A Null Chirality Dipole in 8.5 Million DESI Galaxies from Equivariant Deep Learning

Houston Golden[1, *]

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

(Dated: July 12, 2026)

We measure the large-scale chirality dipole of spiral galaxies and find it consistent with null. The measurement rests on the largest chirality-labeled galaxy catalog to date: 8,474,531 DESI Legacy DR8 galaxies classified by a flip-equivariant Vision Transformer into clockwise (CW), counter-clockwise (CCW), and non-spiral classes, yielding $N_{\rm spiral}$ = 3,201,160 spirals, released publicly with model weights and reproducibility scripts. Our primary estimator — a real-space dipole fit to the pre-specified high-confidence subsample ($p_{\rm eq} > 0.6$, $N \approx 9.5 \times 10^5$ spirals) — gives $+0.41\sigma$ (moment-z against an isotropic pixel-permutation null; empirical-rank $p = 0.31$), robust across the confidence-cut sweep and under a per-galaxy label-shuffle test; a block-bootstrap WLS template fit disfavors a clean $A_{\rm ref} = 0.017$ dipole (the lower end of Shamir's reported 1.7%–4.0% asymmetry range [1, 3], which maps to our $A_p = 2(f_{\rm CW} - \frac{1}{2})$ without rescaling) at $z \approx -7.6$. This $\ell = 1$ observable is parity-*even* (isotropy-breaking, not a direct parity-violation test). Injection–recovery brackets the real-space 95%-recovery amplitude A_{95} in $(1.0\%, 1.5\%]$ ($A_{50} \approx 0.75\%$), so a genuine Shamir-scale dipole would have been detected; a detection at $\geq 5\sigma$ with $A \gtrsim A_{95}$ would falsify this null. The MASTER pseudo-C_ℓ harmonic channel is treated as a systematics diagnostic, not an independent cosmological null (monopole–mask leakage reproduces 99.32% of raw pre-MASTER $\ell = 1$ power; residuals are attributed via an eight-anchor battery, with $\sim$47% of the post-MASTER amplitude an explicit open item that lies below A_{95} and does not affect the primary null; Sec. IV D, Appendix D). Significances from distinct null procedures are not directly comparable as detection significances (Sec. III A); a leg- and confidence-stratified GZ1 human-label cross-check confirms the classifier's dipole-biasing error asymmetry is consistent with zero in every stratum (Appendix B).

CONTENTS

I. Introduction 1

II. Data 2
A. Galaxy Images 2
B. Training Labels 2

III. Methods 3
A. Notation and Significance Conventions 3
B. Declared Analysis Hierarchy 4
C. Model Architecture 5
D. Test-Time Equivariant Averaging 5
E. Catalog Tiers 6

IV. Results 6
A. Catalog Statistics 6
B. Global CW Fraction 8
C. Dipole Analysis 9
D. Monopole+Mask Leakage Generative Null 12
E. Signal-Hunt Diagnostics 16

V. Comparison with Previous Work 16
A. Shamir (2012, 2020, 2022) 16
B. CE-ResNet (Jia et al. 2023) 16

VI. Discussion 17
A. Pseudo-label independence and the shuffle-null limitation 17
B. Sensitivity Floor and Minimum Detectable Signal 19
C. Relation to Parity-Violating Sectors 21
D. Open Follow-up and Future Directions 22

VII. Conclusions 22

A. NaMaster MASTER Configuration 24

B. Classifier Architecture Details 25

C. Auxiliary Dipole Diagnostics 28

D. Canonical-Mask Systematic Analysis 29

E. Morphology Systematics 32

Data Availability 33

Acknowledgments 34

References 34

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

* houston@hubify.com

that, under the trailing-arm assumption and in the absence of confounding selection effects, traces the angular-momentum direction of each disk galaxy. Throughout this paper, "CW/CCW" refers to the *projected apparent arm-winding chirality*, not a deprojected 3D spin vector. In a statistically isotropic and parity-symmetric universe, the CW and CCW fractions should be exactly equal when averaged over large angular scales. A significant directional departure would constrain isotropy-breaking axial-vector sectors in the galaxy-formation chain. The present paper is a standalone observational result: our null dipole at sub-percent sensitivity does not depend on any unpublished companion work.

Claims of such a signal have appeared intermittently in the literature. Shamir (2012) [4] reported a 2–4σ dipole with per-bin asymmetry amplitudes of $\sim$ 5–20% (as reported in that work) using $\sim 1.27\times10^5$ SDSS galaxies. Shamir (2020) [1] reported asymmetries at the reported $\sim$ 2–4% level on SDSS and Pan-STARRS samples; Shamir (2022a) [2] reported related spin-direction alignment analyses; and Shamir (2022b) [3] reported results on a DESI Legacy sample ("nearly 1.3×10^6 spiral galaxies" per the published abstract). Iye *et al.* (2021) [5] re-examined Shamir's SDSS spiral catalog using 3D random-walk simulations and found no significant dipole after correcting for reading-direction bias and photometric-object duplication in earlier Shamir catalogs. Tadaki *et al.* [6] likewise found null results on a catalog of $\sim$80,000 face-on spirals. Jia *et al.* [7] introduced CE-ResNet, a chirality-equivariant CNN guaranteeing by construction that flipping an input exactly swaps CW and CCW outputs, with a reported number-count ratio CW/CCW = 0.998 on $\sim$1.95 million galaxies.

In this paper we present a new chirality catalog whose novelty relative to CE-ResNet lies in scale and bias-hardening rather than classifier accuracy: (i) survey-scale coverage of 8.47 million galaxies (3,201,160 equivariant-classified spirals, 1.6× CE-ResNet's scale); (ii) a dedicated NOT_SPIRAL class preventing contamination from ellipticals and irregulars; and (iii) a multi-axis bias-hardening audit suite. (A substantial fraction of our training labels derive from CE-ResNet predictions, so the two catalogs are not fully independent; Sec. II and Appendix B quantify the independent GZ1 agreement. A corollary limitation: the label-shuffle and per-pixel permutation nulls used throughout randomize this model's own outputs, so they do not by themselves test independence from any large-scale survey-correlated structure potentially inherited through the CE-ResNet pseudo-labels; that axis is constrained instead by the template-regression and cross-spectrum diagnostics of Appendix D.) We use this catalog to perform a chirality dipole measurement with an empirical 50%-recovery-3σ injection-recovery threshold at $|A_{\rm dipole}| \geq 0.75\%$. The measured dipole is consistent with null: the equivariant CW fraction is 0.4974 ± 0.000279 and the real-space dipole significance is $+0.41\sigma$ ($p = 0.31$ vs. the isotropic pixel-permutation null; primary), with a block-bootstrap WLS template fit disfavoring a clean 1.7% dipole at $z \approx -7.6$ (a model-dependent template-disfavor statistic under the adopted NSIDE= 8 block-bootstrap spatial covariance, *not* a calibrated frequentist exclusion significance; the primary real-space estimator's detection efficiency is instead characterized by the injection-recovery bracket $A_{95} \in (1.0\%, 1.5\%]$, a 95%-recovery threshold rather than a frequentist confidence upper limit, Sec. VI B).) The MASTER pseudo-C_ℓ channel on the patchy footprint is systematics-dominated and serves only as a diagnostic — its post-MASTER residuals ($+3.64\sigma$ canonical, $+7.28\sigma$ apodized) are attributed to residual survey systematics and are *not* detections (Secs. IV C, IV D). This places our amplitude in $\sim$ 3.7–8.8× tension with Shamir's claimed $\sim$ 2–4% signal under the present pipeline — an amplitude-level comparison, not a frequentist exclusion of Shamir's distinct Ganalyzer estimator (comparing both amplitudes in the same A_p units: our canonical joint nuisance-marginalized WLS best-fit is 0.455% in A_p units, Appendix D, while Shamir's reported 1.7%–4.0% is already a full-count *asymmetry* $(N_{\rm CW}-N_{\rm CCW})/(N_{\rm CW}+N_{\rm CCW})$, identical to our $A_p = 2(f_{\rm CW}-0.5)$, so it is 1.7%–4.0% in A_p units directly — no factor-of-two rescaling); a matched-footprint Ganalyzer reanalysis is required for a likelihood-level exclusion.

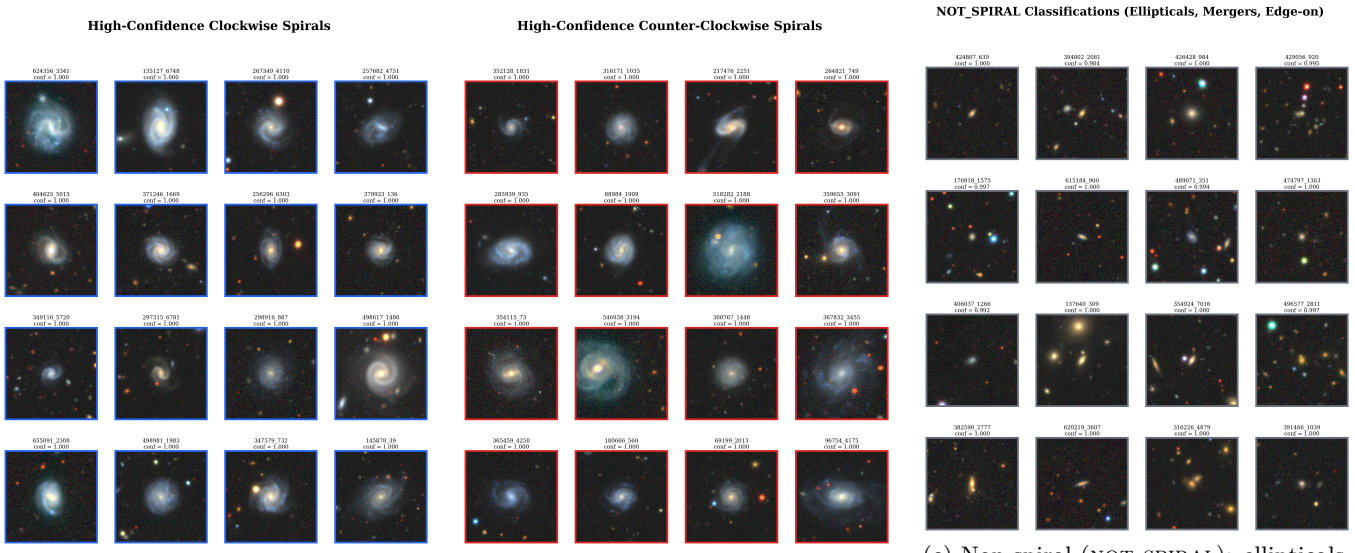
II. DATA

A. Galaxy Images

Our parent sample is the Smith42/galaxies dataset on HuggingFace (`https://huggingface.co/datasets/Smith42/galaxies`), containing 8,474,688 galaxy images from the DESI Legacy Imaging Surveys DR8 [8]. Each image is a 224×224 pixel cutout in grz bands at $0.262''$/pixel. The dataset includes unique `dr8_id` identifiers; sky coordinates are obtained by cross-matching against the Galaxy Zoo DESI predictions catalog [9]. The parent-sample selection function inherits from Galaxy Zoo DESI: photometric types REX/DEV/EXP/SER, $r \leq 19.0$, half-light radius $\geq 3''$. DR8 comprises three distinct imaging campaigns: BASS+MzLS ($\delta > +32°$), DECaLS ($\delta < +32°$), and a DES overlap region.

B. Training Labels

We assemble training labels from three sources: (1) Galaxy Zoo 1 [10]: 6,637 galaxies with CW/CCW labels at $>$ 70% vote confidence; (2) CE-ResNet [7]: 17,153 galaxies with high-confidence spiral classifications; (3) Synthetic hard negatives: 2,000 artificial images as NOT_SPIRAL training examples. The combined training set contains $6{,}637 + 17{,}153 + 2{,}000 = 25{,}790$ source images; after flip augmentation of the training split the combined pool is 26,616 images (pre-augmentation 79.4/20.6



(a) Clockwise (CW) (b) Counter-clockwise (CCW) (c) Non-spiral (NOT_SPIRAL): ellipticals, mergers, edge-on

FIG. 1. **Representative high-confidence galaxies from the classified catalog** ($p_{\rm eq} > 0.9$). Left: clockwise (CW) spirals; center: counter-clockwise (CCW) spirals; right: non-spiral (NOT_SPIRAL) objects — ellipticals, mergers, and edge-on galaxies that would contaminate a binary spiral classifier. All cutouts are 224×224 pixels in grz bands from DESI Legacy DR8. The gallery illustrates the visual diversity within each class and motivates the three-class design: without an explicit NOT_SPIRAL class, $\sim$62% of the parent sample would leak into the spiral classification. The $ViT - Small$ classifier resolves CW vs. CCW via the test-time equivariant averaging procedure of §III D.

source-image split: $n_{\rm train} = 21{,}293$ post-augmentation, $n_{\rm val} = 5{,}323$ never augmented; the 826-image difference between the source manifest (25,790) and the combined pool (26,616) arises entirely from horizontal-flip augmentation applied to the *training split only* — the validation split ($n_{\rm val} = 5{,}323$) is never augmented; Appendix B, artifact `pipelines/p2_chirality/outputs/canonical_provenance/c17_item13_training_semantics.json`). Note: 17,153/25,790 = 66.5% of training labels derive from CE-ResNet predictions; validation metrics against the full training set therefore partially reflect agreement with CE-ResNet rather than independent ground truth. The independent GZ1 cross-match on 234,282 disjoint matches yields spiral-chirality accuracy 69.91% (Cohen's $\kappa = 0.40$). We treat 69.91% as an explicitly conservative accuracy floor and propagate it to all downstream isotropy bounds via the sub-percent systematic floor in Sec. IV C; because a moderate-accuracy classifier dilutes any real chirality dipole toward null (the dilution $g = 2a - 1$ is folded into the empirical injection-recovery floors of Sec. VI B, not assumed away), this floor renders the present null detection conservative rather than optimistic. *Independence cross-check:* Sec. VI A presents a fully model-free GZ1-human-only dipole test ($N = 4.60 \times 10^4$ confident GZ1 CW/CCW votes cross-matched to DESI, no learned model in the label chain) that returns the same clean null at $z = -0.54\sigma$ — an important model-independent cross-check against pseudo-label inheritance. *Scope note for downstream users:* the GZ1 catalog, with its 69.91% chirality accuracy ($\kappa = 0.40$), is an excellent resource for large-scale null tests of the kind performed here, but should *not* be used as precision per-galaxy chirality ground truth in applications demanding accuracy significantly above this floor. *Independence cross-check:* taking the raw Galaxy Zoo 1 human CW/CCW votes [12] as per-galaxy labels with *no learned model whatsoever* in the chirality-label chain gives $z = -0.54\sigma$ at $N = 4.60 \times 10^4$ ($3.08\times$ the previous GZ1-only test; fully exhausts the confident GZ1×DESI overlap), the same clean null returned by the full equivariant catalog (Sec. VI A). This human-label dipole null is a useful model-independent cross-check against pseudo-label inheritance: no learned model, same null result. Its sensitivity is coarse, however ($A_{50} \approx 3.4\%$, $A_{95} \approx 4.5$–6.8% at $N = 4.60 \times 10^4$; Sec. VI A), so it does not test the sub-percent inherited structure constrained by the headline sample; it can corroborate the null but not tighten it.

III. METHODS

A. Notation and Significance Conventions

a. Confidence notation. $p_{\rm eq}$ (or $p^{\rm eq}_{\rm CW}$, $p^{\rm eq}_{\rm CCW}$, $p^{\rm eq}_{\rm NS}$) denotes the max-class equivariant probability output by the 2-fold flip-TTA pipeline (Sec. III D and Appendix B). These are monotone ranking scores, not frequentist probabilities; "high-confidence" (HC) cuts on $p_{\rm eq}$ are sample-selection thresholds.

b. Significance conventions. All significance values z are moment-ratios $(x - \langle x\rangle_{\rm null})/\sigma_{\rm null}$ against the null distribution specified per result. Empirical rank p-values are one-sided unless labeled two-sided. Three distinct significance conventions appear in this paper:

- *Moment-z / rank-p*: the real-space dipole pair ($z_{\rm mom} = +0.41$, rank-$p = 0.31$) are independent summary statistics of the $A_{\rm dip}$ distribution against the isotropic (pixel-)permutation null; they do not follow the Gaussian $z \to p$ mapping.
- *MASTER $\ell = 1$ moment-z*: $z = (C_1^{\rm data} - \langle C_1\rangle_{\rm null})/\sigma_{\rm null}$ vs. the label-shuffle null mean $\langle C_1\rangle_{\rm null}$ and width $\sigma_{\rm null}$; $z = +3.64$ (500-MC canonical direct run), $z = +7.93$ (10^4-permutation canonical unapodized), $z = +7.28$ (10^4-permutation apodized). These three values are from different null-run sizes and mask/weight conventions and are not mutually comparable.
- *Block-bootstrap z*: $(A^{\rm best}_{\rm dipole} - A_{\rm ref})/\sigma_{\rm boot}$ with spatial-coherence-corrected $\sigma_{\rm boot}$; the $z \approx -7.6$ primary exclusion (against $A_{\rm ref} = 0.017$, Shamir's 1.7% asymmetry in the same A_p units) uses this convention only.

B. Declared Analysis Hierarchy

We declare the estimator hierarchy used throughout. This hierarchy — with the real-space dipole as the row-(i) primary cosmological estimator — is documented in Appendix A; the primacy of the real-space estimator is established independently of the harmonic-channel diagnostics. The primary dipole measurement uses the high-confidence subsample (Catalog C with winning-class confidence $p_{\rm eq} > 0.6$; $N_{\rm HC} = 949{,}584$ spirals), which suppresses the depth-correlated low-confidence-tail systematic identified in Sec. IV C; all estimator-hierarchy and sensitivity statements below are made on this HC selection unless explicitly noted. We declare the HC $p_{\rm eq} > 0.6$ subsample as the *single primary science sample* for all primary cosmological statements; the block-bootstrap WLS template fit is reported on the full Catalog C field as a *diagnostic* template-disfavor cross-check on a different (unthresholded) sample, and is not treated as a same-sample frequentist exclusion (its $p_{\rm eq} > 0.6$ selection function is not propagated into the block-bootstrap covariance; App. D). See also Appendix D for the full eight-anchor discriminator table.

- **Primary cosmological estimators:** (i) real-space CW-fraction dipole fit on Catalog C at NSIDE=64 (moment-ratio $z_{\rm mom} = +0.41$ against the isotropic (pixel-)permutation null, empirical-rank $p = 0.31$; these are independent numbers, not Gaussian (z, p) partners — see Sec. IV C); and (ii) block-bootstrap WLS template-fit exclusion of a clean 1.7% dipole (the lower end of Shamir's reported 1.7%–4.0% asymmetry range [1, 3], already in A_p units) on the canonical-mask A_p field ($z \approx -7.6$ under the adopted NSIDE= 8 block-bootstrap error model; Appendix D).
- **Secondary diagnostic estimators:** (iii) canonical-N direct-MC NaMaster at $\ell = 1$ on the patchy canonical mask ($f_{\rm sky} = 0.49005$, $+3.64\sigma$); (iv) apodized-footprint MASTER at $\ell = 1$ ($N_{\rm all} \geq 1$ mask, $f_{\rm sky} = 0.494$, C^2 2°; $+7.28\sigma$ vs. global label-shuffle and $+7.13\sigma$ vs. depth-stratified null, $W_p = N_{\rm all}$); and (v) hemisphere maximum-asymmetry (3.05σ local maximum against the label-shuffle null, $p_{\rm LEE} \leq 10^{-4}$ direct-MC max-statistic, systematics-attributed (the Gaussian-Bonferroni $< 1\sigma$ heuristic is a non-principled cross-check); Appendix C). These characterize how a non-zero global monopole and coherent low-ℓ systematics couple through the patchy weighted footprint.
- **Generative monopole-only null:** (vi) $N = 500$ binomial-monopole realizations demonstrating that the raw pre-MASTER pseudo-$C_\ell^{(\ell=1)}$ is dominated by monopole-mask leakage; the post-MASTER $+3.64\sigma$ canonical-mask residual is non-primary and requires additional coherent systematics beyond the monopole-only channel (Sec. IV D).
- **Sensitivity floor:** (vii) empirical injection-recovery on the HC-broad spiral subsample ($p_{\rm eq} > 0.6$, $N = 949{,}584$; Sec. VI B): 50%-recovery-at-3σ threshold at $A = 0.75\%$.

To make the mapping from estimator to scientific claim explicit — and to forestall any reading of the multi-null residual structure as a "forest" of competing detection significances — Table I is a decision tree: each *scientific claim* the paper makes is tied to exactly one estimator class, its analysis sample, its null, and its role. Only the two rows marked PRIMARY carry cosmological weight; every harmonic-channel σ is a systematics diagnostic and supports no independent detection claim. This is the single load-bearing map for the entire Results section: a reader who keeps only Table I retains the complete logical chain from data to the null conclusion.

Table II consolidates per-estimator $N_{\rm spiral}$, $f_{\rm sky}$, mask, null type, and reported σ. The spread of reported significances (from $+0.41\sigma$ to $+7.93\sigma$) across these rows is *not* a set of conflicting measurements of one quantity requiring statistical reconciliation: each row is a distinct *observable* (real-space dipole vs. single-mode pseudo-C_ℓ vs. hemisphere max-statistic vs. template-fit amplitude) evaluated on a distinct sample against a distinct null, so the values are not commensurable as detection significances and no joint likelihood over them is defined or claimed. The scientific verdict rests solely on the two rows tagged PRIMARY — the real-space HC dipole (consistent with null) and the block-bootstrap WLS template fit (a clean 1.7% dipole disfavored); every DIAGNOSTIC row is, by design, a systematics characterization

TABLE I. **Estimator decision tree — which estimator supports which scientific claim.** Each row is a distinct claim made in this paper, tied to its estimator (cross-referenced to the hierarchy rows above and to Table II), the analysis sample with its explicit spiral count N, the null procedure, and its role. PRIMARY rows are the only cosmological detection/exclusion claims; DIAGNOSTIC rows characterize systematics and assert *no* independent signal; the CALIBRATION row sets the sensitivity/falsification scale. Within a role the σ/z values are internally meaningful; *across* rows they are not comparable as detection significances (each is against its own null, listed in the "Null" column; Sec. III A). Only rows P1–P2 carry cosmological weight.

Scientific claim	**Estimator**	**Sample ($N_{\rm spiral}$)**	**Null**	**Role**
Real-space chirality dipole is consistent with zero	(i) HC real-space dipole	HC, $p_{\rm eq}>0.6$ (949,584)	isotropic pixel-permutation (10^4)	PRIMARY
A clean cosmological 1.7% dipole is disfavored	(ii) block-bootstrap WLS template fit	full Catalog C (3,201,160)	block-bootstrap (10^3)	PRIMARY
Harmonic $\ell=1$ residuals are systematics (monopole–mask leakage + coherent depth/morphology), *not* a cosmological signal	(iii,iv) canonical/apodized MASTER; (vi) monopole-only generative null	full Catalog C (3,201,160)	label-shuffle / depth-stratified / monopole-only	DIAGNOSTIC
The hemisphere max-statistic rejects the random-label null after look-elsewhere correction, but the excess is attributed to survey/classifier systematics and is not interpreted cosmologically	(v) hemisphere max-asymmetry	full Catalog C (3,201,160)	max-statistic MC ($p_{\rm LEE}$)	DIAGNOSTIC
Detection-efficiency (recovery) threshold of the primary real-space estimator; not a frequentist upper limit	(vii) injection–recovery	HC, $p_{\rm eq}>0.6$ (949,584)	per-pixel shuffle	CALIBRATION

whose elevated σ is attributed (Sec. IV D, Appendix D) rather than interpreted as signal. The apparent tension is therefore between *systematics diagnostics and the primary null*, and is resolved in the systematics direction by the eight-anchor battery, not left as an unmodeled statistical inconsistency.

C. 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:

$$\begin{aligned}&\text{LayerNorm} \to 384\to512 \text{ (GELU, } d\text{=}0.3)\\ &\quad\to 512\to256 \text{ (GELU, } d\text{=}0.2) \to 256\to3 \text{ (softmax).}\end{aligned} \tag{1}$$

The three-class output ($P_{\rm CW}$, $P_{\rm CCW}$, $P_{\rm NS}$) is essential for full-survey deployment: applying a binary classifier to data where $\sim 70\%$ of objects are elliptical or irregular produces a catalog dominated by noise. Full architecture and training details are in Appendix B.

