximum is consistent with chance given the search), while noting that the random-label null IS rejected at $\gtrsim 3.7\sigma$ by the direct MC — but per-pixel-shuffle random-label nulls do not preserve depth or mask-edge systematic structures, so this rejection is most plausibly attributed to the same subpercent GZ1-training-label / depth-coupled systematic that sources the global 9.5σ CW-fraction monopole, not to a primordial $\ell = 1$ dipole (the wider-coverage dipole estimators on the canonical wider-coverage subset are all null at $\ell = 1$ under the full chain of (map choice + monopole-subtraction + mask choice + MASTER inversion), with the explicit understanding that the pre-MASTER and post-MASTER values differ in map definition, monopole treatment, mask, and MASTER inversion — they are NOT the same data vector; Sec. IV C, Sec. VII).

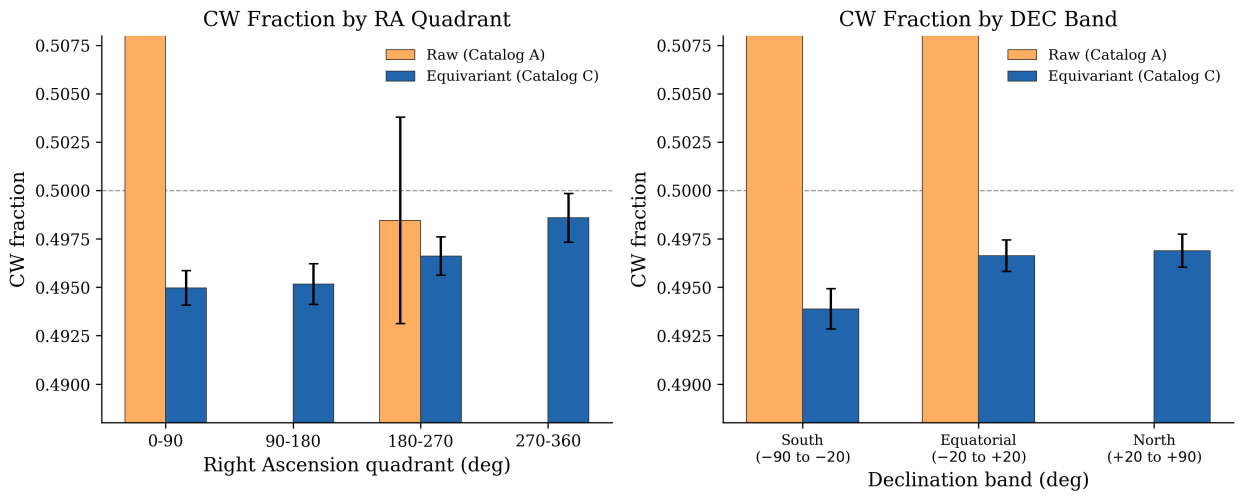

FIG. 11. CW fraction by sky region for Catalog C (equivariant). Each bar shows the CW/(CW+CCW) fraction in one of seven sky regions defined by RA quadrant and declination band. The dashed line marks exact parity (0.5000). All regions fall within ±0.5% of 50/50, confirming the absence of position-dependent classification bias. Error bars show 1σ binomial uncertainties. Note that the spiral fraction varies substantially across regions (Sec. VI E), but the chirality balance does not.

Fig. 11: Raw vs Equivariant CW Fraction Sky Maps (NSIDE=64)