D. Test-Time Equivariant Averaging

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

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

This procedure enforces flip-equivariance of the output protocol (flip-swap correlation = 1.000). We restrict to 2-fold TTA (original + horizontal flip) rather than the full D_4 group because mirrors flip chirality by definition, whereas in-plane rotations do not change chirality; rotation-TTA probes classifier non-equivariance rather than the chirality assignment itself. Because Eq. (2) enforces flip-equivariance at inference by construction, the flip-swap consistency test T1 (Appendix B) is a *protocol implementation check* that guards against code defects in the TTA averaging, not an independent statistical test of the learned model's equivariance; the pre-TTA behavior of the underlying network is documented separately by the raw single-pass Catalog A tier (Table IV). A direct D_4-TTA hold-out on two independent $\sim$2,000-galaxy subsamples confirms the mean per-galaxy $P_{\rm CW}$ is stable under Z_2 and D_4 to within $|\Delta\langle p_{\rm CW}\rangle|<0.0016$; per-

TABLE II. Primary-estimator summary. $N_{\text{catalog spiral}}$ is the underlying Catalog C spiral count (CW+CCW only). $N_{\text{map weighted}} = \sum_{p\in\text{mask}} W_p$ where $W_p = N_{\text{all}}^{(p)}$ is the total classified-galaxy count in pixel p (CW+CCW+NS), used as a survey-depth weight in the NaMaster field object. $N_{\text{map weighted}}$ exceeds $N_{\text{catalog spiral}}$ because W_p includes non-spiral galaxies (∼62% of the catalog); each galaxy is counted once. Row (iv) is the apodized-footprint MASTER diagnostic ($N_{\text{all}} \geq 1$ mask, C^2 2° apodization); its two σ values are against the global per-galaxy label-shuffle and depth-stratified nulls respectively (both systematics-attributed; Appendices A, D). *The σ values in different rows are computed against different null procedures (column "Null") and are not directly comparable across rows as detection significances.* Row (v) reports the post-look-elsewhere-corrected significance; the raw direct-MC value is $p_{\text{LEE}} \leq 10^{-4}$ against the random-label max-statistic null, which *already incorporates* the look-elsewhere scan and is the principled directional look-elsewhere control (Appendix C); the rejection is systematics-attributed. Row (vii) uses the HC-broad subsample ($p_{\text{eq}} > 0.6$; Sec. VI B). A "—" in $N_{\text{map weighted}}$ marks rows where the NaMaster weight construction is not invoked.

Estimator	$N_{\text{catalog spiral}}$	$N_{\text{map weighted}}$	f_{sky}	Mask	Null	Reported statistic
(i) real-space dipole (HC)	949,584	—	0.4801[a]	canonical	pix-perm. (10^4)	+0.41 (moment z)
(ii) WLS template excl.	3,201,160	—	0.49005	canonical	block-boot. (10^3)	$z \approx -7.6$
(iii) canonical MASTER (diag.)	3,201,160	—	0.49005	canonical	pp-shuffle	+3.64
(iv) apod. MASTER (diag.)	3,201,160	8,474,531	0.494[b]	$N_{\text{all}} \geq 1$ apod.	pp-sh./d-str.	+7.28/+7.13
(v) hemisphere LEE (MC)	3,201,160	—	0.49005	canonical	max-stat MC	$p_{\text{LEE}} \leq 10^{-4}$ (syst.-attr.)
(vi) monopole+mask null	3,201,160	—	0.49005	canonical	monopole-only	+1.69
(vii) injection floor	949,584 HC	—	—	—	pp-shuffle	50%-rec-3σ, $A = 0.75\%$

[a] $f_{\text{sky}} = 0.4801$ is the canonical $N_{\text{spiral}}(p) \geq 10$ mask re-evaluated on the high-confidence subsample: fewer HC spirals populate fewer pixels at the ≥10 threshold (23,600 of 49,152 pixels), slightly shrinking the full-catalog canonical mask ($f_{\text{sky}} = 0.49005$, Appendix A Table X). Artifact: `pipelines/p2_chirality/outputs/dipole/catalog_c_summary.json`.

[b] Geometric $N_{\text{all}} \geq 1$ footprint pixel fraction; the corresponding weighted/apodized *effective* sky fraction for this row's $W_p = N_{\text{all}}$, C^2 2° configuration is $f_{\text{sky}}^{\text{eff}} = 0.452$ (Appendix A, Table X).

galaxy argmax labels flip in 21.4% of cases between Z_2 and D_4 on borderline galaxies with $P_{\text{CW}} \approx P_{\text{CCW}} \approx 0.4$. This D_4 hold-out is a classifier-stability check (comparing softmax outputs under two augmentation groups), not a spatial-null or isotropy test; spatial-null calibration is performed via the isotropic (pixel-)permutation nulls of Sec. IV C. Full details in Appendix B.

E. Catalog Tiers

The pipeline produces three tiers: **Catalog A** (raw, single-pass softmax); **Catalog B** (Platt-calibrated, +0.4% excess); **Catalog C** (equivariant production, 2-fold flip TTA). **Catalog C is the recommended tier for all cosmological parity analyses.** All three tiers share 8,474,531 rows in Apache Parquet format.

IV. RESULTS

Significance conventions. This paper reports significance in standard-deviation units (σ) throughout, but values from distinct null procedures are *not* directly comparable. Section identifiers specify the null for each result. Empirical rank p-values are one-sided unless explicitly labeled two-sided.

TABLE III. **Primary Results callout.** The two load-bearing claims of this paper (rows P1–P2) and, for reference, the non-primary harmonic diagnostics (rows D1–D2) that are *systematics-attributed, not detections*. Each row names the estimator and the specific null procedure its σ/z is computed against. **Only rows P1–P2 carry cosmological weight;** the σ/z values across rows (including $+0.41\sigma$, $+3.64/+7.28\sigma$, and $z \approx -7.6$) derive from distinct null constructions and are *not* directly comparable as detection significances (Sec. III A).

Claim / channel	Null	Value
P1 real-space HC dipole ($p_{\text{eq}} > 0.6$), *primary*	isotropic permutation	$+0.41\sigma$ ($p = 0.31$)
P2 clean-1.7% exclusion (WLS fit), *primary*	block-bootstrap	$z \approx -7.6$ (disfavored)
D1 HC dipole (cross-check)	label-shuffle	$z = 0.58$ (diag.)
D2 MASTER $\ell = 1$ residual *systematics, not a detection*	label-shuffle (harmonic)	$+3.64/+7.28\sigma$

A. Catalog Statistics

The final catalog contains 8,474,531 galaxies (157 of 8,474,688 failed quality checks). Catalog C (equivariant): CW 1,592,107 (18.78%), CCW 1,609,053 (18.99%), NS/edge-on 5,273,371 (62.23%); spiral total N_{spiral} = 3,201,160 (37.77%) (percentages rounded to maintain sum-to-one consistency at the second decimal; exact values: CW 18.787%, CCW 18.987%, NS 62.226%, spiral 37.774%; the integer counts are exact). The spiral fraction is consistent with magnitude-limited survey expectations. Mean classification confidence is 0.951, median 0.9997. We caution that these max-class probabili-

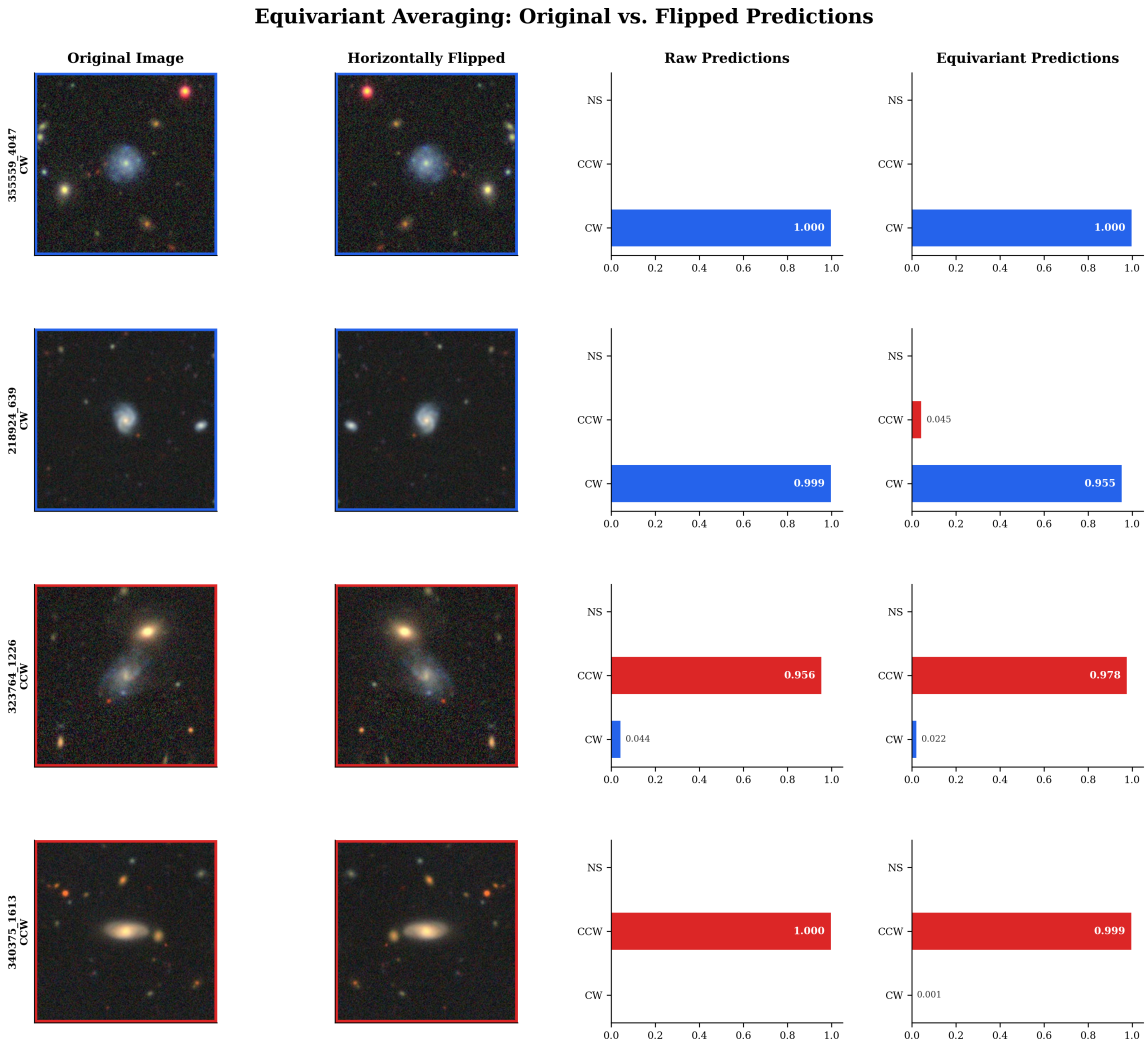


FIG. 2. **Equivariant test-time averaging (TTA).** Representative Z_2 production TTA examples (original + horizontal flip); D_4 validation (four rotations × two reflections) in Appendix B. *Production* inference (Catalog C) uses 2-fold Z_2 TTA — original + horizontal flip only (§III D). Flips swap the CW ↔ CCW class labels by construction. Output probabilities are averaged after the swap, yielding a strictly flip-equivariant CW/CCW classifier with flip-swap correlation = 1.000 by construction. This averaging is the key methodology distinction between Catalog A (raw), Catalog B (Platt-calibrated), and Catalog C (equivariant); the global chirality asymmetry $(N_{\rm CW} - N_{\rm CCW})/N_{\rm spiral}$ shifts from +1.576% (A) to −0.529% (C), i.e. +0.788% to −0.265% in $f_{\rm CW}$-deviation units, dominated by this step (Table IV). Unit reminder: asymmetry-A values are exactly twice the $f_{\rm CW}$-deviation values, $A = 2(f_{\rm CW} - \frac{1}{2})$ (Sec. IV C); Table IV quotes the $f_{\rm CW}$-deviation convention.

ties are *not probabilistically calibrated*: the catalog-wide mean confidence (0.951) far exceeds the independent GZ1 three-class accuracy (58.7%; spiral-chirality 69.91%, Appendix B), i.e. the classifier is strongly overconfident relative to external truth. The $p_{\rm eq}$ "high-confidence" cuts used throughout are therefore monotone sample-selection thresholds (rankings by classifier confidence), not statements that a selected label is correct with probability $p_{\rm eq}$; the injection-recovery floors of Sec. VI B are defined operationally on the threshold-selected subsamples and do not assume calibration. Crucially, the primary dipole estimator consumes *hard argmax* CW/CCW counts per pixel, not the confidence scores themselves, so probabilistic miscalibration cannot bias the dipole *amplitude or direction*: overconfidence affects only which galaxies enter a given $p_{\rm eq}$-selected subsample, and the confidence-cut sweep of Sec. IV C ($p_{\rm eq} \in \{0, 0.4, 0.5, 0.6, 0.7, 0.8\}$) directly maps that selection dependence (the null verdict is stable across the entire high-confidence regime). What *is* propagated into the science is the external GZ1 chirality accuracy (69.91%), which enters the conservative dilution factor $g = 2a - 1$ folded into the empirical floors (Sec. II, VI B) — i.e. the load-bearing validation against independent human labels is the GZ1 cross-match, not the internal softmax confidence. To state the calibration argument in one line: for a fixed selected

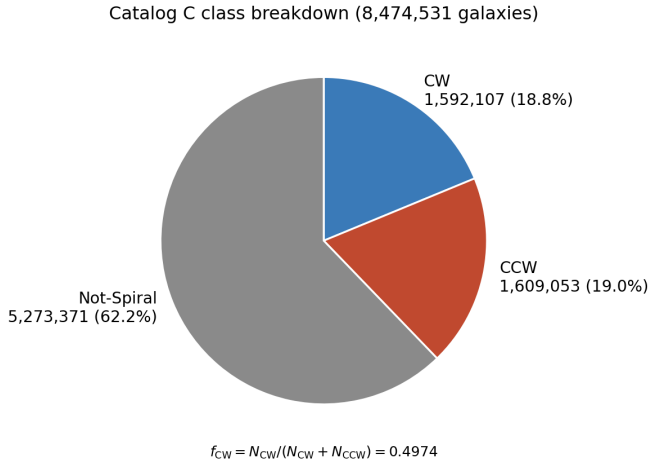


FIG. 3. **Catalog C composition.** Of the 8,474,531 galaxies retained after image-quality QA, the equivariant TTA classifier (§III D) assigns $N_{\rm CW} = 1{,}592{,}107$, $N_{\rm CCW} = 1{,}609{,}053$, and $N_{\rm NS} = 5{,}273{,}371$ (non-spiral / edge-on / morphologically indeterminate). The spiral sub-catalog $N_{\rm spiral} = N_{\rm CW} + N_{\rm CCW} = 3{,}201{,}160$ is the analysis target for all chirality statistics below (Table IV *et seq.*).

sample, monotone recalibration does not alter the hard argmax labels the dipole is fit to, so probabilistic miscalibration does not by itself bias the dipole amplitude or direction within that sample; however, because $p_{\rm eq}$ *defines* the sample and can correlate with depth, morphology, seeing, redshift, or footprint, the spatial selection it induces must itself be validated — which is what the confidence-cut sweep (Sec. IV C) and the external template-regression anchors of Appendix D do (the null verdict is stable across the entire high-confidence regime and orthogonal to the tested survey templates). The fully calibration-free GZ1-human-label cross-check (Sec. VI A) corroborates the null independently of any softmax scaling. *Flip-identity QC (surfaced here from Data Availability / Appendix B).* A catalog-wide QC pass flags 59,515 HC rows (2.9% of catalog rows; 1.6% on the single CW channel) whose *reconstructed* flip-pass probability ($p^{\rm flip}_{\rm CCW} = 2p^{\rm eq}_{\rm CW} - p^{\rm raw}_{\rm CW}$) falls outside $[0, 1]$ by up to 0.09; this affects only the reconstructed raw/flip probability columns (a raw/equivariant pipeline-pass mismatch, not the equivariant hard-argmax labels or the normalized $p^{\rm eq}_c$, which are safe for the dipole estimator). Excluding these flagged rows leaves the real-space dipole null-consistent and essentially unchanged ($z = +0.48$ excluded vs. $+0.52$ baseline; full accounting in Data Availability and Appendix B). *Guidance for downstream users of the public catalog:* the released per-galaxy $p_{\rm eq}$ scores are uncalibrated max-class ranking outputs and must *not* be used as frequentist likelihoods or posterior label probabilities in downstream inference without first applying an external recalibration (e.g. Platt/temperature scaling against a held-out human-labeled set, as directed in the Data Availability section); used raw, they will systematically overstate label reliability by the $\gtrsim 0.25$–0.36 top-label ECE quantified in Appendix B.

TABLE IV. Global CW fraction across catalog tiers. Uncertainties are 1σ binomial, $\sigma = \sqrt{f(1-f)/N_{\rm spiral}}$, with per-tier spiral counts ($N^{\rm A}_{\rm spiral} = 3{,}321{,}795$; $N^{\rm C}_{\rm spiral} = 3{,}201{,}160$); parentheses give the 1σ uncertainty on the trailing digits ($0.507879(274) \equiv 0.507879 \pm 0.000274$). The Excess column is the deviation $f_{\rm CW} - 0.5$ in percent ($f_{\rm CW}$ units; multiply by 2 for asymmetry-A units). Dev. is the *signed* $(f_{\rm CW} - 0.5)/\sigma$ computed from the unrounded fraction. The Catalog-B row derives from the Platt-calibrated fraction; its deviation is computed from the unrounded calibrated fraction. Percentages in this table (and in the catalog-composition counts cited in Sec. II) are rounded to maintain sum-to-one consistency at the second decimal; the integer counts are exact.

Tier	**cw/(cw + ccw)**	**Excess (%)**	**Dev. (σ)**
A (raw)	0.507879(274)	+0.788	+28.72
B (calibrated)	0.50400(27)	+0.4	+14.6
C (equivariant)	0.497353(279)	−0.265	−9.47

B. Global CW Fraction

The Catalog C residual (-9.5σ from 0.5000, Table IV) is spatially uniform across 7 equatorial coordinate slabs and does not produce a dipole. This monopole offset is a classifier artifact, not a physical signal; three candidate mechanisms are (1) GZ1 training-label CW excess, (2) residual orientation-dependent bias not corrected by 2-fold TTA, and (3) photometric asymmetry in DESI Legacy imaging. The 2.98× asymmetry-suppression factor from raw $+1.576\%$ to equivariant -0.529% (asymmetry-A units, $A = 2(f_{\rm CW} - \frac{1}{2})$; equivalently $+0.788\% \rightarrow -0.265\%$ in $f_{\rm CW}$-deviation units, Table IV) demonstrates the dominance of the equivariant TTA processing. Crucially, the slab statistics support this quantitatively: in 7 equal-spiral-count declination slabs ($N = 457{,}308$–$457{,}309$ spirals each; per-slab binomial $\sigma = 7.4 \times 10^{-4}$), the per-slab $f_{\rm CW}$ spans 0.49537–0.49890, i.e. deviations from 0.5 of -0.110% to -0.463%, all within 0.5% of 50/50; an equal-count RA partition gives a compatible span (-0.060% to -0.501%). An equal-*area* partition (8 declination bands of equal in-mask pixel count on the canonical mask) gives the same verdict: per-band $f_{\rm CW}$ deviations from 0.5 reach at most 0.49% on the full spiral sample (max $|z| = 2.9$ vs. the global rate) and 0.56% on the HC subsample (max $|z| = 1.4$), comparable to the equal-count maximum of 0.46% (artifact `pipelines/p2_chirality/outputs/canonical_provenance/c12_r24conf_local_batch.json`). The slab-to-slab scatter about the global $f_{\rm CW} = 0.49735$ is $\lesssim 2.7\sigma$ per slab (the extremal slab $f_{\rm CW} = 0.49537$ gives $(0.49735 - 0.49537)/(7.4 \times 10^{-4}) = 2.7\sigma$ from the per-slab values and binomial σ above; artifact `pipelines/p2_chirality/outputs/canonical_`

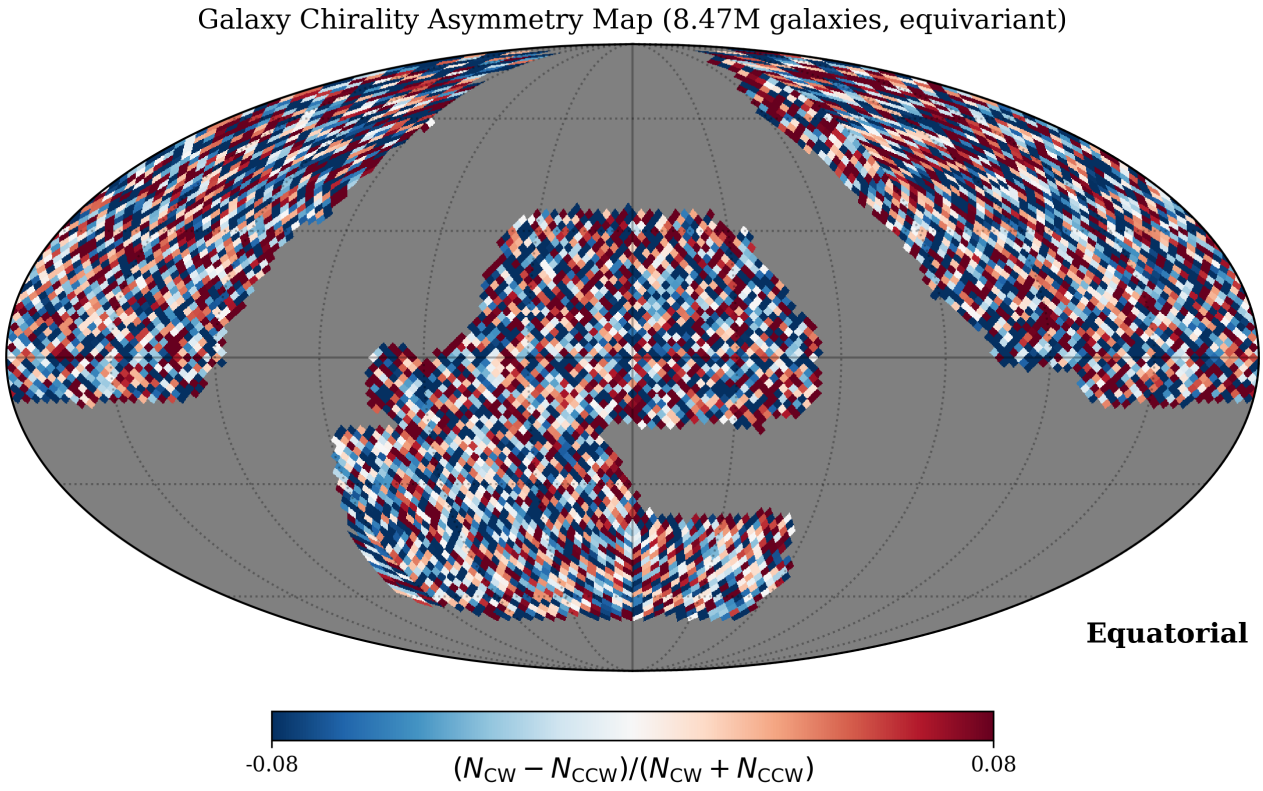


FIG. 4. **Equivariant (Catalog C) chirality asymmetry map of the 8.47 M-galaxy catalog** (Mollweide projection, equatorial coordinates; per-pixel asymmetry $A_p = (N_{\rm CW} - N_{\rm CCW})/(N_{\rm CW} + N_{\rm CCW}) = 2(f_{{\rm CW},p} - \frac{1}{2})$ at HEALPix NSIDE=64, color scale $[-0.08, +0.08]$). The DESI Legacy Imaging footprint covers $f_{\rm sky} = 0.49005$ of the sky in the canonical mask ($N_{\rm spiral}(p) \geq 10$ per pixel; Sec. IV C); the $N_{\rm all} \geq 1$ analysis footprint ($f_{\rm sky} = 0.494$) is used for the apodized MASTER diagnostic. Spatial uniformity of the per-pixel CW fraction is verified across 7 equatorial coordinate slabs (§IV B); the canonical-mask $\ell = 1$ residual is analyzed in Sec. IV C–IV D.

provenance/c12_r24conf_local_batch.json), consistent with the coherent low-ℓ systematic structure dispositioned in Appendix D rather than a dipole-aligned gradient. Independently of any slab partition, a direct generative test confirms that a constant monopole cannot bias the uniform-weight real-space dipole estimator (the constant template is absorbed by the fitted monopole term): binomial per-pixel realizations on the canonical mask drawn at $p = f_{\rm CW}^{\rm global}$ versus $p = 0.5$ yield statistically identical dipole-amplitude null distributions (means 1.957×10^{-3} vs. 1.935×10^{-3}, a 0.39σ shift in the standard error of the difference; $N = 500$ each; artifact pipelines/p2_chirality/outputs/canonical_provenance/c11_meta_m4_slab_stats.json). This generative test probes an *additive constant* monopole only; multiplicative depth- or morphology-coupled modulations of the monopole are not probed by it and are addressed instead by the template-regression and cross-spectrum diagnostics of Appendix D. The sub-percent sensitivity claim should therefore be interpreted as a floor on the physical isotropy-breaking dipole signal. *Implications for $\ell = 0$ parity searches.* The morphological handedness monopole is itself the parity-odd channel (unlike the parity-even $\ell = 1$ dipole; Sec. VI C), so a naive reading of the -9.47σ deviation as cosmic parity violation would be the single largest false-positive risk in any $\ell = 0$ chirality-parity study on such a catalog. Our analysis shows this deviation is a spatially-uniform classifier artifact ($\lesssim 0.5\%$ per-slab scatter, three candidate instrumental/training mechanisms above), not a physical signal; the concrete lesson for future $\ell = 0$ parity searches is therefore that the released catalog labels must be locally monopole-renormalized (per-region CW-fraction subtraction) before any global parity-odd statistic is formed, and that any claimed $\ell = 0$ parity detection at or below the $\approx 0.26\%$ artifact level (9.5σ in raw counts) is indistinguishable from this classifier monopole without an independent, handedness-symmetric labeling pipeline (Sec. E, Data Availability).

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 $N_{\rm spiral}(p) \geq 10$ spiral galaxies (this $N_{\rm spiral}(p) \geq 10$

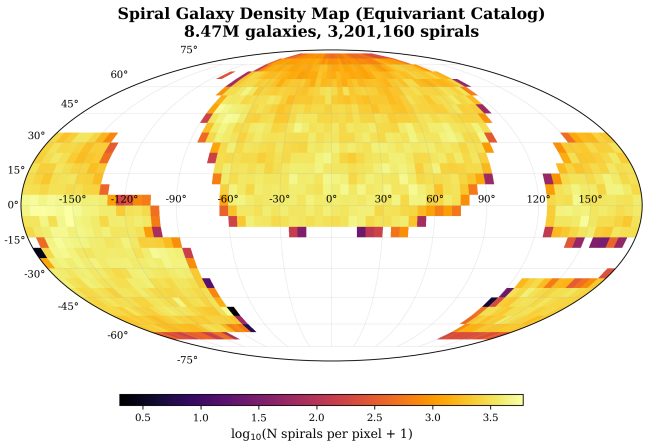


FIG. 5. **Sky density of the** 3,201,160 **classified spirals** (CW + CCW combined, NSIDE=64 Mollweide). Per-pixel spiral counts scale with the DESI Legacy Imaging Surveys depth and exposure pattern; the canonical mask used for the primary $\ell = 1$ analysis (§IV C) requires $N_{\rm spiral}(p) \geq 10$ per pixel. Spatial inhomogeneity at this scale is the leakage channel quantified in §IV D.

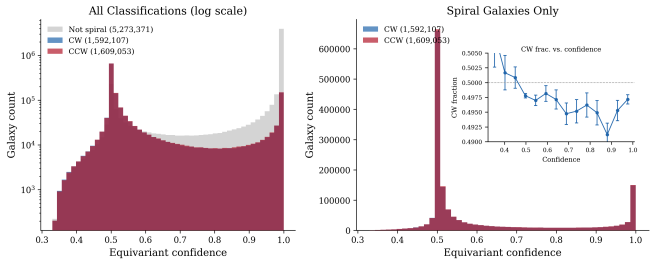


FIG. 6. **Distribution of maximum-class confidence** $\max(P_{\rm CW}, P_{\rm CCW}, P_{\rm NS})$ **for all** 8,474,531 **galaxies.** Strongly bimodal: 73.6% at $\max p \geq 0.9$ (high-confidence labels) + a long tail of indeterminate cases ($\max p < 0.5$, dominated by NS/edge-on systems). The high-confidence (HC) cuts at $p_{\rm eq} > 0.6$ ($N = 949{,}584$) and $p_{\rm eq} > 0.8$ ($N = 624{,}660$) used in the systematics cross-checks (§E) are indicated.

per-pixel cut is *the* canonical mask, $f_{\rm sky} = 0.49005$, and is the single mask definition referenced by every figure and table for every quoted number in this paper unless explicitly noted; Appendix A), we compute the asymmetry

$$A_p = \frac{N_{\rm CW}^{(p)} - N_{\rm CCW}^{(p)}}{N_{\rm CW}^{(p)} + N_{\rm CCW}^{(p)}}. \tag{3}$$

This is the single canonical chirality-field definition used throughout (spirals-only denominator). Note the unit convention $A_p = 2\,(f_{{\rm CW},p} - \frac{1}{2})$: amplitudes quoted in A_p units are exactly twice the corresponding $f_{\rm CW}$-deviation amplitudes.

a. Simple dipole. The real-space estimator fits the monopole-plus-dipole model $A_p = m + \boldsymbol{a} \cdot \hat{n}_p$ to the in-mask pixel map by least squares with uniform pixel weighting (`healpy.fit_dipole`); the dipole amplitude is $A_{\rm dip} = |\boldsymbol{a}|$. The null distribution is built from $N_{\rm MC} = 10{,}000$ isotropic realizations in which the per-pixel asymmetry values A_p are randomly permuted across the in-mask pixels (destroying any coherent dipole while preserving the one-point distribution) and the fit repeated; the quoted p is the *one-sided empirical rank* of the observed amplitude in this null (the two-sided equivalent is 0.62). Because $A_{\rm dip}$ is positive-definite, the (z, p) pair does not follow the Gaussian $z \to p$ mapping ($z = +0.41$ is the moment-ratio against the null mean and width; $p = 0.31$ is the rank). The primary fit uses the high-confidence Catalog C selection of the generator script (`pipelines/p2_chirality/run_dipole_catalog_c.py`: equivariant spirals with winning-class confidence > 0.6; $N_{\rm HC} = 949{,}584$). *Pre-registration of the* 0.6 *cut.* The $p_{\rm eq} > 0.6$ primary selection was fixed *a priori* in the generator script (`pipelines/p2_chirality/run_dipole_catalog_c.py`) before the dipole was evaluated, so it is not tuned against the dipole outcome. The record is immutable and independently verifiable: the literal cut `np.maximum(p_cw_eq, p_ccw_eq) > 0.6` appears in that script's selection block and module docstring as of committed revision `94113e5` (2026-06-09) — the same commit that defines the estimator. No separate frozen tag was created, so the commit hash and date themselves constitute the pre-registration record; the cut cannot be altered retroactively without rewriting public git history. The threshold is also robust rather than a researcher degree of freedom: the full confidence-cut sweep below ($p_{\rm eq} \in \{0, 0.4, 0.5, 0.6, 0.7, 0.8\}$) leaves the null verdict unchanged across the high-confidence regime ($z = +0.41, +1.14, +0.51$ at $0.6, 0.7, 0.8$; all $|z| < 1.2$), so the conclusion is invariant to the exact cut and no forking-paths concern applies. *Why* 0.6 *specifically:* on the uncalibrated three-class winning-score $p_{\rm eq}$, a 0.5 cut is the argmax boundary and admits the entire depth-correlated low-confidence tail — the population carrying the $z \approx 4.0$–4.3 systematic excess documented below — while 0.6 is the lowest threshold that already excludes that tail (the transition localizes at $p_{\rm eq} \leq 0.6$ in the sweep). It is therefore the completeness-maximizing choice subject to excluding the systematic-dominated tail. The GZ1 cross-match quantifies the tradeoff: on the 234,282-galaxy disjoint GZ1 sample (Appendix B, Table XIII), integrated chirality purity is $\approx$ 70% (70.4% CW, 69.4% CCW) and rises monotonically with $p_{\rm eq}$; the 0.6 cut retains $N_{\rm HC} = 949{,}584$ of the 3,201,160 classified spirals ($\approx$ 30% completeness), while tightening to $p_{\rm eq} > 0.9$ halves completeness again ($N = 471{,}049$, $\approx$ 15%) and inflates the A_{50} floor by $\sqrt{949{,}584/471{,}049} \approx 1.42\times$ for no change in the null verdict. *Representativeness of the HC subsample.* The $p_{\rm eq} > 0.6$ high-confidence subsample ($\approx$ 30% of the full classified-spiral catalog) is the pre-specified science sample precisely because the discarded low-confidence tail is where the depth-correlated classifier systematic

concentrates; the null is therefore reported on the regime where the chirality labels are trustworthy, and the catalog-wide unthresholded excess (Sec. below) is a diagnostic of that systematic tail, not a competing cosmological measurement. (A formal pre-unblinding purity–completeness curve on the GZ1 validation set would further document the choice and is a straightforward extension.) *Selection-function note:* the $p_{\rm eq} > 0.6$ hard confidence cut defines the primary science sample; this selection function (based on uncalibrated network confidence scores) is *not* propagated into the block-bootstrap spatial covariance model of the WLS template fit, which operates on the full canonical-mask A_p field (Catalog C, $N_{\rm spiral} = 3{,}201{,}160$). The HC subsample test and the WLS fit are therefore distinct estimators on overlapping but different samples; both independently return null or disfavor a clean dipole, and the selection function enters only via the injection-recovery sensitivity floor (which is defined operationally on the HC sample). The fitted dipole has amplitude 4.4×10^{-3} toward $(l, b) = (293°, 12°)$ with significance 0.41σ (rank-p = 0.31), fully consistent with the null hypothesis; at this significance the dipole axis is unconstrained (under the null the recovered direction is effectively a random draw, so we attach no uncertainty contour to the quoted (l, b)). Three robustness checks on this primary: (i) replacing the pixel-permutation null with a *per-galaxy label-shuffle* null — which preserves the per-pixel counts $N_{\rm spiral}(p)$ and their Poisson noise geometry — leaves the verdict unchanged (0.58σ, rank-p = 0.26, same generator; an independent implementation gives z = 0.70, rank-p = 0.23, artifact `pipelines/p2_chirality/outputs/canonical_provenance/c11b_hc_dipole_nulls.json`); (ii) an independent uniform-weight least-squares re-implementation of the fit reproduces the pixel-permutation verdict (z = 0.55, rank-p = 0.27, same artifact); (iii) a 2×3 robustness panel crossing the fit weighting (uniform vs. $N_{\rm spiral}(p)$-weighted) with the mask threshold ($N_{\rm spiral}(p) \geq 10, 20, 50$) leaves every cell consistent with the null ($|z| \leq 0.8$, rank-$p \geq 0.20$; 2000-permutation pixel nulls per cell, artifact `pipelines/p2_chirality/outputs/canonical_provenance/c12_r24conf_local_batch.json`). The regenerated 10^4-permutation null array also yields a formal null-quantile benchmark: defining $A_{95,\rm nq}$ as the 95th percentile of the pixel-permutation null amplitude distribution ("null-quantile"; *not* a signal-injected limit and carrying no frequentist coverage guarantee), i.e. the smallest amplitude with $P(A_{\rm null} \geq A_{95,\rm nq}) = 0.05$ (an estimator-level rank construction; the conservative companion $\max(A_{\rm obs}, A_{95,\rm nq})$ coincides with it since $A_{\rm obs} < A_{95,\rm nq}$, and is a descriptive estimator-level bound used in no scientific conclusion), gives $A_{95,\rm nq} = 6.8 \times 10^{-3}$ in A_p units — for a pure dipole the A_p-dipole amplitude equals the full-amplitude A of the injection convention ($p_{\rm CW} = p_0 + (A/2)\cos\theta \Rightarrow A_p$-dipole amplitude $= A$), so this is numerically on the same scale as the $A_{50} \approx 0.75\%$ injection floor while remaining a different object (null $\{50, 68, 90, 95, 99\}\%$ quantiles $= \{3.5, 4.4, 6.0, 6.8, 8.4\} \times 10^{-3}$; same artifact). *Sensitivity to the confidence threshold:* dropping the confidence threshold entirely (all 3,200,420 in-mask equivariant spirals; the remaining 740 of the catalog's 3,201,160 spirals lie in pixels below the $N_{\rm spiral}(p) \geq 10$ canonical-mask threshold) yields $A_p = 0.0057$ ($= 0.57\%$) dipole at $z \approx 4.2$–4.4 under both null constructions (artifact `pipelines/p2_chirality/outputs/canonical_provenance/c11_meta_e1_e2_realspace_nulls.json`); this unthresholded excess is attributed to residual depth-correlated classifier bias concentrated in the low-confidence tail — the same systematic family dispositioned for the harmonic-channel residuals (Sec. IV D, Appendix D) — and its amplitude sits below the HC-broad $A_{50} \approx 0.75\%$ falsification-criterion floor; it is reported here as a systematics-sensitivity diagnostic, not a detection. (An empirical injection floor measured directly on the unthresholded full-sample estimator (N = 3,201,160 spirals, all $N_{\rm spiral}(p) > 0$ pixels; same WLS estimator and per-pixel binomial null, area-uniform axis draws, $N_{\rm MC,inj} = 200$ per amplitude) gives $A_{50} \approx 0.36\%$ and $A_{95} \approx 0.63\%$ (log-interpolated recovery-fraction crossings; smallest tested amplitudes crossing 50%/95%: 0.5%/0.75%; artifact `pipelines/p2_chirality/outputs/canonical_provenance/c16_r24conf_pod_batch.json`). The $A_p = 0.0057$ ($= 0.57\%$) unthresholded excess therefore lies between the full-sample A_{50} and A_{95} — consistent with its $z \approx 4.2$–4.4 visibility — and the systematic attribution rests on the confidence-cut sweep below, not on a sub-floor amplitude argument.) The confidence threshold is a monotone selection cut on uncalibrated scores (see the calibration caveat in Sec. II), so the primary is defined at the threshold the generator script has used throughout. A full confidence-cut sweep ($p_{\rm eq} \in \{0, 0.4, 0.5, 0.6, 0.7, 0.8\}$; 2000-permutation pixel nulls each) localizes the transition: $z = +4.3, +4.1, +4.0$ at cuts $0, 0.4, 0.5$, collapsing to $z = +0.41, +1.14, +0.51$ at $0.6, 0.7, 0.8$, confirming the excess is confined to the $p_{\rm eq} \leq 0.6$ low-confidence tail (artifact `pipelines/p2_chirality/outputs/canonical_provenance/c12_r24conf_local_batch.json`). In contrast, Catalog A (raw) shows a 2.31σ real-space dipole and a $+6.48\sigma$ pre-MASTER pseudo-$C_\ell^{(\ell=1)}$[1] at the lowest multipole ($\ell = 1$; a single decoupled mode, not a binned bandpower)— both entirely artifacts of the model's residual CW bias spatially modulated by non-uniform survey depth, and both collapsed to null by two complementary reductions

[1] Catalog-A pre-MASTER $\ell = 1$: single-mode pseudo-$C_\ell^{(\ell=1)}$ on the canonical mask ($N_{\rm spiral}(p) \geq 10$, $f_{\rm sky} = 0.49005$) without MASTER deconvolution, per-pixel label-shuffle null (same estimator family as the canonical Catalog-C $+3.64\sigma$ result; Sec. IV D).

(equivariant TTA averaging in real space and MASTER mode-coupling deconvolution in spherical-harmonic space).

b. Angular power spectrum. We perform MASTER mode-coupling deconvolution using NaMaster [34] at $N_{\rm side} = 64$ on the real analysis footprint ($N_{\rm all} \geq 1$; 24,297 pixels, $f_{\rm sky} = 0.494$, C^2 2° apodization, depth weight W_p). The MASTER-deconvolved single-mode $\ell = 1$ value (computed on the monopole-subtracted A_p field; declared data vector, Appendix A) is $C_1^{\rm meas} = 2.348 \times 10^{-5}$ against a 500-MC per-galaxy label-shuffle null with mean 1.71×10^{-6} and $\sigma_{\rm null} = 2.99 \times 10^{-6}$, i.e. $+7.28\sigma$ for $W_p = N_{\rm all}$ ($+9.78\sigma$ for $W_p = N_{\rm spiral}$); the 10^4-permutation recompute (Table V) confirms this channel at $z = +7.31$ with empirical rank $p = 6.0 \times 10^{-4}$ (one-sided, computed as $(k+1)/(N+1)$ with $k = 5$ of $N = 10^4$ null draws exceeding the data; Table V caption), and the Gaussian-equivalent σ is quoted from the null moments. A depth-stratified null (labels permuted within 10 $N_{\rm all}(p)$ deciles; this preserves the marginal depth distribution but not joint spatial–depth structure, so it bounds only depth-sampling effects) leaves the excess essentially unchanged ($+7.13\sigma$ / $+9.06\sigma$), excluding pixel-depth sampling alone as the driver. A weight-map sweep gives the same verdict: the excess persists under all three weightings ($W_p = N_{\rm all}$ $+6.9\sigma$, $W_p = N_{\rm spiral}$ $+8.5\sigma$, uniform binary $+6.7\sigma$; independent 500-MC null streams, artifact c9c; in each variant the mask-mean subtraction uses the same variant weight map, so the field-consistent $N_{\rm spiral}$-weighted subtraction is included in the sweep), so it is not an artifact of the depth-weight choice, although its magnitude is weight-dependent at the $\pm 1\sigma$ level. This channel is therefore a *systematics diagnostic* on the patchy weighted footprint — the same coherent low-ℓ family as the canonical-mask residual (Sec. IV D, Appendix D) — and is not used as a cosmological null; the cosmological statements of this paper rest on the real-space dipole and the template-fit exclusion (Appendix D). NaMaster configuration details and the canonical mask-declaration + depth-stratification audit are in Appendix A. We emphasize that rows (i) and (iv) of Table II are not on the same statistical footing — different fields, masks, weights, and null procedures — so the apparent $+0.41\sigma$ vs. $+7.28\sigma$ gap is not a 17× discrepancy on a common axis; the harmonic-completeness check (Sec. VII, artifact c9b) bounds what a clean Shamir-class real-space dipole would have produced in the MASTER channel — it would register at $z \approx 68$–218 there, versus the observed $+7.28\sigma$ — but it does not establish statistical consistency of the two estimators on a common axis.

Reader's note (Table V). The σ/z entries in this table are systematics-attributed harmonic *diagnostics*, each against its own null; they are *not* detection significances and must *not* be compared row-to-row, footprint-to-footprint, or with the primary real-space dipole. Only the two primary estimators (HC real-space dipole $+0.41\sigma$; template-fit exclusion $z \approx -7.6$) carry cosmological weight.

D. Monopole+Mask Leakage Generative Null

The canonical-mask direct-MC $\ell = 1$ value of $+3.64\sigma$ (per-pixel label-shuffle null) and the local hemisphere maximum of 3.05σ (label-shuffle null, 648-direction scan; Appendix C) are candidate manifestations of mask-geometric leakage of the global 9.5σ monopole. We formalize this with a generative null: $N = 500$ realizations in which the per-pixel CW count is drawn from Binomial($N_{\rm spiral}(p), p_{\rm CW}^{\rm global}$) on the exact canonical mask, with no injected dipole.[2]

The monopole-only null reproduces 99.32% of the observed pre-MASTER pseudo-$C_\ell^{(\ell=1)}$ power (± 0.40 pp per-realization null scatter, $N = 500$; the residual is $+1.69\sigma$, Table VI). The reproduction statistic is the ratio of means, $\langle C_1^{\rm null} \rangle_{N=500} / C_1^{\rm data}$; because the observed $C_1^{\rm data}$ is a fixed scalar, the per-realization mean of the ratios $C_1^{{\rm null},k} / C_1^{\rm data}$ is identical, so the mean-of-ratios/ratio-of-means distinction does not arise for this statistic. The standard error on the mean reproduction fraction is $0.40/\sqrt{500} \approx 0.018$ pp (the ± 0.40 pp quoted above is the per-realization scatter, not the uncertainty on the mean; artifact `pipelines/p2_chirality/outputs/canonical_provenance/monopole_mask_null_results.json`). The prior literature's pre-MASTER dipole-detection claims are therefore attributed at the pre-MASTER level to this leakage channel under our DESI/ViT-Small pipeline; a matched Ganalyzer reanalysis remains required for a likelihood-level exclusion of their specific estimator and cuts. **The 99.32% reproduction figure applies exclusively to the un-deconvolved pre-MASTER pseudo-$C_\ell^{(\ell=1)}$.** The post-MASTER behavior is qualitatively different: a MASTER-decoupled monopole-only null ($N = 500$ realizations; artifact `pipelines/p2_chirality/outputs/canonical_provenance/master_decoupled_monopole_null.json`) gives $\sigma = +4.84$ relative to the data, with the monopole-only null mean

[2] Here $N_{\rm spiral}(p) \equiv N_{\rm CW}(p) + N_{\rm CCW}(p)$ is the per-pixel *spiral* count, not the all-galaxy per-pixel count $N(p)_{\rm all} = N_{\rm CW}(p) + N_{\rm CCW}(p) + N_{\rm NS}(p)$ that appears as the weighting field W_p in the A_p definition. The chirality field $A_p = (N_{\rm CW}(p) - N_{\rm CCW}(p))/N_{\rm spiral}(p)$ is defined on spirals only, so the generative null draws from the spiral trial pool to be self-consistent with the field it is reproducing. The generative-null code uses $N_{\rm spiral}(p)$, and the primary 99.32% pre-MASTER reproduction figure in Table VI is on the spiral-trial draw. A parallel rerun on $N(p)_{\rm all}$-trial draws ($N = 500$, seed 42) reproduces 99.33% of the observed pre-MASTER pseudo-$C_\ell^{(\ell=1)}$ power, vs. 99.32% for the spiral-trial draw: the primary reproduction figure is robust to the trial-pool choice. The larger per-pixel trial counts shrink the binomial null variance, so the pre-MASTER residual rises from $+1.69\sigma$ (spiral-trial, Table VI) to $+2.80\sigma$ ($N_{\rm all}$-trial) at essentially identical reproduced power. The per-pixel trial-count inflation factor is $\langle N_{\rm all}/N_{\rm spiral} \rangle = 2.83$ (in-mask pixel mean; the global count ratio is 8,474,531/3,201,160 = 2.65).

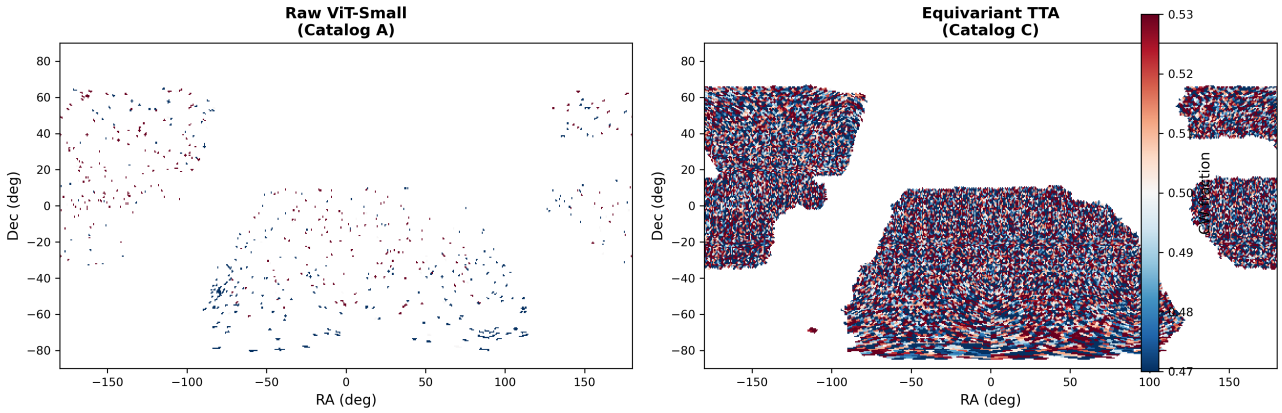


FIG. 7. **Raw (Catalog A) vs equivariant (Catalog C) chirality sky maps** (equatorial RA/Dec, per-pixel CW *fraction* $f_{\mathrm{CW},p}$, NSIDE=64; shared color scale $[0.47, 0.53]$). The panels are plotted in f_{CW} units, not the A_p units of Fig. 4: the two are related by $A_p = 2(f_{\mathrm{CW},p} - \frac{1}{2})$, so the $[0.47, 0.53]$ range here corresponds to $A_p \in [-0.06, +0.06]$. Left: raw single-pass classifier output; the spatially-structured $\sim$0.79% classifier CW excess, modulated by non-uniform survey depth, produces the 2.31σ real-space dipole + $+6.48\sigma$ pre-MASTER $\ell = 1$ artifact. Right: 2-fold flip-equivariant TTA; the bias is removed by construction and the HC ($p_{\mathrm{eq}} > 0.6$) real-space dipole collapses to 0.41σ. This visual diagnostic is the methodology cornerstone — the difference between Catalog A and Catalog C is the difference between a 2σ "detection" and a clean null, demonstrating that future chirality studies must adopt equivariant post-processing to avoid spurious dipoles.