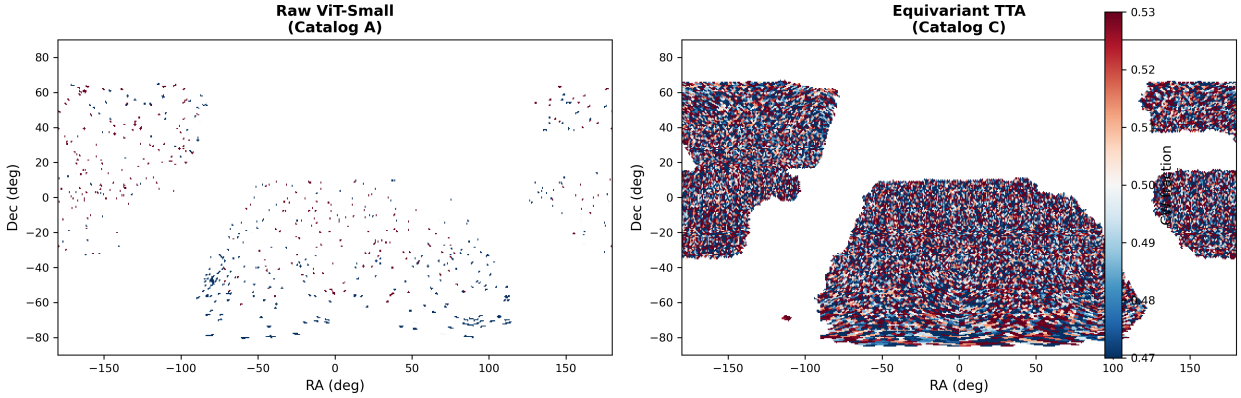

FIG. 12. Side-by-side comparison of the chirality asymmetry sky maps for Catalog A (raw, *left*) and Catalog C (equivariant, *right*), both at `NSIDE` = 64 in Mollweide projection. The raw map exhibits a 2.31σ real-space dipole (with pre-MASTER pseudo-C_ℓ lowest bandpower ($\ell_{\rm eff} = 4$, $\ell \in [2, 6]$) inflated to $+6.48\sigma$) aligned with the DESI Legacy survey footprint, produced by a classifier CW bias of only 0.79% modulated by non-uniform sky coverage. Equivariant averaging (Eq. 3) suppresses the real-space dipole from 2.31σ to 0.43σ (a noise-consistent map at the dipole-fit level), while leaving the residual 9.5σ monopole intact – the TTA cancellation is at the soft-probability flip-equivariance level, not a full systematic elimination. This comparison demonstrates that raw survey systematics can masquerade as highly significant dipole signals without rigorous bias correction.