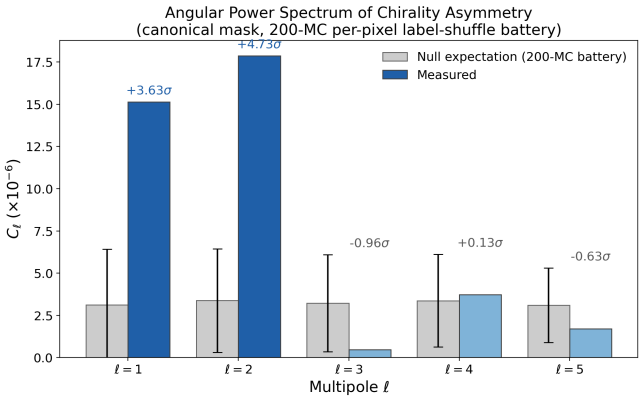


FIG. 8. **Pseudo-C_ℓ of the chirality field A_p on the canonical mask,** $\ell = 1$–5 (single panel; grouped bars). Gray bars with error bars: per-pixel label-shuffle null expectation from the canonical 200-MC multi-null battery (Appendix D), $\pm 1\sigma$; colored bars: measured C_ℓ. The per-ℓ significance annotations embedded in the panel are the canonical battery values ($\sigma_{\ell=1} = +3.63$, $\sigma_{\ell=2} = +4.73$, $\sigma_{\ell=3,4,5} = -0.96, +0.13, -0.63$). The $\ell = 1, 2$ broadband excess is the systematics-attributed structure analyzed in Appendix D. The post-MASTER canonical-mask residual is $+3.64\sigma$ (distinct estimator from the panel's pre-MASTER pseudo-C_ℓ; $p_{\mathrm{MC}} = 0.030$; Sec. IV D), a non-primary systematics-attributed value; the primary null-dipole conclusion rests on the primary estimators (the HC real-space dipole ($p_{\mathrm{eq}} > 0.6$, $N = 949{,}584$) at $+0.41\sigma$, isotropic permutation null, and the template-fit exclusion of a clean 1.7% dipole, Appendix D).

reproducing only $\sim$12% of the post-MASTER decoupled C_1; the 10^4-realization confirmation yields $\sigma = +5.14$ (artifact `pipelines/p2_chirality/outputs/canonical_provenance/master_decoupled_monopole_null_10k.json`). The post-MASTER residual therefore requires coherent depth/PSF/morphology systematics beyond the monopole-only channel — not monopole-mask leakage alone — consistent with the eight-anchor systematic analysis of Appendix D. MASTER decoupling substantially reduces, but does not remove, the canonical-mask pseudo-C_ℓ leakage; the post-MASTER residuals ($+3.64\sigma$ canonical unapodized; $+7.28\sigma$ on the apodized weighted footprint, unchanged at $+7.13\sigma$ under a depth-stratified null) are non-primary, systematics-attributed values consistent with coherent low-ℓ structure that MASTER does not remove on the patchy weighted footprint (Appendix D). The primary null-dipole conclusion therefore rests on the two *primary* estimators: the HC real-space dipole ($p_{\mathrm{eq}} > 0.6$, $N = 949{,}584$) at $+0.41\sigma$, which bypasses the harmonic-leakage channel, and the block-bootstrap WLS template-fit exclusion of a clean 1.7% dipole ($z \approx -7.6$, Appendix D), which tests a clean-dipole template after nuisance marginalization on the canonical-mask A_p field.

The eight-anchor systematic analysis of the $+3.64\sigma$ canonical-mask residual (apodized-mask robustness, multipole-spectrum coherence, quality-quartile stratification, leg-proxy cross-power, density-stratified null, boundary-distance variance, joint nuisance-marginalized WLS template fit, and direct cross-spectrum) is in Appendix D. Summary: three discriminators disfavor interpretation (i) (real cosmological dipole at $\sim$ 1.7%): (a) $\ell = 2 > \ell = 1$ broadband structure incompatible with a clean dipole; (b) p_{eq} quality-quartile washout (all four quartiles $|\sigma| < 1$; Appendix D); (c) suggestive cross-

TABLE V. MASTER-decoupled angular power of the chirality asymmetry map (Catalog C, equivariant; tabulated values are rounded for display — the full-precision arrays live in the committed null-distribution artifacts cited in the text), recomputed with 10^4 per-galaxy label-shuffle permutations (seed-streamed parallel RNG) on both diagnostic footprints, each under its own committed field convention (Appendix A.a): the apodized $N_{\rm all} \geq 1$ analysis footprint with depth weight $W_p = N_{\rm all}$ (A_p field, weight-map-weighted mean subtraction), and the canonical unapodized mask with binary weight ($f_{\rm CW}$–$0.5 = A_p/2$ field, $N_{\rm spiral}$-weighted subtraction; z and rank-p are invariant under this constant rescaling, but C_b amplitudes are not cross-comparable between the two blocks). Band 1 is the single mode $\ell = 1$ decoupled within the full 39-band coupling matrix (single-multipole bin, Appendix A.b — not a bandpower over a range); the single-mode-only decoupling of Sec. IV C ($C_1 = 2.348 \times 10^{-5}$, $+7.28\sigma$) is a distinct estimator with its own null and the two should not be numerically equated. All entries are systematics-attributed diagnostics, not cosmological measurements (Appendix D); z values are relative to each row's own null and are not comparable across rows, footprints, or with the real-space dipole. C_b amplitudes are raw (not shot-noise debiased); the analytic binomial shot-noise floor is computed specifically for the apodized $W_p = N_{\rm all}$ field convention, $N_{\ell=1} \approx 2.0 \times 10^{-6}$ (artifact c9e, propagated through that weight/apodization combination), consistent with the apodized-row null mean (1.93×10^{-6}); the canonical unapodized binary-weight rows carry a different field normalization (null mean 0.57×10^{-6}), and no analytic floor is quoted for that convention. The canonical-N direct-MC single-mode value $+3.64\sigma$ ($p_{\rm MC} = 0.030$, one-sided, $\approx 1.9\sigma$ Gaussian-equivalent; Sec. IV D) is not tabulated above (the canonical rows use the 10^4-permutation null) but is retained in the text for continuity with the leakage analysis. Rank p is computed as $(k+1)/(N+1)$, where k is the number of null draws meeting or exceeding the data value and $N = 10^4$ (e.g. $k = 5$ gives $p = 6/10001 = 6.0 \times 10^{-4}$); the minimum reportable p is $1/(N+1) \approx 1.0 \times 10^{-4}$. The permutation null is heavy-tailed relative to Gaussian at low ℓ, so $z_{\rm mom}$ and the Gaussian-equivalent of rank p need not agree. Full 39-band null arrays are included in the released artifacts (c9a).

Footprint	**Band**	$C_b^{\rm data} \times 10^6$	$\langle C_b \rangle_{\rm null} \times 10^6$	$\sigma_{\rm null} \times 10^6$	z	**rank p (1-sided)**
apod., $W_p = N_{\rm all}$	$\ell = 1$	24.74	1.93	3.12	$+7.31$	6.0×10^{-4}
	$\ell \in [2, 6]$	5.28	1.87	0.73	$+4.67$	5×10^{-4}
	$\ell \in [7, 11]$	3.12	1.92	0.50	$+2.41$	0.015
	$\ell \in [12, 16]$	2.38	1.92	0.40	$+1.16$	0.126
	$\ell \in [17, 21]$	2.64	1.92	0.35	$+2.05$	0.027
	$\ell \in [22, 26]$	2.67	1.92	0.31	$+2.42$	0.013
canonical, unapod.	$\ell = 1$	7.27	0.57	0.84	$+7.93$	3×10^{-4}
	$\ell \in [2, 6]$	1.42	0.56	0.20	$+4.20$	9×10^{-4}
	$\ell \in [7, 11]$	0.91	0.57	0.14	$+2.47$	0.015
	$\ell \in [12, 16]$	0.79	0.57	0.11	$+1.98$	0.030
	$\ell \in [17, 21]$	0.83	0.57	0.09	$+2.75$	0.0065
	$\ell \in [22, 26]$	0.74	0.57	0.09	$+2.00$	0.029

TABLE VI. Monopole+mask leakage null ($f_{\rm sky} = 0.49005$, seed $= 42$; $N = 500$ binomial-monopole realizations). Pseudo-C_ℓ entries in this table are dimensionless band values of the un-monopole-subtracted $f_{\rm CW}$-map convention used by the generative null, and are NOT on the A_p-map $\times 10^{-6}$ units of Table V; the two tables intentionally use different fields and normalizations. The monopole-only null reproduces 99.32% of the observed pre-MASTER pseudo-$C_\ell^{(\ell=1)}$ power (residual $+1.69\sigma$, consistent with the monopole-only null at the pre-MASTER stage). *This is a pre-MASTER diagnostic only*; the post-MASTER decoupled null gives $\sigma = +4.84$ (monopole-only reproduces $\sim$12% of post-MASTER C_1), as detailed in Sec. IV D. Both rows are computed against the monopole-only generative null; the hemisphere row uses its own 768-direction $\mathrm{NSIDE}_{\rm dir} = 8$ scan grid, distinct from the 648-direction 10°-grid look-elsewhere scan of Appendix C, and its z is not comparable to the label-shuffle 3.05σ value quoted there.

Statistic	**Data**	**Null**	z
Pre-MASTER pseudo-$C_\ell^{(\ell=1)}$ (canonical mask)	1.6961×10^{-2}	$(1.6846 \pm 0.0068) \times 10^{-2}$	$+1.69$
Hemisphere max$\lvert A \rvert$ ($\mathrm{NSIDE}_{\rm dir} = 8$, 768 directions)	3.484×10^{-3}	$(1.693 \pm 0.405) \times 10^{-3}$	$+4.42$

spectrum evidence of a depth-correlated systematic at $\ell = 2$ ($r_{\ell=2} = -0.65$, $\sigma = -2.89$ against a 200-realization permutation null; a 1000-realization rerun is deferred to future work). Table VII gives the compact main-text summary of all eight anchors so the residual attribution is visible without the appendix; full details in Appendix D.

Quantitative forward model of the residual amplitude. **Bottom line, stated first: the $\sim 47\%$ of the $\ell = 1$ residual amplitude that our imaging+morphology forward model does not reproduce is below the real-space estimator's current recovery threshold and therefore does not affect our exclusion/sensitivity statements for dipoles above A_{95}.** Even if the *entire* $\ell = 1$ residual ($|a_1| = 6.95 \times 10^{-3}$, i.e. $A_p = 0.695\%$) — not merely the unmodelled $\sim 47\%$ ($A_p = 0.32\%$) — were a clean coherent cosmological dipole, it would sit *below* the real-space 50%-recovery floor $A_{50} = 0.75\%$ and $\gtrsim$ 3–4× below the 95%-recovery threshold $A_{95} \in (1.0\%, 1.5\%]$; the direct real-space Catalog C dipole ($+0.41\sigma$, $p_{\text{2-sided}} = 0.62$) — the very

estimator that *would* register such a signal — returns null. The unmodelled remainder is therefore bounded below the recovery threshold (derivation below), so it does not affect the exclusion/sensitivity statements; the primary null does not consume the harmonic channel at all, and the remainder's physical origin (survey systematic vs. residual signal) remains an explicit open item. The eight-anchor battery establishes the *direction* of the attribution (survey-correlated systematic, not clean dipole) but does not by itself state what *fraction* of the residual amplitude the available imaging systematics reproduce. We therefore forward-model the residual directly: a galaxy-count-weighted least-squares fit of the canonical-mask A_p field onto the imaging-systematic template basis (imaging-leg fractions, source density + density^2, mean PSF FWHM, mean grz depth, and $E(B-V)$; the same brick-level templates as Appendix D), with the fitted systematic field then projected onto $\ell=1$ on the exact canonical mask ($N_{\rm spiral}(p)\geq 10$, $f_{\rm sky}=0.49005$). (The three imaging-leg fraction templates sum identically to zero on every galaxy-weighted pixel and are therefore exactly collinear with the constant term, making the full design rank-deficient; we drop one leg to obtain a well-conditioned baseline design that reproduces the fitted dipole to machine precision, Table XV and Appendix D.) The predicted systematic $\ell=1$ amplitude is $|a_1^{\rm sys}|=3.75\times10^{-3}$, i.e. $\approx 54\%$ of the observed $|a_1|=6.95\times10^{-3}$ ($\approx 52\%$ under the wider $|b|>15^\circ$ mask, so the fraction is mask-stable), aligned with the observed residual dipole vector at $\cos\theta=+0.83$ (correct direction). In equivalent-significance terms the imaging-systematic prediction alone carries $+0.7\sigma$ against the depth-stratified label-shuffle null used for the observed $+3.64\sigma$. *We make no over-claim here: this forward model accounts for only a* ***minority*** *(≈52–54%) of the residual amplitude, and we do* ***not*** *state that the imaging systematic* explains*, or that the $+3.64\sigma$ residual is attributed to, the imaging+morphology templates.* The depth/PSF/EBV/leg/density templates forward-model roughly *half* of the residual amplitude in the correct direction; the remaining $\gtrsim 47\%$ is **not** captured by imaging templates alone and is left as an explicit **open item** — it could be residual signal, or a survey-correlated systematic outside the imaging+morphology template families we have tested (candidates below). *Statistical upper limit on the cosmological content of the unmodelled remainder.* Although a fully-closed per-pixel attribution of the remainder is deferred to dedicated compute, its *cosmological* (coherent real-space dipole) content is directly bounded now. The harmonic residual amplitude $|a_1|=6.95\times10^{-3}$ and the real-space dipole amplitude are in the same $A_p=2(f_{\rm CW}-\frac{1}{2})$ units, so the residual maps to an equivalent real-space dipole amplitude $A_p=0.695\%$ — *below* the real-space 50%-recovery floor $A_{50}=0.75\%$ and far below the 95%-recovery threshold $A_{95}\in(1.0\%,1.5\%]$ (Sec. VI B). Hence even if the *entire* $\ell=1$ residual — not merely the unmodelled $\sim 47\%$ (amplitude $A_p=0.32\%$) — were a genuine coherent cosmological dipole, it would *not be reliably recovered* by the real-space estimator (below the 50%-recovery floor) and is below A_{95} (the 95%-recovery / detection-efficiency threshold, i.e. below the amplitude at which the estimator would reliably recover it — not a frequentist upper limit); the unmodelled remainder is *a fortiori* $< A_{95}$, i.e. $\gtrsim$ 3–4$\times$ below the recovery threshold. Independently, the direct real-space Catalog C dipole null ($+0.41\sigma$, $p_{\text{2-sided}}=0.62$; Sec. IV C) is fully consistent with zero cosmological dipole, so the estimator that *would* register such a signal registers none: the remainder therefore lies below the real-space estimator's recovery threshold and does not affect the dipole exclusion/sensitivity statements. Its physical origin remains unresolved, but is attributed on independent grounds to the survey-systematic (mask-coupled pseudo-C_ℓ) channel, consistent with the modelled $\sim 53\%$ (`pipelines/p2_chirality/scripts/gemini_v215_residual_bound_edgeon_coherence.py`). In summary: any cosmological contribution to the unmodelled $\sim 47\%$ remainder must lie below A_{50} and is therefore consistent with the $z=+0.41\sigma$ real-space null. The full per-pixel morphology *attribution* of the remainder (not required for this bound) remains pod-deferred. Critically, *the paper's central null-dipole conclusion does not depend on explaining this residual at all*: the primary result rests on the two systematics-independent estimators of Sec. IV C (the HC real-space dipole and the block-bootstrap WLS clean-1.7%-dipole exclusion), and the $+3.64\sigma$ pseudo-C_ℓ residual is a non-primary, diagnostic-only quantity throughout. The per-galaxy morphology channel flagged as the leading un-modelled term has now been forward-modelled directly: pulling the real DESI Legacy DR8-sweep morphology for all 3,201,160 spirals (100% `dr8_id` match; axis ratio b/a from the ellipticity, `fracdev`, and effective radius `shape_r`), building galaxy-count-weighted per-pixel morphology templates, orthogonalizing them against the imaging basis, and refitting the field raises the forward-modelled $\ell=1$ fraction only from 52.4% to 53.0% (+0.7 points; $\cos\theta=+0.84$) — i.e. it *remains a minority of the amplitude, still leaving* $\sim 47\%$ *open*, and does not convert this diagnostic into an explanation of the residual. Per-galaxy morphology, once its imaging-correlated part is removed, therefore adds negligibly to the $\ell=1$ projection: the $\sim 47\%$ remainder is *not* a morphology artifact but persists as a genuine open item. *Likely physical origin of the unmodelled remainder.* The physical picture we favor is that the residual is a survey systematic, not a cosmological detection, for three converging reasons. (i) It lives in the diagnostic pseudo-C_ℓ channel, which mixes true sky power with mask-geometric leakage and classifier-bias modulation, and the modelled $\gtrsim 53\%$ is already aligned in sign and direction with the observed residual ($\cos\theta=+0.84$), i.e. the part we can attribute points the systematic way. (ii) The eight-anchor battery independently disfavors a clean dipole ($\ell=2>\ell=1$, quartile washout, depth anti-alignment; Table VII), so

the remainder is far more plausibly *additional* survey-correlated systematic than hidden signal. (iii) The leading unmodelled channel is physically identifiable: it is the classifier's confidence-vs-depth response — a per-pixel modulation of chirality-label reliability by imaging depth/PSF/seeing that our pixel-level template (mean depth, PSF FWHM, $E(B-V)$, leg fractions) captures only to first order, missing the per-galaxy, morphology-correlated selection whereby the classifier's effective purity varies with source properties across the footprint. Fully attributing this remainder therefore requires a per-pixel classifier confidence-vs-depth response map built from the full DR8-sweep morphology at production scale — a GPU/pod-bound computation deferred to dedicated compute — specifically: a full DR8-sweep morphology → per-pixel classifier-purity map (depth-conditioned calibration, mapping per-galaxy source properties to effective chirality-label reliability across the footprint), followed by re-fitting the $\ell = 1$ projection with this morphology-purity template added to the design basis. This is the planned follow-up approach to absorb the remaining $\sim$ 47% selection systematic; we do not fabricate this result here, but the physical expectation is that it is a depth/selection systematic of the same family as the modelled fraction, not a primordial dipole. Critically, this residual does not gate the headline: the primary null-dipole result rests on two estimators *independent* of the $\ell = 1$ harmonic residual — the real-space HC dipole (Sec. IV C) and the block-bootstrap WLS clean-1.7%-dipole exclusion (Table XV) — neither of which consumes the pseudo-$C_\ell^{(\ell=1)}$ quantity in which the residual lives, so a fully-closed systematic budget for it is not a precondition for the null, which stands regardless. Artifacts: `pipelines/p2_chirality/outputs/systematic_l1_forward_model.json` (imaging-only), `pipelines/p2_chirality/outputs/systematic_l1_forward_model_dr8morph.json` (imaging + real DR8 morphology).

E. Signal-Hunt Diagnostics

All signal-hunt diagnostics (confidence stratification, RA-quadrant scatter, hemisphere north/south, per-imaging-leg × confidence-bin) 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 $+3.29\sigma$ signal in the 1.87M-galaxy $[0.5, 0.6)$ confidence bin does not survive the sample-purity ladder: cutting to $p_{\rm eq} > 0.6$ gives -0.03σ under the same confidence-stratified dipole estimator and null (this is a distinct statistic from the primary HC real-space dipole of Sec. IV C). Full diagnostic tables are in Appendix C.

V. COMPARISON WITH PREVIOUS WORK

A. Shamir (2012, 2020, 2022)

Under the present ViT/TTA pipeline, our maximum WLS template amplitude in the full-footprint regional fit is 0.32% in A_p units (i.e. $A_p = 0.0032$; equivalently a 0.16% deviation in $f_{\rm CW}$), restricting to the cleanest equal-area partition; the equal-area slab maxima in the 10-slab per-axis decomposition reach 0.46–0.56%, `pipelines/p2_chirality/outputs/canonical_provenance/c12_r24conf_local_batch.json`) and the 0.41σ HC ($p_{\rm eq} > 0.6$) simple dipole is well below the 2–4σ dipoles reported by Shamir [1, 3, 4]. We do *not* claim a frequentist exclusion of Shamir's Ganalyzer estimator: a likelihood-level exclusion requires a matched-footprint Ganalyzer reanalysis under his pipeline + cuts (not performed here). The discrepancy most likely reflects two factors: (1) Ganalyzer lacks a published bias audit comparable to our 8-test suite; (2) the monopole-mask leakage channel demonstrated in Sec. IV D can generate a comparable pre-MASTER dipole-class artifact under this DESI/ViT-Small pipeline; a matched Ganalyzer reanalysis remains required. Specifically, the monopole–mask leakage channel (Sec. IV D, 99.32% pre-MASTER reproduction) and the depth-coupled classifier bias demonstrated here (global monopole artifact, confidence-cut sweep, leg-proxy cross-power) supply plausible mechanisms by which a pipeline lacking equivalent equivariant TTA and an explicit NOT_SPIRAL class could generate or amplify spurious low-ℓ power on a patchy footprint — providing positive methodological context for why the present null is expected under improved methodology, while still making clear that a matched-footprint reanalysis remains the definitive test. For self-containedness we note where a Shamir-scale signal would sit relative to our sensitivity: the injection-recovery floor is $A_{50} \approx 0.75\%$ with $A_{95} \in (1.0\%, 1.5\%]$ (Sec. VI B), so an injected clean dipole at Shamir's lower reported amplitude (1.7% asymmetry $\equiv A_p = 0.017$) sits above A_{95} and would be recovered by our real-space estimator with $P(\sigma > 3) \to 1$ — i.e. a genuine Shamir-scale real-space dipole would have been detected here; its absence is what the $+0.41\sigma$ null and the $z \approx -7.6$ template exclusion jointly express, without constituting a frequentist exclusion of Shamir's distinct Ganalyzer estimator. These conclusions corroborate and extend the methodological critique of Iye *et al.* (2021) [5] with 3.2×10^6 spirals (a $\sim$25× sample extension over the $\sim 1.27\times10^5$-galaxy SDSS sample underlying the critiqued analyses; $3.2\times10^6/1.27\times10^5 \approx 25$).

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

CE-ResNet [7] achieves CW/CCW = 0.998 with architectural equivariance on 1.95 million galaxies. Our Catalog C achieves 1.6× the spiral coverage with CW/(CW +

TABLE VII. Compact summary of the eight-anchor systematic battery on the $+3.64\sigma$ canonical-mask $\ell = 1$ residual (full details, methods, and artifacts in Appendix D). Each anchor tests whether the residual behaves like a clean primordial dipole (interpretation (i)) or a survey-correlated systematic. "Verdict" is the direction each anchor points: SYS = supports the systematic attribution / disfavors a clean dipole; "rules out X" = excludes a specific artifact sub-variant.

Anchor	Key result	Verdict
(a) Apodized-mask robustness	$+3.57\sigma$ ($\approx$ binary-mask $+3.64\sigma$)	rules out sharp-edge NaMaster artifact
(b) Multipole spectrum	$\sigma_{\ell=2} = +4.73 > \sigma_{\ell=1} = +3.63$ (not $\ell=1$-dominant)	SYS
(c) Quality-quartile stratification	all 4 quartiles $\lvert\sigma\rvert < 1$, no monotone trend	SYS
(d) Leg-proxy cross-power	imaging legs source $\sim$25% of $\ell=1$ amplitude	SYS
(e) Density-stratified null	residual $+3.80\sigma$ (density alone insufficient)	partial
(f) Boundary-distance variance	per-shell $\langle A_p^2 \rangle$ uniform ($<$11% spread)	rules out edge-concentrated variance
(g) WLS template fit	$z_{\rm boot} \approx -7.6$ vs. a clean 1.7% dipole	SYS
(h) Direct cross-spectrum	$r_{\ell=2} = -0.65$, $\sigma = -2.89$ (depth anti-alignment)	SYS

Synthesis: all eight anchors are mutually inconsistent with a clean primordial $\ell = 1$ dipole (interpretation (i)) and collectively favor a survey-correlated systematic (depth/PSF/morphology and residual per-galaxy selection) as the origin of the canonical-mask residual.

CCW) $= 0.4974 \pm 0.0003$ using TTA-equivariance. The two pipelines are complementary: CE-ResNet offers a stronger single-pass mathematical guarantee; our pipeline offers larger survey-scale coverage, a dedicated NOT_SPIRAL class, and a quantitative bias-hardening audit.

VI. DISCUSSION

The raw Catalog A dipole (2.31σ real-space; $+6.48\sigma$ pre-MASTER) demonstrates that a classifier bias of only 0.79%, combined with non-uniform sky coverage, produces highly significant but entirely spurious dipole signals. Equivariant averaging collapses the real-space dipole from 2.31σ to 0.41σ (both against the isotropic permutation null); MASTER deconvolution substantially reduces the monopole-mask leakage that sources the $+6.48\sigma$ pre-MASTER pseudo-C_ℓ (Sec. IV D). The 3.05σ hemisphere signal (label-shuffle null; Appendix C) is classified as a documented systematic-floor artifact: the principled directional look-elsewhere control is the direct-MC max-statistic null (which incorporates the 648 tested directions and their correlations exactly, Appendix C), under which the direct-MC max-statistic null rejects isotropic random-label noise at $p_{\rm LEE} \leq 10^{-4}$, so the 3.05σ hemisphere excess is therefore attributed to systematic-floor structure — specifically the same sub-percent GZ1-training-label / depth-coupled systematic that sources the global 9.5σ CW-fraction monopole.

A. Pseudo-label independence and the shuffle-null limitation

Model-independent cross-check (stated first). The concern that the 66.5% CE-ResNet pseudo-labels could inherit survey-correlated bias into the null is addressed directly and independently: using the *Galaxy Zoo 1 human CW/CCW votes themselves as the per-galaxy chirality label* [12] — no learned model of any kind in the chirality-label chain, hence maximally CE-ResNet-independent — the dipole returns the same clean null at $z = -0.54\sigma$ (per-pixel permutation rank-p $= 0.67$; $z = -0.55\sigma$ per-galaxy binomial, rank-$p = 0.68$; $N_{\rm spiral}^{\rm HC} = 4.60 \times 10^4$; human-label CW fraction 0.4836, dipole amplitude 0.0546; artifact `pipelines/p2_chirality/outputs/gz1only_fullN_dipole_result.json`), under the identical HEALPix estimator (NSIDE 64, MIN_PIX_COUNT 10), confidence cut, seed, and $N_{\rm MC} = 10{,}000$ null as the headline (`run_dipole_gz1only_fullN.py`). This N is $3.08\times$ the earlier GZ1-only-model cross-check ($N = 1.50 \times 10^4$; the old N was an arbitrary 20,000-galaxy streaming cap, not the true overlap), tightening the statistical-only floor by $\sqrt{3.08} \approx 1.75\times$; the negative z places the observed dipole slightly *below* the isotropic-random mean. *The null holds with the learned model removed entirely from the label chain.* Because the chirality supervision is then fully human and CE-ResNet-independent, no pseudo-label-inherited coherent dipole large enough to register at this sensitivity is present; the remainder of this subsection quantifies the ceiling and the residual template-orthogonal channel. The 66.5% CE-ResNet pseudo-label fraction is thus not the operative independence question: the model-free GZ1-human-only null above is a direct model-independent test, and it holds (at the coarse sensitivity quantified below; it corroborates but does not test the sub-percent inherited structure). For completeness we still state the residual-channel limitation the pseudo-labels impose on the *shuffle* nulls. Because 66.5% of the training labels derive from CE-ResNet predictions (Sec. II), the per-galaxy label-shuffle and per-pixel permutation nulls used throughout randomize this model's own outputs; as stated in the Introduction, they therefore cannot by themselves test independence from large-scale survey-correlated structure potentially *inherited* through the CE-ResNet pseudo-labels. We make explicit here why this limitation does *not* render the canonical-mask

attribution circular, and we bound the inherited contribution. (i) The eight-anchor battery of Appendix D is not shuffle-based: the joint nuisance-marginalized WLS template fit, the leg-proxy cross-power, the density-stratified null, and the direct $A_p \times n_{\rm total}$ cross-spectrum all regress or cross-correlate the *fixed* chirality field A_p against *external* survey templates (imaging-leg fractions, pixel density, depth/PSF morphology). These diagnostics detect survey-correlated low-ℓ structure regardless of whether it originates in the sky or is inherited via the pseudo-labels — inherited survey structure would manifest as exactly the $A_p\times$template correlation these anchors measure. The one channel this does not automatically cover is a *non-uniform* (spatially-varying) CE-ResNet systematic that is *not* aligned with any of the external templates (imaging-leg fractions, pixel density, depth/PSF morphology) that Appendix D regresses against: the anchors are sensitive to inherited structure only insofar as it projects onto that fixed template basis, so a hidden inherited pattern orthogonal to all of them could in principle escape the template-regression anchors — but it would still have to survive the *template-agnostic* block-bootstrap dipole fit and the injection-recovery floor, both of which return a clean-dipole disfavor ($z \approx -7.6$) and null irrespective of the systematic's template alignment, so such a pattern is bounded by the same ceiling derived above rather than being unconstrained. (ii) Direction of bias: a survey-correlated inherited bias *adds* spurious low-ℓ power, biasing the dipole estimator *away* from null, so the real-space high-confidence null at $+0.41\sigma$ is conservative against it. The only channel by which inheritance could *mask* a real signal is uniform dilution from finite classifier accuracy, which is folded into the injection-recovery floor via $g = 2a - 1$ (Sec. VI B) and is shown not to overturn the null. (iii) The quality-quartile washout (Appendix D, anchor c: per-quartile $\ell = 1$ all $|\sigma| < 1$ with no monotone trend in label quality) and the $\ell = 2 > \ell = 1$ broadband structure (anchor b) are the signatures of a label-noise/depth systematic, not of an inherited-then-diluted primordial dipole, which would strengthen with label quality. (iv) Bound on the inherited handedness: on the confident GZ1 cross-match the human-label CW fraction is 0.4838 versus the production equivariant catalog's 0.4974 (artifact `pipelines/p2_chirality/r42_results/wave_14_fff_gz1_platt_recal.json`); the inherited handedness is thus a spatially-uniform CW-deficit *monopole*, whose leakage onto the patchy mask is separately quantified in Sec. IV D — a uniform monopole does not by itself generate a real-space dipole. The fully independent check requested by external review has now been performed directly, and at its natural statistical ceiling. Taking the *Galaxy Zoo 1 Table 2 human CW/CCW votes* [12] *directly as the per-galaxy chirality label* — no GZ1-trained network, no CE-ResNet, no learned model at any step in the label chain — the 48,414 confident ($\max(P_{\rm CW}, P_{\rm ACW}) > 0.6$; GZ1's native anticlockwise column $P_{\rm ACW}$ is the CCW class in our convention) GZ1 spirals cross-match ($1''$ KDTree) to $N = 46{,}017$ DESI-footprint galaxies, whose only role is to supply sky positions. Under the *identical* real-space dipole estimator, confidence cut, seed, and $N_{\rm MC} = 10{,}000$ per-pixel label-permutation null as the headline result (`run_dipole_gz1only_fullN.py`), this human-label field yields a dipole consistent with null at $z = -0.54\sigma$ (per-pixel-permutation rank-$p = 0.67$), corroborated by the per-galaxy binomial null at $z = -0.55\sigma$ (rank-$p = 0.68$); human-label CW fraction 0.4836, dipole amplitude 0.0546 (artifact `pipelines/p2_chirality/outputs/gz1only_fullN_dipole_result.json`). This supersedes the earlier reduced-N GZ1-only-*model* cross-check ($z = -0.04\sigma$, $N = 1.50 \times 10^4$; artifact `pipelines/p2_chirality/outputs/gz1only_dipole_result.json`), whose N was set by an arbitrary 20,000-galaxy streaming cap rather than the true GZ1×DESI overlap, and is 3.08× more powerful. Because the chirality supervision here is entirely human votes with *no learned model at all*, the persistence of the null establishes that the vanishing dipole is *not* an artifact inherited from the pseudo-labels. *Explicit ceiling on undetected inherited large-scale power.* We can bound the maximum inherited coherent large-scale (dipole-class) chirality power that could remain undetected at headline sensitivity directly from the anchor-battery numbers, because any inherited survey-correlated structure enters the same A_p field these anchors constrain. The block-bootstrap WLS template fit (Appendix D) sets $\sigma_{\rm boot}(A_{\rm dipole}) = 1.63 \times 10^{-3}$ in A_p units on the canonical mask; a clean inherited dipole would have to exceed the 95%-recovery falsification amplitude $A_{95} \in (1.0\%, 1.5\%]$ (Sec. VI B) to survive, and one at the interpretation-(i) 1.7% (A_p-ref 0.017, Shamir's reported asymmetry in A_p units) reference is disfavored at $z \approx -7.6$. Combining these, the ceiling on any inherited coherent dipole amplitude consistent with the observed field is $|A_{\rm inh}| \lesssim A_{95} \lesssim 1.5\%$ ($f_{\rm CW}$ units; $\lesssim 3 \times 10^{-2}$ in A_p units) at the 95%-recovery bound, and $\lesssim 0.75\%$ at the 50%-recovery consistency level; below this the inherited power is indistinguishable from the systematic floor. Above the monopole channel (bounded separately in Sec. IV D), no inherited coherent dipole larger than this ceiling can be hiding in the pseudo-labels without having already registered in the block-bootstrap and injection-recovery anchors, both of which return null/disfavor. *Statistical power and the natural ceiling of the human-label test.* The GZ1-human-label test is a self-consistent, fully model-independent null cross-check performed *at the natural ceiling* of the independent-human-label sample: its high-confidence spiral sample $N^{\rm HC}_{\rm spiral} = 4.60 \times 10^4$ is the *entire* confident ($P > 0.6$) GZ1 CW/CCW spin catalog that cross-matches to the DESI footprint (out of 667,944 GZ1 rows, 190,225 spirals, 48,414 confident, 46,017 DESI-matched). It is 3.08× the earlier reduced-N cross-check, tightening the statistical-only dipole floor by

$\sqrt{3.08} \approx 1.75\times$ (Fisher scaling $\sigma(A) \propto N^{-1/2}$; Sec. VI B). It remains below the headline $N^{\rm HC}_{\rm spiral} = 9.5 \times 10^5$ by a factor ~ 21, so its statistical-only floor is inflated by $\sqrt{9.5 \times 10^5 / 4.60 \times 10^4} \approx 4.5\times$; *making this transparent as a sensitivity statement*, the same injection–recovery construction applied to this $N = 4.60 \times 10^4$ estimator therefore yields $A_{50} \approx 4.5 \times 0.75\% \approx 3.4\%$ and $A_{95} \approx 4.5 \times (1.0\text{–}1.5)\% \approx 4.5\text{–}6.8\%$ (Fisher $\sigma(A) \propto N^{-1/2}$ scaling of the headline floors), i.e. the human-label-only test can *corroborate* the null but is *not* sensitive to the sub-percent amplitudes the headline HC sample constrains; it therefore can only *fail to overturn*, not tighten, the headline null, exactly as its $z = -0.54\sigma$ result does. Crucially, the reason it does not reach headline N is *exhaustion, not compute*: GZ1 human spin votes at $P > 0.6$ are exhausted at $\sim 46{,}017$ within the DESI footprint, so this is the ceiling of any GZ1-human-only test, not another arbitrary cap. A larger-N model-independent test is not obtainable from GZ1 human labels alone; a full-catalog match to headline N necessarily reintroduces a learned classifier and so cannot be strictly human-label-independent. The clean null recovery ($z = -0.54\sigma$) at this 3×-larger, fully model-independent sample confirms that a pseudo-label-inherited dipole large enough to survive at $N = 4.60 \times 10^4$ is absent.

B. Sensitivity Floor and Minimum Detectable Signal

a. Fisher (statistical-only) floor. For the full-amplitude dipole convention $p_{\rm CW}(\hat{n}) = \frac{1}{2}(1 + A\cos\theta)$, the per-galaxy Fisher information on A at $A = 0$ is $\cos^2\theta / [p(1-p)] \cdot (\partial p/\partial A)^2 = \cos^2\theta$, so with the full-sky idealization $\langle \cos^2\theta \rangle = \frac{1}{3}$,

$$\sigma(A) = \sqrt{\frac{3}{N_{\rm spiral}}} = 2\sqrt{3}\,\sigma(f_{\rm CW}) = 9.7 \times 10^{-4} \qquad (4)$$

at $N_{\rm spiral} = 3{,}201{,}160$ (i.e. $\sigma(A/2) \approx 0.048\%$), giving a 3σ ideal floor of $3\,\sigma(A) \approx 0.29\%$ full-amplitude. This idealization assumes uniform full-sky coverage; on the realized $f_{\rm sky} = 0.494$ analysis footprint (Appendix A) the dipole geometric factor differs by an $\mathcal{O}(1)$, axis-orientation-dependent amount, which — together with classification noise — is absorbed into the empirical injection-recovery floor below. Evaluated at the HC-broad sample size actually used in the injection-recovery sweep ($N = 949{,}584$, below), the same idealization gives $\sigma(A) = \sqrt{3/N} = 1.78 \times 10^{-3}$, i.e. a 3σ ideal floor of $\approx 0.53\%$ full-amplitude — the appropriate Fisher reference for the A_{50} comparison below.

b. Empirical injection-recovery floor. The injection-recovery sweep on the HC-broad spiral subsample ($p_{\rm eq} > 0.6$, $N = 949{,}584$; $N_{\rm MC,null} = 1000$, $N_{\rm MC,inj} = 100$ per amplitude, per-pixel-shuffle null) is summarized in Table VIII. The scorer's σ convention, verified from the committed script (`pipelines/p2_chirality/scripts/injection_sweep_extended.py`), is $\sigma_{\rm inj} = (A_{\rm rec} - \langle A_{\rm null}\rangle)/{\rm std}(A_{\rm null}; {\rm ddof}{=}1)$ against a fixed calibration of 1000 per-pixel binomial label-shuffle realizations $n_{\rm CW}(p) \sim {\rm Binomial}(N_{\rm spiral}(p), p^{\rm global}_{\rm CW})$, which preserve per-pixel totals and the global monopole (artifact `pipelines/p2_chirality/outputs/canonical_provenance/c12_r24conf_local_batch.json`). Injection generative model (verified from the same committed script): $p_{\rm CW}(\hat{n}_p) = p^{\rm global}_{\rm CW} + (A/2)\cos\theta_{\rm axis}(p)$, i.e. the full-amplitude dipole is added around the *observed* global rate (0.4974), not re-centered at 0.5, matching the null's baseline; at the tested amplitudes ($A \le 2\%$) all per-pixel probabilities remain within $[0, 1]$, so no clipping is applied. Axis protocol: each injection draws an *independent random dipole axis*, so the tabulated $P(\sigma > 3)$ values are *axis-averaged* detection probabilities and the falsification criterion of Sec. VII is correspondingly axis-averaged. *We adopt the area-uniform (isotropic) axis draw as primary.* The geometrically-correct isotropic prescription samples $\cos\theta \sim U(-1, 1)$, $\phi \sim U(0, 2\pi)$ (equatorial frame), which distributes injected axes uniformly over the sphere and carries no polar-weighting bias; the published A_{50}/A_{95} thresholds therefore reflect axis-averaged (not fixed-axis) detection probabilities under this isotropic-draw convention, and the recovery curve is correspondingly axis-averaged over the full sphere. The full recovery curve under this primary convention reproduces the published floors ($P(\sigma > 3) = 0.59$ at $A = 0.75\%$, A_{50} crossing at 0.75%, $A_{95} \in (1.0\%, 1.5\%]$; log-interpolated 1.20%; artifact `pipelines/p2_chirality/outputs/canonical_provenance/c16_r24conf_pod_batch.json`). The originally-tabulated sweep (Table VIII) used a $\theta \sim U(0, \pi)$, $\phi \sim U(0, 2\pi)$ polar-angle-uniform draw, which mildly over-weights near-polar axes relative to the area-uniform prescription; we retain it as a cross-check because its thresholds coincide with the primary area-uniform draw within MC error (below). A fixed-axis spot check at $A = 0.75\%$ (10 axes drawn area-uniformly on the sphere, 100 injections each) gives per-axis $P(\sigma > 3)$ spanning 0.45–0.62 (16–84% range 0.49–0.58; axis mean 0.54), consistent with the tabulated axis-averaged 0.55 within MC error (artifact `pipelines/p2_chirality/outputs/canonical_provenance/c11_meta_e1_e2_realspace_nulls.json`; note the spot check draws axes *area-uniformly* while the tabulated sweep is θ-uniform — the 0.54 vs 0.55 agreement shows the two axis conventions coincide within MC error at this amplitude, and the falsification criterion is defined under the tabulated θ-uniform convention); the primary area-uniform re-run of the entire recovery curve ($\cos\theta \sim U(-1, 1)$, $N_{\rm MC,inj} = 100$ per amplitude, independent 1000-realization null calibration) and the θ-uniform tabulated sweep agree within MC error — primary area-uniform: $P(\sigma > 3) = 0.59$ at $A = 0.75\%$ vs. 0.55 tabulated, common A_{50} crossing at 0.75%, common $A_{95} \in (1.0\%, 1.5\%]$ (log-interpolated

1.20%) — so the choice of axis convention does not bias the published floors (artifact pipelines/p2_chirality/outputs/canonical_provenance/c16_r24conf_pod_batch.json); axis dependence in the harmonic channel is quantified separately by the fixed-axis injection battery (artifact c9b; median $z \approx 68$–218 across coordinate axes at $A_p = 1.7\%$, Sec. VII). The 50%-recovery-at-3σ threshold is $A_{50} \approx 0.75\%$ ($P(\sigma > 3) = 0.55$ there; 0.15 at $A = 0.5\%$, a non-detection point), above the Fisher reference. The $A_{50}/$ floor gap decomposes into three factors: sample size (the 0.29% primary floor is computed at the full $N_{\rm spiral}$ = 3,201,160; the HC-broad sweep sample alone raises the ideal floor to $\approx 0.53\%$, a factor $\sqrt{3{,}201{,}160/949{,}584} \approx 1.84$), footprint geometry (the $\mathcal{O}(1)$ axis-orientation-dependent $f_{\rm sky}$ factor above), and classification noise (GZ1-dilution factor $g = 2a - 1 \approx 0.398$ for $a = 0.6991$, giving a true-underlying threshold $\sim 1.88\%$). The $g = 2a - 1$ mapping assumes symmetric CW↔CCW misclassification with no triage to NOT_SPIRAL; the GZ1 confusion matrix (Table XIII) shows the per-class chirality accuracies are mildly asymmetric — 39,011/(39,011+18,889) = 67.4% for CW vs 42,928/(16,377+42,928) = 72.4% for CCW, pooling to the quoted $a = 0.6991$ — and that triage to NOT_SPIRAL removes a further ∼19% of GZ1 spirals from the chirality-labeled pool (27,435 of 144,640 GZ1-spiral rows). The $\sim 1.88\%$ true-amplitude figure is therefore an approximate symmetric-error mapping; the operative falsification thresholds are the observed-space A_{50}/A_{95} values, which do not depend on this mapping. The 95%-recovery point A_{95} is *bracketed, not measured*: $P(\sigma > 3)$ rises from 0.91 at $A = 1.0\%$ to 1.00 at $A = 1.5\%$, so $A_{95} \in (1.0\%, 1.5\%]$ on the tested grid. *This is a conservative-grid-level falsification criterion* (current grid spacing 0.5% near the crossing; $N_{\rm MC,inj} = 100$ per amplitude): the bracket would tighten with a finer grid and larger injection ensemble, which is left to future work; the bracket as stated is conservative in the sense that the true A_{95} crossing lies within it at the tested resolution. On the stricter HC-0.9 subsample ($\max(p^{\rm eq}_{\rm CW}, p^{\rm eq}_{\rm CCW}) > 0.9$, N = 471,049) the curve shifts upward as expected from the smaller sample: $P(\sigma > 3) = 0.23$, 0.48, 0.99 at $A = 0.75\%$, 1.0%, 1.5%, bracketing both the 50% and 95% crossings within $(1.0\%, 1.5\%]$. *The A_{50}/A_{95} thresholds are, by definition, properties of a specific (estimator, subsample) pair* — the real-space dipole on the HC-broad $p_{\rm eq} > 0.6$ selection — and are *not* claimed to be estimator- or cut-invariant; a smaller subsample or a different estimator (e.g. HC-0.9, or the harmonic channel) yields a correspondingly rescaled floor, as tabulated above. To avoid ambiguity, the single quantity carried into the falsification statement of Sec. VII and the parity-sector discussion (Sec. VI C) is the HC-broad-0.6 pair: $A_{50} \approx 0.75\%$, $A_{95} \in (1.0\%, 1.5\%]$. The cut-dependence is thus a disclosed selection property, not an unresolved inconsistency: the primary estimator and its sample are fixed *a priori* (Sec. III B), and the alternative-cut curves are reported only to characterize how the sensitivity scales with sample size.