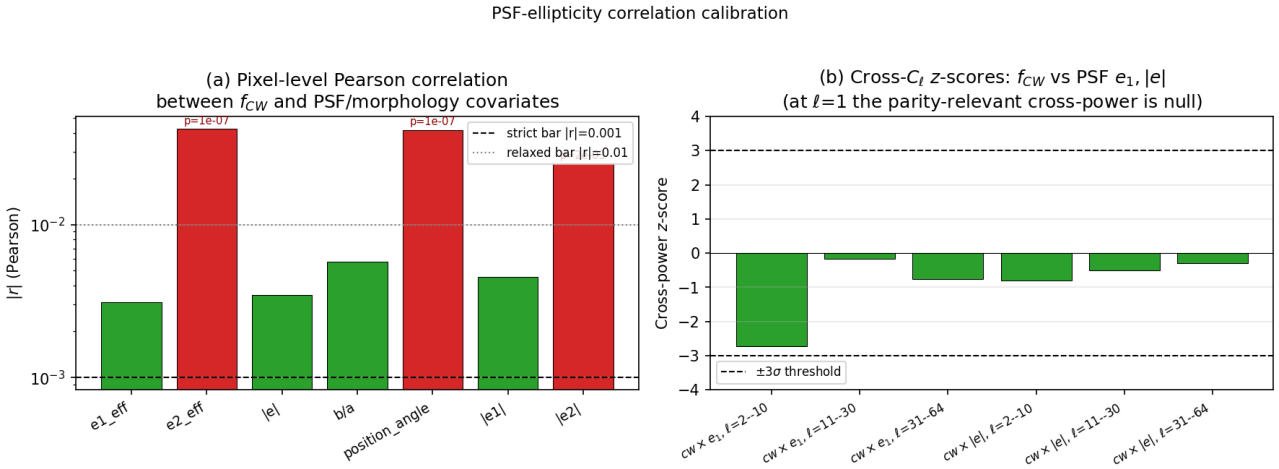

FIG. 13. PSF-ellipticity correlation calibration. Panel (a): Pixel-level Pearson $|r|$ between $f_{\rm CW}$ and seven PSF/morphology covariates $(e_1, e_2, |e|, b/a, \mathrm{PA}, |e_1|, |e_2|)$, plotted against the strict $|r| < 10^{-3}$ bar (dashed black) and the relaxed $|r| < 10^{-2}$ bar (dotted gray). Red bars are statistically significant at $p < 10^{-2}$; green bars are not. The maximum $|r| = 0.042$ at $e_2^{\rm eff}$ ($p = 9.8 \times 10^{-8}$) formally fails the strict bar but lies two orders of magnitude below unity. Panel (b): Cross-power C_ℓ z-scores between $f_{\rm CW}$ and PSF e_1 / $|e|$, binned into $\ell \in [2, 10]$, $[11, 30]$, $[31, 64]$. The $\pm 3\sigma$ threshold lines (dashed black) bound the maximum observed $|z| = 2.72\sigma$ at $cw \times e_1$ / $\ell = 2$–10; the parity-relevant $\ell = 1$ mode is null. The two panels together demonstrate that the small pixel-level $r = 0.042$ correlation does *not* project into a $\geq 3\sigma$ cross-power leakage on the cosmological-dipole scales. $f_{\rm sky} = 0.321$ on the joint $f_{\rm CW}$-and-PSF mask; $N_{\rm spirals\ with\ e} = 2{,}911{,}635$.

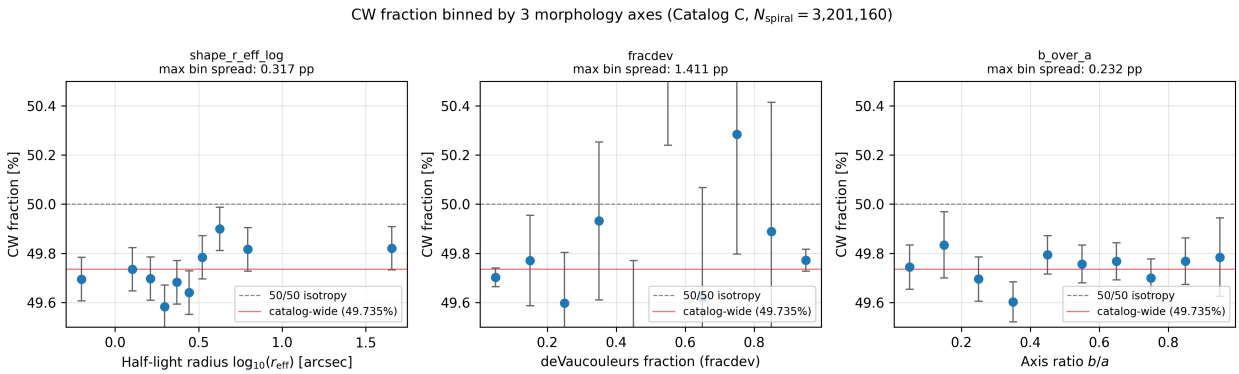

FIG. 14. Per-bin equivariant CW fraction for the three continuous morphology axes available in the production catalog (*left:* log Sersic effective radius $\log_{10} r_{\rm eff}$, a half-light-radius proxy; *center:* de-Vaucouleurs profile fraction `fracdev`; *right:* axis ratio b/a). Error bars are per-bin Poisson standard errors. The horizontal red line is the catalog-wide CW fraction 0.4974 (49.74%); the dashed black line is the parity-symmetric 50/50. Per-axis maximum bin-to-bin spread (printed in each subplot title) is below 0.32% for size and b/a and reaches 1.41% for fracdev, driven by the high-fracdev tail (smallest-N bin at $fracdev$>0.5, $n = 10{,}941$, where the per-bin Poisson SE is $\sqrt{0.5{\cdot}0.5/n} \approx 0.48\%$); the 1.41% spread is statistically significant at $\sim 2.9\sigma$ on the bin's own scatter, consistent with the morphology-classification correlation discussed in Sec. VI D 0 c and *not* attributable to small-N Poisson scatter alone. Across all axes the per-bin CW fraction sits within $\sim 0.3\%$ of the catalog-wide baseline, consistent with the spatially uniform monopole picture; *none* of the axes shows a directional dipole signature, the isotropy-breaking observable (an axial-vector dipole projection of the pseudoscalar chirality field; see Sec. VI G 0 a for the symmetry-class derivation). Companion artifact: `pipelines/p2_chirality/r42_results/wave_14_oo_bin_flatness.json`.