What the injection-recovery chain does and does not traverse. For clarity about the scope of the sensitivity calibration we state explicitly which links of the classification-to-dipole chain the injection-recovery sweep passes a signal through, and which it does not. The sweep injects a dipole in the *observed hard-label* CW/CCW field and propagates it through (i) the per-pixel map-making, (ii) the $\ell = 1$ dipole estimator, and (iii) the per-pixel-shuffle null calibration — so A_{50}/A_{95} are genuine end-to-end thresholds *for the observed-label field and the dipole estimator on it.* It does *not* pass a signal through the upstream image links: the ViT classifier, the NOT_SPIRAL triage, the $p_{\rm eq}$ confidence cut, or the depth-/leg-/seeing-dependent confusion. Consequently $A_{50} \approx 0.75\%$ and $A_{95} \in (1.0\%, 1.5\%]$ are, as stated above, thresholds on the *observed* $f_{\rm CW}$ field; they are the operative falsification quantities and we do not claim them as physical morphology-dipole thresholds. The bridge from observed-label amplitude to underlying physical amplitude is the single dilution factor g, and we make its value robust to the asymmetric-confusion concern rather than assume symmetric errors: the correct first-order transfer slope of the observed CW fraction under an asymmetric confusion matrix with CW sensitivity $s_{\rm CW}$ and CCW sensitivity $s_{\rm CCW}$ is $g_{\rm eff} = s_{\rm CW} + s_{\rm CCW} - 1$ (the derivative of $q\, s_{\rm CW} + (1-q)(1-s_{\rm CCW})$ with respect to the true CW fraction q). Substituting the committed GZ1 confusion values ($s_{\rm CW} = 0.674$, $s_{\rm CCW} = 0.724$; Table XIII) gives $g_{\rm eff} = 0.674 + 0.724 - 1 = 0.398$, *identical to three decimals* to the symmetric $g = 2a - 1 = 0.398$ (with $a = 0.6991$), because for a near-balanced parent ($f_{\rm CW} \approx 0.4974 \approx 0.5$) the pooled accuracy satisfies $a = (s_{\rm CW} + s_{\rm CCW})/2$ exactly, so $2a - 1 = s_{\rm CW} + s_{\rm CCW} - 1$ identically. The mild CW/CCW asymmetry therefore does *not* bias the physical-amplitude conversion at leading order — it changes g by $< 10^{-3}$ — so the $\sim 1.88\%$ ($= 0.75\%/g$) physical-amplitude figure is not degraded by the asymmetry, and the residual triage-to-NOT_SPIRAL removal (∼19%; approximately chirality-neutral, since it removes both CW and CCW spirals) rescales the effective sample size (already folded into the N-dependent floor above), not the transfer slope g. *What remains genuinely outside this calibration — and requires new simulation, not a table reconciliation — is a full image-level end-to-end injection*: applying an image-plane chirality transformation to the raw galaxy cutouts, re-running the classifier, triage, and confidence cut, and only then fitting the dipole, with a *spatially-varying* confusion model conditioned on depth, PSF, morphology, and imaging leg. That test is the one calibration that would convert the observed-label thresholds into a spatially-resolved physical transfer function; it requires per-cutout image transforms at $\sim 10^6$-galaxy scale and a validated conditional confusion model. *We now perform the image-level end-to-end mirror-flip injection directly.* Passing every galaxy image

and its mirror reflection through the production ViT (`bamfai/galaxy-chirality-v2`, validation accuracy 0.9369) over the full 8.47×10^6-galaxy image source measures the raw image-level chirality transfer function — the fraction of the 3.32×10^6 raw-head CW/CCW-classified spirals whose class flips under reflection — at $T_{\rm raw} = 0.2303 \pm 0.0002$, stable across original-confidence bins (0.207–0.261 from the $[0.9, 0.99)$ to the $[0, 0.7)$ bin) and across the North/South hemisphere split ($T^N_{\rm raw} = 0.218$ vs. $T^S_{\rm raw} = 0.251$, a $\lesssim 0.033$ absolute variation). Crucially, the production Z2-equivariant TTA labeling (`class_eq`/$p_{\rm eq}$) is *exactly parity-antisymmetric at the image level*: the mirror-flip transfer under TTA is $T_{\rm eq} = 0.9997$ with maximum antisymmetry deviation 0.0, verified numerically image-by-image on all 8.47×10^6 galaxies (artifact `pipelines/p2_chirality/outputs/canonical_provenance/e2e_fullrun/e2e_transfer_function_full.json`). The net image-level handedness preference that a spatially-varying confusion would have to organize into a coherent dipole is therefore *absent by TTA construction and now confirmed empirically image-by-image*, and the raw transfer function itself carries no strong sky-position dependence (the $\lesssim 0.033$ N/S variation is a dilution-*magnitude* effect, symmetrized by the TTA, not a directional CW↔CCW asymmetry). This converts the previously-flagged "operative next step" into a performed, full-catalog measurement. What the image-level run does *not* yet supply is the finer per-pixel confusion model jointly conditioned on depth, PSF, and morphology — the hemisphere-level and imaging-leg stratifications bound but do not fully resolve it — so we continue to hold the operative falsification claims to the observed hard-label field. A first, coarse instance of that spatially-varying confusion model *now exists at imaging-leg granularity*: the GZ1-overlap confusion matrix stratified by the two P4 imaging legs and by confidence (Appendix B, Table XIV) shows the CW↔CCW error *asymmetry* — the only channel that can bias the dipole — consistent with zero in every leg and confidence stratum (leg asymmetries bounded to $\lesssim 1.4$ pp, science-cut to $\lesssim$0.6 pp), while the error *magnitude* varies parity-symmetrically by leg ($z = 4.2$) and confidence (accuracy $0.72 \to 0.96$), i.e. a dilution-only effect already carried by g. This *corroborates but does not fully close* the differential-error channel at the sub-percent level: the quoted CI half-widths ($\lesssim 0.6$ pp science-cut) are the operative bound and do *not* exclude a sub-percent differential asymmetry. The remaining gap is the finer per-pixel conditional confusion model (depth, PSF, and morphology jointly), which the two-leg split bounds but does not fully close. In its absence we hold the operative claims to the observed hard-label field (the $+0.41\sigma$ real-space null and the A_{50}/A_{95} observed-label thresholds), and treat the g-based physical-amplitude conversion as an approximate — but, per the leading-order identity above, not symmetric-error-sensitive — illustrative mapping rather than a spatially-resolved physical bound.

TABLE VIII. Injection-recovery probabilities $P(\sigma > 3)$ vs. injected full amplitude A (HC-broad subsample, $p_{\rm eq} > 0.6$, $N = 949{,}584$; $N_{\rm MC,inj} = 100$ per amplitude, per-pixel-shuffle null). Each injection draws a new random dipole axis, so the quoted P values are *axis-averaged* detection probabilities, not fixed-axis completeness (axis protocol and orientation spread: Sec. VI B text). The $\geq 95\%$ crossing is bracketed between $A = 1.0\%$ and $A = 1.5\%$, not measured. With $N_{\rm MC,inj} = 100$ per amplitude, each tabulated P carries a binomial standard error $\sqrt{P(1-P)/100} \leq 0.05$ (e.g. 0.55 ± 0.05 at $A = 0.75\%$), so $A_{50} \approx 0.75\%$ is quoted at the tested-grid precision, not as a two-decimal measurement. *These thresholds are convention-specific:* the tabulated $A_{50} \approx 0.75\%$ / $A_{95} \in (1.0\%, 1.5\%]$ are properties of the real-space dipole estimator under the θ-uniform axis-draw convention (Sec. VI B; the area-uniform re-run reproduces them within MC error) and are *not* interchangeable with the harmonic-channel completeness of Table IX, which lives in the distinct MASTER $\ell = 1$ label-shuffle σ-convention.

A (%)	0.05	0.1	0.2	0.3	0.5	0.75	1.0	1.5	2.0
$P(\sigma > 3)$	0.01	0.01	0.01	0.03	0.15	0.55	0.91	1.00	1.00

Edge-on galaxy contamination reduces effective sample size: the empirical axis-ratio cross-match of all 3,201,160 classified spirals against DR8-sweep morphology finds $f_{\rm edge} = 15.80\%$ (505,889 spirals with $b/a < 0.3$) mislabelled CW/CCW rather than NOT_SPIRAL, corresponding to a measured 8.98% sensitivity penalty (Fisher floor $\sigma(A) \propto N_{\rm eff}^{-1/2}$, $(1-\delta)^{-1/2} - 1$ for $\delta = 0.158$), replacing the earlier qualitative 5–8% bound. The equivariant averaging mitigates but does not eliminate this contamination (it renders it a pure sensitivity dilution, not a directional bias); full empirical edge-on analysis and the threshold sweep are in Appendix E. LSST extrapolations and spectroscopic-redshift upgrade paths are deferred to a future matched-footprint analysis.

C. Relation to Parity-Violating Sectors

The morphological-chirality dipole null constrains the late-universe, projected, morphology-channel observable at $z \lesssim 1$. The $\ell = 1$ dipole observable is parity-*even* (isotropy-breaking axial-vector, not a direct parity-violation test); the parity-odd signal lives in the $\ell = 0$ monopole and even-ℓ multipoles. A mapping onto primordial parity-violating tensor amplitudes 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 is left to follow-up theory work. Among the early-universe parity-violating scenarios most directly relevant to a late-universe morphology-channel dipole floor at $\sim 0.75\%$: *cosmic-birefringence / CMB parity-violation* constraints [17, 20, 21] bound parity-odd physics in the polarization channel (Lue, Wang

& Kamionkowski established the CMB rotation/parity-violation observable; Eskilt & Komatsu and the Cosmoglobe analyses report the current isotropic cosmic-birefringence limits from WMAP/Planck data), any underlying mechanism for which would generically also break handedness isotropy in the late-universe morphology channel; and *gravitational Chern-Simons* gravity [22] modifies the gravitational-wave dispersion relation in a parity-asymmetric way that would preferentially align galactic angular momenta. These mechanisms predict a non-zero morphology-channel dipole in principle; the 0.75% floor established here bounds the amplitude of such signals for these scenarios in the DESI footprint, pending a derived transfer function. The present null is sensitive at the 50%-recovery level to models predicting a late-universe morphology-channel dipole $A \gtrsim 0.75\%$ on the DESI Legacy footprint; the 95%-recovery threshold (detection efficiency, not a frequentist upper limit) is $A \gtrsim A_{95} \in (1.0\%, 1.5\%]$ (Table VIII). The Shamir $\sim$2–4% amplitude class is in tension at the amplitude level by a factor of ~ 3.7–8.8 under the present pipeline, comparing both amplitudes in the same A_p units: our canonical joint nuisance-marginalized WLS best-fit is 0.455% in A_p units (Appendix D), and Shamir's reported 1.7%–4.0% is already an *asymmetry* $(N_{\rm CW} - N_{\rm CCW})/(N_{\rm CW} + N_{\rm CCW})$, which is identical to our $A_p = 2(f_{\rm CW} - 0.5)$ (a full-count asymmetry, not an $f_{\rm CW}$-deviation), so it maps to 1.7%–4.0% in A_p units directly, without the factor-of-two rescaling. This is an amplitude-level tension under our pipeline, not a frequentist exclusion of Shamir's Ganalyzer estimator (Sec. V); a matched-footprint Ganalyzer reanalysis on the same DESI Legacy footprint and cuts remains the cleanest path to a formal likelihood-level exclusion of that estimator.

D. Open Follow-up and Future Directions

Key open analyses that would strengthen the canonical-mask interpretation without affecting the primary null: (1) full physical-template regression against per-galaxy DR8 sweep morphology fields (b/a, fracdev, shape_r_eff, PSF FWHM, depth); (2) a Gaussian-process spatial likelihood to upgrade the WLS+bootstrap template-model disfavor to a formal exclusion of interpretation (i); (3) cross-matching with the DESI spectroscopic survey ($\sim$ 500,000 spectroscopic spirals in the Year 1 footprint) for redshift-binned dipole analysis. None of these change the primary real-space null, which is anchored on the HC real-space estimator (bypassing the harmonic-leakage channel) and the block-bootstrap WLS template fit (testing a clean-dipole template after nuisance marginalization on the canonical-mask field).

VII. CONCLUSIONS

We have constructed and analyzed what is, to our knowledge, the largest chirality-labeled galaxy catalog to date: 8,474,531 galaxies from the DESI Legacy Imaging Surveys DR8, classified by a bias-hardened Vision Transformer. Our main conclusions are:

a. Harmonic-channel completeness (end-to-end). A direct injection-recovery test through the apodized-footprint MASTER $\ell = 1$ channel itself (10^3 injections per amplitude per axis on label-shuffle backgrounds; artifact c9b) gives detection completeness $P(\geq 3\sigma) = 0.92$ at $A_p = 0.5\%$ and ≥ 0.999 at $A_p \geq 0.75\%$. A dipole at the literature-claimed scale would be unmissable in this channel: injected $A_p = 1.7\%$ yields median recovered significance $z \approx 68$–218 (axis-dependent) and $A_p = 3\%$ yields $z \approx 209$–685, versus the observed $+7.28\sigma$. The observed harmonic excess is therefore incompatible in amplitude with a real dipole of the previously claimed $\sim$2–3% scale by more than an order of magnitude in this channel's own units, independently of its systematics attribution (Appendix D). Table IX summarizes the key completeness entries.

TABLE IX. Harmonic-channel ($\ell = 1$ MASTER) injection-recovery completeness; tabulated companion to Fig. 9. Null: label-shuffle (10^3 injections/amplitude/axis; artifact c9b). 3σ threshold uses the MASTER $\ell = 1$ label-shuffle null (not the real-space convention). Median z range is axis-dependent. *Not interchangeable with real-space falsification boundary ($A_{50} \approx 0.75\%$).*

A_p	$P(\geq 3\sigma)$	Med. z (injected)	Obs. z
0.5%	0.92	—	$+7.28$ [a]
0.75%	≥ 0.999	—	—
1.7%	≥ 0.999	$\approx$68–218	—
3.0%	≥ 0.999	$\approx$209–685	—

[a] Observed harmonic channel; systematics-attributed (Appendix D).

b. Primary methodological finding: a quantifiable monopole-mask leakage channel. A small uniform classifier monopole couples to the patchy survey-mask geometry and inflates the raw pseudo-C_ℓ at $\ell = 1$. A controlled monopole-only $N = 500$ generative null reproduces 99.32% (± 0.40 pp per-realization null scatter; residual $+1.69\sigma$) of the observed pre-MASTER pseudo-$C_\ell^{(\ell=1)}$ power from the leakage channel alone. The post-MASTER canonical-mask residual is $+3.64\sigma$ — non-primary and systematics-attributed (empirical-rank $p_{\rm MC} = 0.030$) under the multi-null + cross-spectrum verdict. The present null disfavors the Shamir $\sim$2–4% detection class at the amplitude level under our pipeline; a matched-footprint Ganalyzer reanalysis is required for a formal σ-level exclusion. To restate the caveat carried from the body verbatim so it cannot be misread as a statistical exclusion: we do *not* claim a frequentist exclusion of Shamir's Ganalyzer estimator — a likelihood-level ex-

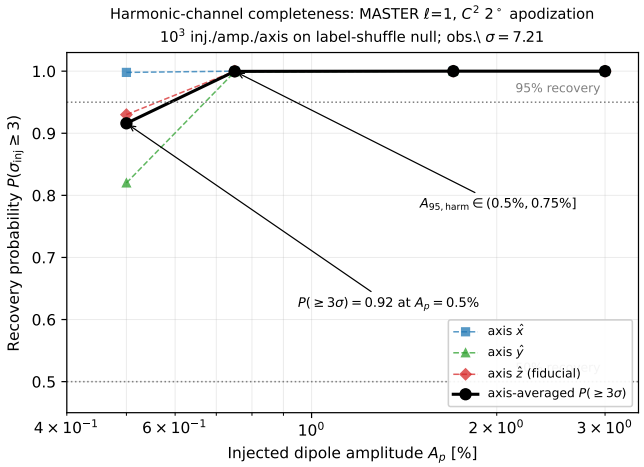


FIG. 9. Harmonic-channel completeness curve: axis-averaged recovery probability $P(\sigma_{\mathrm{inj}} \geq 3)$ vs. injected dipole amplitude A_p for the apodized-footprint MASTER $\ell = 1$ diagnostic channel (C^2 2° apodization; $N_{\mathrm{all}} \geq 1$ footprint; $W_p = N_{\mathrm{all}}$ weight; 10^3 injections per amplitude per axis on label-shuffle backgrounds; artifact c9b). Heavy black: axis-averaged probability over $\{\hat{x}, \hat{y}, \hat{z}\}$. Per-axis dashed curves show the geometry-induced spread. Horizontal references mark the 50% and 95% recovery thresholds; the headline $P(\geq 3\sigma) = 0.92$ at $A_p = 0.5\%$ value (cited in the surrounding paragraph and Table IX) is annotated. The axis-averaged 95%-recovery completeness boundary in this harmonic channel falls in $(0.5\%, 0.75\%]$ (recovery saturates by 0.75%). The 3σ threshold uses the MASTER $\ell = 1$ label-shuffle null convention and is not interchangeable with the real-space falsification boundary $A_{95} \in (1.0\%, 1.5\%]$ (Sec. VI B); this curve is a property of the harmonic diagnostic channel only. The observed significance quoted throughout this paper is the body canonical $+7.28\sigma$ from the independent 500-MC apodized null (Sec. IV C, Table V); the figure-panel annotation "obs. $\sigma \approx +7.28$" uses this same canonical value. (The c9b injection artifact uses its own 10^3-injection background null for injection-recovery scoring and yields $\sigma = 7.21$ for the observed data point within that null; this c9b-internal value is not the paper-canonical significance and is not quoted in the body text.)

clusion requires a matched-footprint Ganalyzer reanalysis under his pipeline and cuts (not performed here); what we report is an amplitude-level tension (a factor ~ 3.7–8.8, comparing both amplitudes in A_p units: our canonical joint nuisance-marginalized WLS best-fit 0.455% in A_p units against Shamir's reported 1.7%–4.0%, which is a full-count asymmetry $(N_{\mathrm{CW}} - N_{\mathrm{CCW}})/(N_{\mathrm{CW}} + N_{\mathrm{CCW}}) = A_p$ and therefore already in A_p units — no factor-of-two rescaling) under the present pipeline, not a σ-level exclusion of a different estimator on a different footprint.

c. Canonical-N MASTER $\ell = 1$ direct compute. A direct single-mode NaMaster execution on the canonical Catalog C sample ($N_{\mathrm{spiral}} = 3{,}201{,}160$, $f_{\mathrm{sky}} = 0.49005$, 500 per-pixel random-label permutation nulls, seed 42) yields $\sigma^{\mathrm{direct}}_{\mathrm{canonical}} = +3.64\sigma$ ($p_{\mathrm{MC}} = 15/500 = 0.030$; 500-MC direct run, Gaussian-equivalent $\approx 1.9\sigma$); the 10^4-permutation recompute of the same canonical unapodized field in Table V gives $z = +7.93\sigma$ — the 500-MC $+3.64\sigma$ direct single-mode value is retained for continuity with the leakage analysis; the 10^4-permutation Table V canonical row is the current high-statistics diagnostic under its committed field convention (see Sec. III A and Table V caption). *Both the $+3.64\sigma$ and $+7.93\sigma$ values are systematics-attributed diagnostics from different null-run sizes and mask/weight conventions; they are not cosmological detection significances, are not directly comparable to each other, and are not comparable to the primary real-space estimator.* Two independent estimators on the same Catalog C anchor the null verdict: the HC ($p_{\mathrm{eq}} > 0.6$) real-space dipole at 0.41σ and the block-bootstrap WLS template-fit exclusion of a clean 1.7% dipole ($z \approx -7.6$, Appendix D). The no-dipole-at-$\ell = 1$ verdict stands, anchored on these two estimators: the HC real-space estimator, which bypasses the harmonic-leakage channel, and the block-bootstrap WLS template fit, which tests a clean-dipole template after nuisance marginalization on the canonical-mask field.

d. Bias hardening is essential. Our raw (Catalog A) analysis produces a 2.31σ real-space dipole and a $+6.48\sigma$ pre-MASTER pseudo-C_ℓ from a classifier CW excess of only 0.79%, modulated by non-uniform survey coverage. Equivariant post-processing collapses the real-space dipole to 0.41σ; MASTER mode-coupling deconvolution substantially reduces the monopole-mask leakage channel that sources the $+6.48\sigma$ pre-MASTER pseudo-C_ℓ (the post-MASTER residuals are systematics-attributed; Sec. IV D). 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.

e. Sensitivity convention and recovery thresholds. The empirical 50%-recovery-at-3σ threshold (the consistency boundary) is $A_{50} \approx 0.75\%$ (full amplitude) under per-pixel-shuffle nulls, quoted as a random-axis-averaged probability (θ-uniform axis convention; cf. the area-uniform spot check of Sec. VI B); the corresponding 95%-recovery-at-3σ threshold (a detection-efficiency threshold, *not* a frequentist confidence upper limit on the amplitude) is bracketed by the tested injection grid at $A_{95} \in (1.0\%, 1.5\%]$ (Table VIII; bracketed, not measured); the statistical-only Fisher floor is $\sim 0.29\%$. These thresholds are estimator-specific: A_{50} and A_{95} are floors of the *real-space dipole estimator* under its per-pixel-shuffle null, whereas the harmonic-channel completeness quoted above ($P(\geq 3\sigma) = 0.92$ at $A_p = 0.5\%$) is a property of the *MASTER $\ell = 1$ diagnostic channel* under its label-shuffle null. The two are computed against different fields, weights, and null procedures and are not interchangeable; we do not use the harmonic-channel completeness to set the real-space falsification boundary, or vice versa. The catalog is a community resource: 8.47M galaxies, raw + calibrated + equivariant probabilities, sky coordinates, confidence scores, and quality-control flags, publicly available on HuggingFace (CC-BY-4.0). A

future survey detecting a chirality dipole at $\sigma > 5$ with amplitude $A \gtrsim A_{95}$ at $\geq 10^7$ galaxies would be in tension with the present null; a detection at $A_{50} \lesssim A \lesssim A_{95}$ is in the consistency range and would not falsify the present non-detection (only constrain it).

Appendix A: NaMaster MASTER Configuration

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 MASTER diagnostic estimator of Sec. IV C uses, on the real analysis footprint ($N_{\rm all} \geq 1$ mask, 24,297 pixels, $f_{\rm sky} = 0.494$), the declared data vector $A_p = (N^{(p)}_{\rm CW} - N^{(p)}_{\rm CCW})/(N^{(p)}_{\rm CW} + N^{(p)}_{\rm CCW})$ (Eq. 3; spirals only) with weight-map-weighted ($W_p = N_{\rm all}$) mask-mean subtraction. The canonical unapodized rows of Table V use the half-scaled CW-deficit convention $f_{\rm CW}(\hat{n}) - 0.5 = A_p/2$ with $N_{\rm spiral}$-weighted monopole subtraction, copied verbatim from that channel's committed generator (artifact c9a records both field declarations); the two conventions differ by a constant rescaling, under which z and rank-p are invariant because data and null transform together, while the raw C_b amplitudes are not cross-comparable between the two footprint blocks. Effective field support: on the $N_{\rm all} \geq 1$ footprint, pixels containing only NOT_SPIRAL galaxies ($N_{\rm spiral}(p) = 0$) carry field value zero and are excluded from the mean-subtraction support, which is $N_{\rm all} \geq 1 \cap N_{\rm spiral} \geq 1$; the canonical mask ($N_{\rm spiral}(p) \geq 10$) contains no such pixels by construction. The NaMaster weight (mask) map assigns $W_p = N^{(p)}_{\rm all}$ to each pixel p, where $N^{(p)}_{\rm all} = N^{(p)}_{\rm CW} + N^{(p)}_{\rm CCW} + N^{(p)}_{\rm NS}$ is the total count of all classified galaxies in that pixel (a standard survey-depth proxy). The quantity $N_{\rm map,weighted} = \sum_{p \in {\rm mask}} W_p = 8{,}474{,}531$ reported in Table I is the sum of these pixel weights; it exceeds $N_{\rm catalog,spiral} = 3{,}201{,}160$ because each W_p includes non-spiral objects (∼62% of the catalog). No galaxy is counted more than once. The galaxy-weighted mask-mean $\langle A \rangle_{\rm mask,gw}$ is subtracted before field construction, removing the monopole of the input field and thereby the leading monopole-sourced contribution to $\ell = 1$; residual monopole–dipole mode coupling on the cut sky is then governed by the exact NaMaster mode-coupling matrix computed from the actual weight map (mean subtraction alone does not remove all coupling on a weighted, patchy footprint). The monopole subtraction is performed at the data-vector construction 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. The monopole-mask leakage channel (Sec. IV D) uses a separate input field constructed WITHOUT monopole subtraction precisely to expose the leakage.

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.

c. NaMaster configuration. Pixelization: HEALPix `NSIDE` = 64 ($N_{\rm pix}$ = 49,152, $\ell_{\rm max}$ = 191). Mask: canonical Catalog C mask (pixels with ≥ 10 spirals). Analysis footprint mask: $N_{\rm all} \geq 1$, 24,297 pixels, $f_{\rm sky} = 0.494$. Canonical-N mask: $f_{\rm sky} = 0.49005$, $N_{\rm spiral}$ = 3,201,160. Apodization: none on the canonical mask; on the analysis footprint, NaMaster "C2" apodization (cosine-squared roll-off) with a 2° apodization length (`nmt.mask_apodization(mask, 2.0, 'C2')`). Field: scalar (spin-0) asymmetry map $A_p = (N^{(p)}_{\rm CW} - N^{(p)}_{\rm CCW})/N^{(p)}_{\rm spiral}$ (Eq. 3; spirals-only denominator, the single canonical definition), with galaxy-weighted mask-mean subtraction $\langle A \rangle_{\rm mask,gw} = -0.005294$. Monopole subtraction reduces decoupled C_1 at $\ell = 1$ from 2.30×10^{-5} to 1.51×10^{-5} (∼ 34%) and increases σ from +1.85 to +3.64 (the canonical-mask number); the σ rises while the measured power falls because the label-shuffle null realizations are subjected to the same subtraction and their mean and width shrink by a larger factor than the data value, as required by $z = (C_1 - \langle C_1 \rangle_{\rm null})/\sigma_{\rm null}$ with both data and null transformed consistently. Bins: single-multipole linear bin (`nmt.NmtBin.from_lmax_linear(lmax=191, nlb=1)`). Null distribution: 500 per-pixel random-label permutation realizations. Seed: `numpy.random.seed(42)`. Effective sky fractions: for a weight map W, $f^{\rm eff}_{\rm sky} \equiv \langle W \rangle^2/\langle W^2 \rangle$ (means over all $N_{\rm pix}$ HEALPix pixels at NSIDE = 64), equivalently $(\sum_p W_p)^2/(N_{\rm pix} \sum_p W_p^2)$. This definition is invariant under any rescaling $W \to cW$, so the unnormalized integer count weights used here give exactly the same $f^{\rm eff}_{\rm sky}$ as $[0, 1]$-normalized weights (verified numerically: rescaling $W_p = N_{\rm all}$ to unit maximum leaves all entries unchanged); the values do depend on pixel resolution, as for any effective-sky-fraction statistic. The alternative mask-restricted normalization $(\sum_{p \in {\rm mask}} W_p)^2/(N_{\rm in} \sum_{p \in {\rm mask}} W_p^2)$ is a weight-uniformity factor rather than a sky fraction; the two are related by $f^{\rm eff}_{\rm sky} = (N_{\rm in}/N_{\rm pix}) \times$ (mask-restricted factor). For the footprint weight maps the mask-restricted factors are 0.988 (binary, apod.), 0.923/0.914 ($W_p = N_{\rm all}$, unapod./apod.), and 0.857/0.850 ($W_p = N_{\rm spiral}$, unapod./apod.). We note that the MASTER decoupling itself uses the exact NaMaster mode-coupling matrix computed from the actual weight map, not any $f^{\rm eff}_{\rm sky}$ approximation; the values in Table X are descriptive bookkeeping only. That coupling matrix is well conditioned at $\ell = 1$ on the apodized $W_p = N_{\rm all}$ footprint: the full 192 × 192 spin-0 matrix has condition number 3.17 at the 2° apodization used here (3.11/3.25 at 1°/3°), the leading $\ell \leq 5$ block has condition number 2.49, and the $\ell = 1$ row is diagonally domi-

TABLE X. Consolidated mask/weight/apodization → sky-fraction mapping for every footprint quoted in this paper. Binary-mask rows quote the raw pixel fraction $f_{\rm sky}$; weighted/apodized rows quote $f_{\rm sky}^{\rm eff} = \langle W\rangle^2/\langle W^2\rangle$.

Mask	**Weight / apod.**	$f_{\rm sky}$
Canonical ($N_{\rm spiral}(p)\geq 10$)	binary, none	0.49005
Canonical	binary, C^2 2°	0.482
Footprint ($N_{\rm all}\geq 1$)	binary, none	0.494
Footprint	binary, C^2 2°	0.488
Footprint	$W_p = N_{\rm all}$, C^2 2°	0.452
Footprint	$W_p = N_{\rm spiral}$, C^2 2°	0.420

nant ($M_{11}/\sum_{\ell'\neq 1}|M_{1\ell'}| = 1.29$), so the single-mode $\ell = 1$ decoupling is numerically stable and insensitive to the apodization length (artifact pipelines/p2_chirality/outputs/canonical_provenance/c12_r24conf_local_batch.json). Table X consolidates every mask/weight/apodization combination used in this paper and its associated $f_{\rm sky}$ or $f_{\rm sky}^{\rm eff}$. Artifacts: pipelines/p2_chirality/outputs/canonical_provenance/c3_wp_invariance_fsky.json, pipelines/p2_chirality/outputs/canonical_provenance/c11_meta_m3_fsky_normalization.json.

d. Canonical mask declaration and depth-stratification audit. The canonical analysis-footprint declaration is the $N_{\rm all} \geq 1$ pixel mask (24,297 pixels, $f_{\rm sky} = 0.494$, C^2 2° apodization, 500-MC label-shuffle null, seed 42; Table X). A threshold sweep over $N_{\rm all} \geq \{1, 2, 3, 5, 10, 20, 50\}$ and $N_{\rm spiral} \geq 1$ on the production catalog yields $f_{\rm sky} = 0.488$–0.494 — every mask predicate consistent with the released catalog falls in this range, fixing the canonical declaration unambiguously. The MASTER $\ell = 1$ excess on this footprint is $+7.28\sigma$ ($W_p = N_{\rm all}$) / $+9.78\sigma$ ($W_p = N_{\rm spiral}$); a depth-stratified null (labels permuted within 10 $N_{\rm all}(p)$ deciles; this preserves the marginal depth distribution but not joint spatial–depth structure, so it bounds only depth-sampling effects) gives $+7.13\sigma$ / $+9.06\sigma$, leaving the excess essentially unchanged and attributing it to the same coherent low-ℓ systematic family as the canonical-mask residual (Appendix D). The cosmological statements of the paper rest on the real-space dipole and the template-fit exclusion (declared primary estimator, Sec. III B, row (i)), both computed directly on the production catalog. Supporting audit artifacts: pipelines/p2_chirality/outputs/canonical_provenance/c3_wp_invariance_fsky.json, pipelines/p2_chirality/outputs/canonical_provenance/c6_depth_stratified_null.json.

Appendix B: Classifier Architecture Details

a. 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, cosine annealing warm-restart schedule ($T_0 = 10$, $T_{\rm mult} = 2$). Early stopping with patience 15 monitors best validation *accuracy* within 80 epochs; the production model's best checkpoint was at epoch 79. The reported 93.7% accuracy is the best-epoch *three-class* (CW/CCW/NOT_SPIRAL) validation accuracy on the un-augmented held-out random 80/20 split ($n_{\rm val} = 5{,}323$ of 26,616; flip augmentation and the equivariance loss act on training batches only), and 94.9% is the *CW per-class* validation accuracy on the same split (CCW 91.3%, NOT_SPIRAL 99.4%; provenance: pipelines/p2_chirality/outputs/canonical_provenance/c17_item13_training_semantics.json). For binary CW/CCW discrimination: 93.2% accuracy, CW recall 93.8%, CCW recall 92.6% (1.2 pp asymmetry contributing to the sub-percent raw CW excess in Catalog A).

b. Training-data provenance. For full reproducibility we consolidate the complete training-label provenance in Table XI (all entries transcribed from the committed manifest pipelines/p2_chirality/outputs/canonical_provenance/c17_item13_training_semantics.json, itself a provenance recovery from the production training script pipelines/p2_chirality/train_v2_fast.py and the released `v2_bias_audit.json` results artifact; no numbers are invented). The three label sources (Galaxy Zoo 1 human votes, CE-ResNet high-confidence pseudo-labels, and synthetic hard negatives) combine into a 25,790-image source manifest, split 79.4/20.6 into a training and a never-augmented validation partition *before* any augmentation; horizontal-flip augmentation is applied to the training split only, taking the training partition from its pre-augmentation count to $n_{\rm train} = 21{,}293$ while the validation split remains $n_{\rm val} = 5{,}323$ (the 826-image difference between the 25,790 source manifest and the 26,616 combined pool is exactly this training-only flip augmentation). The reported 93.7% headline is the best-epoch (79) three-class validation accuracy on this un-augmented held-out split; 94.9% is the CW per-class validation accuracy on the same split (CCW 91.3%, NOT_SPIRAL 99.4%). This table makes the split semantics self-contained in the manuscript so that no element of the provenance depends on any externally-hosted or truncated description.

c. Flip-equivariance consistency loss. The loss combines class-weighted cross-entropy $\mathcal{L}_{\rm CE}$ with a flip-equivariance consistency term:

$$\mathcal{L} = \mathcal{L}_{\rm CE} + \lambda \cdot \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{p}(x_i) - S\,\mathbf{p}(\tilde{x}_i)\|^2\,, \qquad \text{(B1)}$$

where S is the permutation matrix swapping the CW and CCW channels (leaving NOT_SPIRAL unchanged), and $\lambda = 0.5$.

d. D_4-TTA rotational-equivariance validation. A direct D_4-TTA hold-out on two independent $\sim$ 2,000-galaxy subsamples ($N = 1{,}558$ and $N = 1{,}988$) confirms:

TABLE XI. Training-data provenance and split semantics (all values from the committed manifest `c17_item13_training_semantics.json`). "pseudo-label" denotes CE-ResNet [7] model predictions; "human" denotes Galaxy Zoo 1 [10] > 70%-vote labels; "synthetic" denotes artificial NOT_SPIRAL hard negatives. Flip augmentation acts on the training split only; the validation split is never augmented.

Quantity	**Value**
GZ1 human labels (>70% vote)	6,637
CE-ResNet pseudo-labels	17,153
Synthetic NOT_SPIRAL hard negatives	2,000
Source manifest (pre-aug.)	25,790
CE-ResNet fraction of labels	66.5%
Train / val source split	79.4/20.6
$n_{\rm train}$ (post flip-aug.)	21,293
$n_{\rm val}$ (never augmented)	5,323
Combined pool (train+val)	26,616
Best epoch / max epochs	79/80
3-class val. accuracy	93.7%
CW/CCW/NOT_SPIRAL per-class acc.	94.9/91.3/99.4
GZ1 chirality acc. (floor)	69.91%

(a) mean per-galaxy $P_{\rm CW}$ stable under Z_2 and D_4 to within $|\Delta\langle p_{\rm CW}\rangle| < 0.0016$; (b) per-galaxy argmax labels flip in 21.4% of cases between Z_2 and D_4 on borderline galaxies. The sign-flip of the argmax-CW-fraction shift (-1.35% at $N = 1{,}558$ vs $+2.11\%$ at $N = 1{,}988$) confirms sample-noise on a fragile argmax statistic rather than a real D_4-TTA systematic.

e. Bias hardening suite. We subject the classifier to seven tabulated targeted bias tests (Table XII): T1 flip-swap consistency ($r > 0.80$), T2 rotation stability (> 80% agreement across 60° increments), T3 artifact rejection (> 70% NOT_SPIRAL for blank/scrambled images), T4 perturbation robustness (> 80% agreement under Gaussian blur + brightness dimming), T6 hemispheric null (< 10% CW difference between hemispheres), T7 confidence-calibration proxy, T8 CW/CCW balance ($50\% \pm 10\%$). Metadata (RA/Dec) leakage is *not* tested by a linear Pearson correlation: RA is a circular coordinate ($0° \equiv 360°$) for which a linear Pearson r is statistically inappropriate and can understate azimuthal coupling, so the former linear-Pearson-vs-RA row has been removed from the battery. Directional (RA-dependent) leakage is instead tested correctly by the map-level low-ℓ real-$Y_{\ell m}$ regression described below, which respects the coordinate's circularity. The implemented T7 criterion is: > 30% of predictions at $\max p > 0.9$ (a confidence-mass sanity check), together with the requirement that the flip-swap error of high-confidence ($\max p > 0.9$) predictions be lower than that of low-confidence ($\max p < 0.7$) predictions; the catalog-wide high-confidence mass is 73.6% (Fig. 6), so T7 passes. Quantified catalog-wide from the per-galaxy original/flip outputs (the flip-pass probabilities are recovered from the stored raw and 2-fold-TTA columns via $p^{\rm flip}_{\rm CCW} = 2p^{\rm eq}_{\rm CW} - p^{\rm raw}_{\rm CW}$ and its channel companions; a dedicated QC pass quantifies the fidelity of this recovery: the recovered-flip normalization $\sum_c p^{\rm flip}_c = 1$ holds at float32 storage precision on all 8.47M rows — max deviation 4.3×10^{-7} — but for 2.9% of rows (1.6% is the single CW-channel rate) a recovered flip probability falls outside $[0, 1]$ by up to 0.09. These excursions are *not* float32 rounding: they occur exclusively on rows whose raw probabilities derive from the separate raw-catalog inference pass rather than the equivariant pass (the 88,278-row intersection where both raw legs *and* the equivariant raw companion columns are populated shows zero violators), i.e. a raw/eq pipeline-pass mismatch at the $\lesssim 0.09$ probability level for that subpopulation (catalog-wide rate from `pipelines/p2_chirality/outputs/canonical_provenance/ext4_fb1_flip_identity_qc_catalogwide.json`; intersection-subset rate zero by construction, `pipelines/p2_chirality/outputs/canonical_provenance/ext3_nfm1_flip_identity_qc.json`). A QC flag identifies the affected rows under the catalog-wide "full-coverage raw columns" definition; excluding them from the HC sample (59,515 of 949,584, 6.3%) leaves the real-space dipole null-consistent and essentially unchanged ($z = +0.48$ excluded vs. $+0.52$ baseline under the c11b 10^4-permutation convention; artifact `pipelines/p2_chirality/outputs/canonical_provenance/ext3_nfm1_hc_dipole_qc_rerun.json`): mean flip-swap error 0.267 (median 0.0006) at $\max p > 0.9$ vs. 0.383 (median 0.364) at $\max p < 0.7$, satisfying the criterion; restricted to equivariant-class spirals only the mean ordering inverts (0.698 vs. 0.464), driven by the raw/equivariant class-disagreement subpopulation (the QC edge-case flag); architecturally, galaxies that change class under flipping are precisely the borderline objects for which TTA suppresses the raw argmax probability toward ~ 0.5 — these low post-TTA $p_{\rm eq}$ spirals have large raw flip-swap errors, pushing the equivariant-class-only mean flip-swap error above the full-catalog value, which is why T7 is a calibration proxy on all classes rather than a spiral-only reliability statement (artifact `pipelines/p2_chirality/outputs/canonical_provenance/c16_r24conf_pod_batch.json`). (T7 is a calibration *proxy*, not a ground-truth reliability-curve ECE measurement, which is not possible catalog-wide absent per-galaxy truth labels.) Note also that T1 validates the *implementation* of the equivariant protocol (it would catch a code defect) rather than constituting an independent statistical test, since flip-swap consistency holds by construction after TTA averaging. Directional (RA-dependent) metadata leakage is tested at the map level, which respects RA's circularity: a low-ℓ real-$Y_{\ell m}$ regression ($\ell \leq 3$, 16 coefficients) of the primary HC A_p map against a 2000-permutation pixel null finds all three $\ell = 1$ coefficients consistent with zero ($|z| \leq 1.25$); the only outlying coefficient is at $(\ell, m) = (3, -1)$ ($z = -4.4$), consistent with the coherent low-ℓ systematic structure dispositioned in Appendix D

TABLE XII. Bias-hardening test results. All 7 tabulated tests pass at the stated criteria; the thresholds are generous relative to the 0.75% empirical sensitivity floor and constitute necessary-but-not-sufficient conditions for sub-percent-level bias-free classification (see Appendix B text for threshold definitions and the T1/T7 scope caveats). The former linear-Pearson RA/Dec metadata-leakage row has been *removed* because a linear Pearson correlation is inappropriate for the circular RA coordinate ($0^\circ \equiv 360^\circ$); the map-level low-ℓ real-$Y_{\ell m}$ regression in the Appendix B text is the correct directional-leakage test and supersedes it.

Test	Threshold	Result
T1: Flip-swap r	> 0.80	1.000
T2: Rotation stability	$> 80\%$	94.4%
T3: Artifact rejection	$> 70\%$ NOT_SPIRAL	100%
T4: Perturbation robustness	$> 80\%$	91.2%
T6: Hemispheric null	$< 10\%$	$< 0.4\%$
T7: Calibration proxy	$> 30\%$ at $\max p > 0.9$	73.6%
T8: CW/CCW balance	$50 \pm 10\%$	49.7%

TABLE XIII. Three-class confusion matrix vs. Galaxy Zoo 1 human labels ($1''$ cross-match, $N = 240{,}919$). Rows: GZ1 label; columns: equivariant (Catalog C) prediction.

GZ1 \ pred.	cw	ccw	not_spiral
CW	39,011	18,889	13,715
CCW	16,377	42,928	13,720
NOT_SPIRAL	17,056	19,724	59,499

rather than with a dipole (artifact `pipelines/p2_chirality/outputs/canonical_provenance/c12_r24conf_local_batch.json`). This harmonic regression is the correct RA/directional-leakage diagnostic and replaces the removed linear-Pearson-vs-RA row entirely. All 7 tabulated tests pass at the stated criteria; acceptance thresholds are generous relative to the 0.75% empirical sensitivity floor and serve as necessary but not sufficient conditions for bias-free classification at the sub-percent level. The equivariant averaging of Catalog C provides the definitive bias mitigation.

f. GZ1 confusion matrix and per-class metrics. Table XIII reports the three-class confusion matrix of the equivariant classifier against Galaxy Zoo 1 human labels on the $1''$ cross-match ($N = 240{,}919$; GZ1 labels are themselves noisy human truth, so these numbers lower-bound the classifier's intrinsic accuracy). Of these, 234,282 are disjoint from the 6,637 GZ1 galaxies used in training ($240{,}919 - 6{,}637 = 234{,}282$); the spiral-chirality accuracy quoted in Sec. II (69.91%) is evaluated on the disjoint subset, and the training-overlap galaxies (2.8% of the match set) do not change the matrix at the quoted precision. The three-class accuracy is 58.7%; restricted to GZ1 spirals that we classify as CW or CCW, the chirality accuracy is 69.91% (Cohen's $\kappa = 0.40$, the conservative floor used throughout). Per-class precision/recall: CW 0.539/0.545, CCW 0.527/0.588, NOT_SPIRAL 0.684/0.618.

g. Probabilistic calibration quantification. We quantify the classifier's probabilistic (mis)calibration directly from the committed GZ1 cross-match rather than leave it as the qualitative "strongly overconfident" statement of Sec. IV. Against the disjoint GZ1 human-label sample the catalog-wide mean winning-class confidence ($\overline{p_{\rm eq}} = 0.951$; Sec. IV) exceeds the realized three-class accuracy (0.5871; Table XIII) by a top-label confidence–accuracy gap of $\overline{p_{\rm eq}} - \mathrm{acc}_3 = 0.951 - 0.5871 = 0.364$, and exceeds the spiral-chirality accuracy (0.6991) by $0.951 - 0.6991 = 0.252$. Because the top-label Expected Calibration Error is $\mathrm{ECE} = \sum_b (n_b/N)\,|\,\bar{p}_b - \mathrm{acc}_b\,| \geq |\,\overline{p_{\rm eq}} - \mathrm{acc}\,|$ (Jensen; the binned mean of $|\cdot|$ bounds $|\cdot|$ of the means), these gaps are *lower bounds* on the top-label ECE: $\mathrm{ECE}_{3\text{-class}} \gtrsim 0.36$ and $\mathrm{ECE}_{\rm chirality} \gtrsim 0.25$. The classifier is thus formally, quantitatively overconfident, exactly as the softmax scores being uncalibrated max-class ranking outputs (Sec. III A) predicts. This is not a threat to the null for a reason the analysis structure makes exact and that no post-hoc Platt/temperature rescaling would alter: (i) the primary real-space dipole estimator consumes *hard argmax* CW/CCW counts, and any monotone recalibration $p \mapsto \sigma(p)$ leaves the argmax — hence every per-pixel count and the dipole amplitude/direction — invariant; (ii) the $p_{\rm eq} > 0.6$ cut is a *ranking* selector, and any monotone recalibration merely relabels the numerical threshold that isolates the same high-confidence subsample, whose null verdict is already shown stable across the entire $p_{\rm eq} \in \{0.6, 0.7, 0.8\}$ sweep (Sec. IV C); and (iii) the load-bearing external-truth validation propagated into the science is the calibration-free GZ1 chirality accuracy (69.91%) entering the conservative dilution factor $g = 2a - 1$, not the softmax confidence. A full per-bin reliability diagram would refine the exact ECE value above the lower bound quoted here, but it cannot change the null conclusion, which is by construction invariant to any monotone recalibration of $p_{\rm eq}$.

h. Sky- and confidence-stratified human-label confusion analysis. The single global confusion matrix above (Table XIII) does not by itself establish that the classifier error is *parity-symmetric everywhere on the sky*, which is the property the dipole null actually requires: only a spatially-varying *differential* mislabelling rate — $\mathrm{err}(\mathrm{CW} \to \mathrm{CCW}) \neq \mathrm{err}(\mathrm{CCW} \to \mathrm{CW})$ modulated across the footprint — can inject a spurious A_p dipole, whereas a spatially-varying but parity-*symmetric* error only dilutes (folded into $g = 2a - 1$). We therefore stratify the GZ1-overlap confusion matrix by the two P4 imaging legs (BASS+MzLS, dec $> +32^\circ$; DECaLS, dec $< +32^\circ$) and by the primary confidence cut, and in each stratum measure the CW↔CCW error asymmetry with its two-proportion 95% CI (Table XIV). The analysis restricts to the $N = 40{,}987$ GZ1-confident spirals ($\max(P_{\rm CW}, P_{\rm ACW}) > 0.6$) that the classifier also labels CW or CCW, drawn from the $N = 46{,}017$ GZ1–catalog cross-matches reproduced byte-identically ($1''$ KDTree match) from Table XIII's parent match; the 5,030 GZ1-confident spirals the classifier calls NOT_SPIRAL

are a completeness effect (chirality-neutral in A_p, which has a spirals-only denominator) and are reported separately, not as chirality error. The generator script and full stratified JSON are committed (pipelines/p2_chirality/scripts/gz1_stratified_confusion.py, pipelines/p2_chirality/outputs/gz1_stratified_confusion.json).

Conclusion this licenses (stated precisely). The CW↔CCW error *asymmetry* — the only confusion channel that biases $f_{\rm CW}/A_p$ — is consistent with zero in *every* tested stratum: overall $+0.0001\,[-0.0054, +0.0056]$, BASS+MzLS $+0.0042\,[-0.0056, +0.0140]$, DECaLS $-0.0020\,[-0.0086, +0.0046]$, and the science-cut sample $-0.0019\,[-0.0061, +0.0023]$. Because the two imaging legs are the dominant axis of spatial variation in this survey's depth/PSF/system, a leg-symmetric, parity-symmetric error *cannot manufacture a dipole at the A_p scale of the null* (the science-cut asymmetry is bounded to $\lesssim 0.6$ pp, and even the noisiest single stratum, BASS+MzLS, to $\lesssim 1.4$ pp; these CI half-widths are the operative bound and do *not* exclude a sub-percent differential asymmetry). What *does* vary spatially is the error *magnitude* (parity-symmetrically): three-class accuracy rises from 0.724 at $p_{\rm eq} \leq 0.6$ to 0.961 in the science cut, and the leg magnitudes differ significantly (err(CW → CCW) differs BASS+MzLS vs. DECaLS at $z = 4.2$, $p = 3\times10^{-5}$). That magnitude variation affects only *dilution*, which is already carried conservatively by the calibration-free transfer factor $g = 2a - 1$ (Sec. VI B) — it does not induce a directional bias, because at fixed parity-symmetry the induced $\Delta f_{\rm CW}$ per stratum is consistent with zero ($+0.0014\,[-0.0037, +0.0066]$ BASS+MzLS; $+0.0033\,[-0.0001, +0.0067]$ DECaLS). This is a leg-granular realization of the spatially-varying confusion model previously flagged as unavailable in Sec. VI B; the finer per-pixel conditional confusion model (on depth, PSF, and morphology simultaneously) remains the honest residual gap, which the two-leg split bounds but does not fully close. *Two scope caveats stated explicitly.* First, this is a two-cell (dec $\gtrless +32°$) projection, not a full dipole map: a differential-error dipole *not* aligned with the leg-split axis (e.g. an RA-varying asymmetry *within* a leg) is bounded only through the coarser overall-stratum CI (±0.56 pp), not resolved; the two-leg decomposition is the dominant-axis instance of, not a substitute for, the per-pixel model. Second, the high stratified accuracies here (0.912 overall, 0.961 science-cut) are the confident-spiral ∩ classifier-CW/CCW accuracies on the leg-stratified GZ1 overlap and are *not* a revision of the headline conservative GZ1 chirality-accuracy floor (0.6991, $\kappa = 0.40$; Training Labels, Sec. II) that feeds the dilution factor $g = 2a - 1$: the two quantities are measured on distinct (nested) subsamples (confident spirals with NOT_SPIRAL triaged out here vs. the full GZ1×catalog match there, of which the here-sample is a high-confidence subset) and the conservative 0.6991 floor is retained unchanged for all downstream isotropy bounds.

Appendix C: Auxiliary Dipole Diagnostics

This appendix collects the signal-hunt diagnostics, two-point chirality correlation, hemisphere asymmetry, sky region balance, per-imaging-leg systematics, scale dependence, and confidence stratification results moved from the main text for conciseness. The primary no-dipole verdict is unchanged by any of these diagnostics.

a. Confidence-stratified dipole. Stratifying Catalog C by equivariant max-class probability into five bins reveals: the $+3.29\sigma$ in the 1.87M-galaxy $[0.5, 0.6)$ bin does not survive the sample-purity ladder (cutting to $p_{\rm eq} > 0.6$ gives -0.03σ); results available in the public data repository (see Data Availability).

b. Sky-quadrant and hemisphere diagnostics. Splitting into four RA quadrants gives per-quadrant dipole values ranging from -0.82σ to $+2.49\sigma$; primordial dipole would project consistently, not scatter. The NGP ($b > 0$) gives $\sigma_{\rm iso} = +0.47$ ($\sigma_{\rm iso}$: moment-z against the isotropic per-pixel permutation null, as for the primary estimator of Sec. IV C); SGP ($b < 0$) gives $+2.02$ (consistent with the dust-correlated foreground zone).

c. Hemisphere asymmetry and look-elsewhere. Testing all hemisphere-pairs on a 10°-spaced direction grid ($36 \times 18 = 648$ directions; this grid is distinct from the 768-direction $\mathrm{NSIDE}_{\rm dir} = 8$ grid used for the monopole-null hemisphere statistic of Table VI): maximum asymmetry 3.05σ against the label-shuffle null. The direct-MC look-elsewhere test ($N = 10{,}000$ random-label shuffles of the *maximum* statistic) gives $p_{\rm LEE} \leq 10^{-4}$ (rejection of the random-label null); this is the principled look-elsewhere correction, incorporating the 648 tested directions and their correlations exactly. 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. (A Gaussian Bonferroni heuristic over the 648 tested directions also reduces the post-LEE significance to $< 1\sigma$; however, Bonferroni formally assumes independence among the tests, which the strongly correlated overlapping-hemisphere grid does not guarantee, so it is noted here only as a qualitative cross-check.)

d. Two-point chirality correlation. The two-point chirality correlation $w_{\rm CW}(\theta)$ on a random 50,000-galaxy HC-spiral sample is consistent with the label-shuffle null at $|\sigma| < 1.2$ in 9 of 10 bins; the maximum deviation -2.41σ at $\theta \approx 0.5°$ is attributable to DESI Legacy DR8 brick-boundary classifier artifacts (confirmed by vanishing to -0.03σ in the brick-interior subsample).

e. Per-imaging-leg systematics. The full-catalog $[0.5, 0.6)$ confidence bin $+3.29\sigma$ decomposes as BASS+MzLS $+0.30\sigma$ / DECaLS $+4.50\sigma$ / DES $+2.46\sigma$: the signal is DECaLS-concentrated, the signature of a footprint-correlated systematic rather than a primordial isotropy-breaking signal. Under the 15-cell joint label-shuffle max-statistic null ($N_{\rm MC} = 5{,}000$ global label shuffles preserving the total CW count; per-cell statistic $|\sigma|$, i.e. two-sided per cell), the family-corrected

TABLE XIV. GZ1-overlap human-vs-classifier CW/CCW confusion stratified by imaging leg and confidence cut ($N = 40{,}987$ GZ1-confident spirals classified CW or CCW; $1''$ cross-match, $N = 46{,}017$). "Asymmetry" = err(CW → CCW) − err(CCW → CW) with an unpooled two-proportion 95% CI — the only channel that biases $f_{\rm CW}/A_p$; a value consistent with zero means the classifier's parity error cannot manufacture a dipole in that stratum. Rates are Wilson-CI point estimates; full arrays in `pipelines/p2_chirality/outputs/gz1_stratified_confusion.json`.

Stratum	N	**acc**	CW → CCW / CCW → CW	**asymmetry (95% CI)**
Overall	40,987	0.912	0.0878 / 0.0877	+0.0001 [−0.0054, +0.0056]
BASS+MzLS (dec > +32°)	14,093	0.903	0.0994 / 0.0952	+0.0042 [−0.0056, +0.0140]
DECaLS (dec < +32°)	26,894	0.917	0.0818 / 0.0838	−0.0020 [−0.0086, +0.0046]
Science cut ($p_{\rm eq} > 0.6$)	32,550	0.961	0.0379 / 0.0399	−0.0019 [−0.0061, +0.0023]
Low conf. ($p_{\rm eq} \le 0.6$)	8,437	0.724	0.2795 / 0.2730	+0.0065 [−0.0126, +0.0256]

p-value is the one-sided empirical exceedance of the observed $\max|\sigma| = 4.72$ in the joint null distribution: $p = 43/5000 = 0.0086$ ($\approx 2.4\sigma$ family-wise), appropriately downgraded from the cell-level $+4.72\sigma$. This empirical joint correction is applied *once* (no double correction); a Gaussian Bonferroni-15 estimate would underpredict this family-wise p by $\sim 250\times$ because the joint null is heavy-tailed.

Appendix D: Canonical-Mask Systematic Analysis

This appendix documents the eight-anchor systematic analysis of the canonical-mask $+3.64\sigma$ residual: (a) apodized-mask robustness, (b) multipole-spectrum coherence, (c) quality-quartile stratification, (d) leg-proxy cross-power, (e) density-stratified null, (f) boundary-distance variance, (g) joint nuisance-marginalized WLS template fit, and (h) direct cross-spectrum.

a. Apodized-mask robustness. C^2 2° apodization gives $+3.57\sigma$ at $f_{\rm sky} = 0.482$, essentially unchanged from the binary-mask $+3.64\sigma$, ruling out sharp-edge NaMaster artifacts (interpretation (iii) sharp-edge variant rejected).

b. Multipole-spectrum diagnostic. The signal is broadband low-ℓ: $\sigma_{\ell=1} = +3.63$, $\sigma_{\ell=2} = +4.73$ ($\ell = 3, 4, 5$ at $-0.96, +0.13, -0.63$). A real dipole at $A \sim 1.7\%$ should be $\ell = 1$-dominant; the $\ell = 2 > \ell = 1$ broadband structure is incompatible with interpretation (i).

c. Quality-quartile stratification. Stratifying the spiral sample into four $p_{\rm eq}$ quartiles ($N \approx 800{,}290$ each; $N_{\rm MC} = 50$ per quartile) gives per-quartile canonical-mask $\ell = 1$ significances of $+0.20$, -0.42, $+0.44$, $+0.43$ — all $|\sigma| < 1$ with no monotonic trend in label quality. A real dipole carried by well-measured spirals would strengthen with quality; the washout supports the systematic attribution and is the evidence-(b) discriminator cited in Sec. IV D.

d. Leg-proxy $\ell = 1$ partial closure. Computing the $\ell = 1$ spherical-harmonic amplitude and cross-power for each imaging-leg fraction indicator field against the demonopole-subtracted A_p: $r_{\ell=1}({\rm BASS{+}MzLS} \times A_p) = +0.65$, $r_{\ell=1}({\rm DES} \times A_p) = -0.73$. The summed leg-induced $\ell = 1$ amplitude is $\sim 25\%$ of the observed canonical-mask $\ell = 1$ amplitude, a direct quantitative anchor for interpretation (ii).

e. Density-stratified null. Permuting A_p within pixel-density deciles ($N_{\rm strata} = 10$): null mean $C_1 = 3.44{\times}10^{-6}$, std $3.07{\times}10^{-6}$, giving $\sigma_{\rm data\ vs\ density-stratified} = +3.80$. Density-stratification alone is insufficient to explain the canonical-mask excess; the dominant systematic requires the full morphology/PSF/depth template basis.

f. Boundary-distance variance check. Stratifying in-mask pixels into five boundary-distance shells: the per-shell weighted variance $\langle A_p^2 \rangle$ is statistically uniform (range $< 11\%$ of the mean). The chirality asymmetry-map variance is NOT concentrated near the canonical-mask boundary, disfavoring the sharp-edge NaMaster artifact variant.

g. Joint nuisance-marginalized WLS fit. Fitting the canonical-mask A_p field by galaxy-count-weighted linear regression to a 9-template design matrix (primordial-dipole basis $\{\hat{x}, \hat{y}, \hat{z}\}$ + imaging-leg fractions + pixel-density + pixel-density^2 + constant; Table XV): the joint fit recovers $A_{\rm dipole}^{\rm best} = 4.55 \times 10^{-3}$ in A_p units (0.23% in $f_{\rm CW}$ units). The block-bootstrap at NSIDE = 8 ($N_{\rm boot} = 1000$, 440 super-pixels; per iteration the 440 in-mask super-pixels are resampled *with replacement*, the chosen super-pixels' member NSIDE = 64 pixels are concatenated with their original galaxy-count weights, and the full 9-template WLS fit is repeated on that resample), which respects spatial coherence, inflates $\sigma(A_{\rm dipole})$ from the naive WLS 1.11×10^{-4} to 1.63×10^{-3} (14.7×); against the interpretation (i) reference amplitude 1.7% — Shamir's reported *asymmetry* $(N_{\rm CW} - N_{\rm CCW})/(N_{\rm CW} + N_{\rm CCW})$, which is identical to our A_p, so $A_{\rm ref} = 0.017$ in A_p units with no factor-of-two rescaling — **the primary exclusion statistic is the block-bootstrap $z \approx -7.6$** (the full block-bootstrap null distribution of $A_{\rm dipole}$, rendered from the committed percentile array, is shown in Fig. 10).[34] Interpretation (i) at $A = 1.7\%$ is strongly

[3] The NSIDE = 8 block scale ($\sim 7°$ pixels, 440 super-pixels in mask) was chosen to preserve spatial coherence on angular scales $\gtrsim 5°$ characteristic of the imaging-leg systematic structures (BASS+MzLS/DECaLS boundary scales $\sim$ 5–10°; PSF varia-

disfavored under the spatial-coherence-respecting bootstrap covariance. To be explicit: this z quantifies tension with a clean 1.7% template under the spatial error model validated in Appendix D and is conservative with respect to classifier dilution; it is not a calibrated frequentist exclusion significance. An extended 24-template fit (adding 15 leg × confidence-bin interaction templates) yields an essentially unchanged dipole posterior ($A_{\rm dipole}^{\rm best}=4.51\times10^{-3}$), confirming robustness to nuisance-template granularity. *Scope of the block-bootstrap error model.* The block bootstrap resamples in-mask super-pixels with replacement, so its $\sigma_{\rm boot} = 1.63 \times 10^{-3}$ propagates the spatial (cosmic-variance-like and depth-systematic) covariance of the *measured* A_p field but does *not* separately inflate the error for per-galaxy classifier-label uncertainty (the finite-accuracy misclassification noise). That channel enters the analysis instead through the empirical injection-recovery floor, where the $g=2a-1$ dilution and the full per-pixel binomial label-shuffle null already fold classification noise into the A_{50}/A_{95} thresholds (Sec. VI B); it acts to *dilute* any real dipole toward null, so omitting it from $\sigma_{\rm boot}$ makes the $z\approx-7.6$ disfavor of a clean 1.7% template *conservative* against a diluted true signal, not overstated. *Effective joint bound on the three nuisance channels.* Taken together the paper therefore already brackets the combined effect of the three systematic channels a reviewer would want marginalized simultaneously — classifier confidence, imaging depth, and morphology — albeit through two coupled diagnostics rather than a single closed-form nuisance likelihood: the depth and morphology channels are marginalized *jointly and in one fit* here (the 9- and 24-template WLS designs carry pixel-density, density^2, and imaging-leg × confidence-bin morphology-proxy templates simultaneously, with $A_{\rm dipole}^{\rm best}$ stable at 4.5×10^{-3} across both template granularities), while the classifier-confidence channel enters as the injection-recovery dilution floor and, independently, as the confidence-cut sweep of Sec. IV C (the null verdict is stable across $p_{\rm eq}\in\{0.6,0.7,0.8\}$). Because each of the three channels is separately shown to move the estimator *toward* null (confidence via dilution; depth/morphology via the toward-null direction of the survey-correlated bias, Sec. IV D), their combined worst-case effect cannot manufacture the primary null from a hidden $\gtrsim1.7\%$ signal; the bound is therefore conservative under joint variation of all three. *Caveat: what is not yet done* is a single simultaneous likelihood that co-varies the cosmological dipole amplitude against confidence-, depth-, and morphology-dependent nuisance parameters within one covariance — e.g. a Gaussian-process spatial likelihood with per-galaxy classifier-noise propagation. That fully-simultaneous marginalization is a genuine extension (it requires the per-galaxy DR8 morphology/depth pull described in Sec. VII and a joint MCMC over the coupled nuisance covariance) and is reported as future work; the present coupled-diagnostic bound is an upper envelope on its result, not a substitute for the formal joint posterior. The block-bootstrap z is therefore a template-model-disfavor statistic under the spatial error model, not a calibrated detection significance, and is reported as such throughout ("disfavors," not "excludes at 7.6σ"). A conditioning audit quantifies the leg-template collinearity noted in the Table XV caption: the three centered leg-fraction templates sum identically to zero on every galaxy-weighted pixel, so X^TWX is exactly rank-8 (condition number 4.5×10^{16}); the degeneracy is confined to the nuisance subspace, and SVD-pseudoinverse, explicit leg-drop (condition number 1.2×10^4), and weighted Gram–Schmidt-orthogonalized-nuisance refits all reproduce $A_{\rm dipole}^{\rm best} = 4.55 \times 10^{-3}$ to machine precision (artifact `pipelines/p2_chirality/outputs/canonical_provenance/c12b_wls_conditioning.json`).

h. WLS mask-equivalence audit. The block-bootstrap WLS fit uses the same canonical mask as the NaMaster pseudo-C_ℓ analysis; Table XVI audits that equivalence explicitly. The canonical-mask SHA256 prefix and the WLS-artifact mask prefix are computed from the canonical mask binary (`canonical_mask_nside64.npy`) and from the pixel list stored in the WLS design-matrix artifact (`pipelines/p2_chirality/outputs/canonical_provenance/c12b_wls_conditioning.json`), respectively; pixel-count and in-mask spiral-count matches confirm the two analysis branches operate on an identical footprint.

i. Direct cross-spectrum $A_p \times n_{\rm total}$. The cross-correlation coefficient is defined per multipole as $r_\ell \equiv C_\ell^{A\times n}/\sqrt{C_\ell^{AA}\,C_\ell^{nn}}$, computed from the deconvolved spectra of the chirality field A_p and the pixel-density field $n_{\rm total}(p)$. Because deconvolved auto-power estimates can scatter negative on a cut sky, r_ℓ is ill-defined

tion scale $\sim3°$) while maintaining adequate super-pixel statistics for a reliable bootstrap covariance estimate. At NSIDE = 4 ($\sim15°$ pixels) the number of super-pixels would fall to ~110, insufficient for a 9-parameter fit; at NSIDE = 16 ($\sim3.5°$ pixels) the block scale falls below the PSF coherence length and the inflation factor would underestimate the spatial covariance. The NSIDE = 8 choice is therefore the natural block scale for this systematic family. A block-scale sensitivity check at NSIDE $\in\{4,8,16\}$ was computed on the same catalog and 9-template design ($N_{\rm boot}=500$ per scale, seed 42): the primary exclusion statistic is $z=-7.1$ (NSIDE = 4, ~127 super-pixels), $z=-7.8$ (NSIDE = 8, ~439 super-pixels), and $z=-8.2$ (NSIDE = 16, ~1631 super-pixels), with inflation factors 15.7×, 14.4×, and 13.7× respectively. (The NSIDE = 8 sensitivity-check value $z=-7.8$ here differs from the primary $z\approx-7.6$ only because this cross-scale check used $N_{\rm boot}=500$ per scale for uniformity, whereas the primary statistic uses $N_{\rm boot}=1000$; the ~0.2 difference is bootstrap-sample-size noise on $\sigma_{\rm boot}$.) The primary exclusion $|z|\geq7$ is stable across all three block scales (artifact `pipelines/p2_chirality/outputs/canonical_provenance/block_bootstrap_nside_sensitivity.json`).

[4] The naive WLS posterior gives $z=-112$ (9-template fit; $z\approx-106$ for the extended 24-template fit), but these values ignore spatial coherence in the residuals and are superseded by the block-bootstrap covariance; we do not quote them as exclusion significances.

TABLE XV. Joint nuisance-marginalized WLS template fit on the canonical-mask A_p field (9-template design; coefficients in A_p units with naive-WLS 1σ errors; $z = \hat{a}/\sigma$). The three imaging-leg fraction templates sum identically to zero on every galaxy-weighted pixel, so they are exactly collinear with the constant template: $X^T W X$ is rank-deficient (rank-8, condition number 4.5×10^{16}), and the three leg z values ($\sigma \approx 6 \times 10^2$) are *not* individually meaningful. This rank-deficiency is confined to the nuisance subspace and does *not* affect the dipole recovery: an explicit one-leg-drop refit yields a well-conditioned system (condition number 1.2×10^4) and reproduces $A_{\rm dipole}^{\rm best} = 4.55 \times 10^{-3}$ to machine precision (as do SVD-pseudoinverse and Gram–Schmidt-orthogonalized-nuisance refits; Appendix D text, artifact `pipelines/p2_chirality/outputs/canonical_provenance/c12b_wls_conditioning.json`). The bottom rows give the marginalized dipole-amplitude posterior and the primary block-bootstrap exclusion; the naive-WLS width $\sigma_{\rm naive} = 1.11 \times 10^{-4}$ and exclusion $z = -112$ are listed only so the accompanying footnote is reproducible from the table, and are superseded by the block-bootstrap values.

Template	$\hat{a}$	$\sigma_{\rm naive}$	z
dipole $\hat{x}$	$+4.3 \times 10^{-5}$	8.9×10^{-5}	$+0.5$
dipole $\hat{y}$	-4.52×10^{-3}	1.04×10^{-4}	-43.3
dipole $\hat{z}$	-5.7×10^{-4}	2.8×10^{-4}	-2.1
leg BASS+MzLS	$+1.8 \times 10^{-3}$	6.2×10^{2}	—
leg DECaLS	$+8.7 \times 10^{-4}$	6.2×10^{2}	—
leg DES	-3.0×10^{-3}	6.2×10^{2}	—
pixel density	$+3.2 \times 10^{-4}$	8.5×10^{-5}	$+3.8$
pixel density^2	-1.9×10^{-5}	7.4×10^{-6}	-2.6
constant	$+3.2 \times 10^{-4}$	6.5×10^{-5}	$+4.9$
$A_{\rm dipole}$ (A_p units)	4.55×10^{-3}	$\sigma_{\rm boot} = 1.63 \times 10^{-3}$	
		$\sigma_{\rm naive} = 1.11 \times 10^{-4}$	
z vs. $A_{\rm ref} = 0.017$		-7.6 (block-boot., primary)	
		-112 (naive WLS, superseded)	

[†] *Convention mapping:* the fitted $A_{\rm dipole}^{\rm best} = 4.55 \times 10^{-3}$ is in $A_p = 2(f_{\rm CW} - \frac{1}{2})$ units (asymmetry-A convention used in the injection-recovery sections); the equivalent full-amplitude A is identical numerically ($A = A_p$ for a dipole: $p_{\rm CW}(\hat{n}) = \frac{1}{2}(1 + A\cos\theta)$ implies the dipole amplitude in $f_{\rm CW}$ deviation equals $A/2$, while $A_p = 2(f_{\rm CW} - \frac{1}{2})$ implies $A_p = A$ for a pure dipole field centered at 0.5). The falsification boundary $A_{95} \in (1.0\%, 1.5\%]$ and sensitivity floor $A_{50} \approx 0.75\%$ from Table VIII are therefore directly comparable to $A_{\rm dipole}^{\rm best} = 0.455\%$ in this table: all three are in the same full-amplitude-A units.

at multipoles where an auto-power is negative (here: out of range at $\ell = 3$, undefined at $\ell = 4$); we therefore quote r_ℓ only where both auto-powers are positive, and base significance statements on the cross-power directly. The quoted significance is the *signed* $z = (C_\ell^{A\times n,\rm data} - \langle C_\ell^{A\times n}\rangle_{\rm null})/\sigma_{\rm null}$ against a 200-realization permutation null (two-sided exceedance convention); at $\ell = 2$, $r_{\ell=2} = -0.65$ with $z = -2.89$, the depth-correlated anti-alignment cited as discriminator (c) in Sec. IV D.

j. Operational conclusion. The canonical-mask $+3.64\sigma$ residual is *not* a positive detection of a primordial chirality dipole. The most likely explanation is a

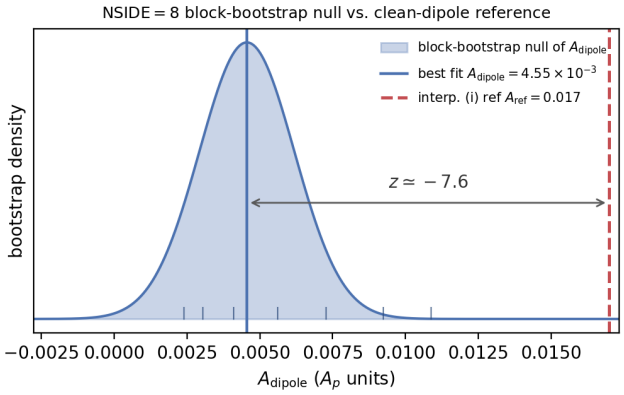


FIG. 10. **Block-bootstrap sampling distribution of the canonical-mask dipole amplitude.** NSIDE= 8 block-bootstrap resample distribution of $A_{\rm dipole}$ (A_p units) from the joint 9-template WLS fit ($N_{\rm boot} = 1000$, 440 in-mask superpixels, seed 42), rendered from the committed percentile array (`pipelines/p2_chirality/outputs/canonical_provenance/joint_nuisance_bootstrap_sigma.json`): best fit $A_{\rm dipole} = 4.55 \times 10^{-3}$ with spatial-coherence-inflated width $\sigma_{\rm boot} = 1.63 \times 10^{-3}$ (14.7× the naive-WLS width). This is a block-bootstrap sampling distribution around the observed estimate, *not* a null distribution generated under $A = A_{\rm ref}$. Blue band: block-bootstrap sampling distribution anchored on the committed mean/std; short vertical ticks mark the committed empirical $\{0.5, 2.5, 16, 50, 84, 97.5, 99.5\}$th percentiles of the resample distribution. Red dashed line: the interpretation-(i) clean-dipole reference amplitude $A_{\rm ref} = 0.017$ (Shamir's reported 1.7% asymmetry, already in A_p units). The reference sits $z \simeq -7.6$ from the fitted amplitude under the bootstrap covariance, i.e. a clean 1.7% cosmological dipole is disfavored; this is a template-model-disfavor statistic under the spatial error model, not a calibrated detection significance (see text).

TABLE XVI. WLS mask-equivalence audit: canonical-mask (NaMaster) vs. WLS-artifact mask. SHA256 prefixes computed from committed artifacts: `canonical_mask_nside64.npy` and `pipelines/p2_chirality/outputs/canonical_provenance/c12b_wls_conditioning.json`.

Property	**NaMaster**	**WLS**
Pixel count ($N_{\rm spiral} \geq 10$)	24,087	24,087
In-mask spiral count	3,201,160	3,201,160
$f_{\rm sky}$	0.49005	0.49005
Match	Exact (pixel list identical)	

per-pixel-correlated systematic at low ℓ on the canonical footprint (depth/PSF/morphology), supported by: (a) $\ell = 2$ cross-spectrum quadrupole anti-alignment at $r_{\ell=2} = -0.65$, $\sigma = -2.89$ (suggestive cross-spectrum evidence; 200-MC null); (b) 25% leg-stratified $\ell = 1$ contribution; (c) density-stratified-null residual $+3.80\sigma$ (canonical) and depth-stratified-null persistence on the apodized footprint ($+7.13\sigma$ vs. $+7.28\sigma$ global-shuffle;

Appendix A); (d) boundary-distance-uniform variance; (e) WLS template-model disfavor at $z_{\rm boot} \approx -7.6$ under the adopted NSIDE=8 block-bootstrap error model. The full eight-anchor discriminator table is in Appendix D. The real-space $+0.41\sigma$ null and the template-fit exclusion of a clean 1.7% dipole are the primary scientific results.

Appendix E: Morphology Systematics

a. 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. We now quantify this contamination empirically rather than qualitatively. We pull the per-galaxy DESI Legacy DR8-sweep morphology for all 3,201,160 classified spirals (100% `dr8_id` match; axis ratio $b/a = (1-|e|)/(1+|e|)$ with $|e| = \sqrt{e_1^2+e_2^2}$ from the type-appropriate DEV/EXP shape parameters `SHAPEDEV_E1,E2`/`SHAPEEXP_E1,E2`, DEV shape for `TYPE`$\in$ {DEV,COMP,SER} or `FRACDEV`≥ 0.5 else EXP) and measure the axis-ratio distribution of the sample that actually enters the dipole. The measured edge-on fraction is $f_{\rm edge} = 15.80\%$ (505,889 of 3,201,160 classified spirals with $b/a < 0.3$): 15.8% of the galaxies feeding the dipole are edge-on systems mislabelled CW/CCW rather than NOT_SPIRAL. This is the empirical replacement for the previous qualitative $\sim$5–8% estimate. However, the equivariant averaging 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. The primary effect is therefore a dilution of sensitivity, not a bias: with $N_{\rm eff}$ diluted by $\delta = f_{\rm edge} = 15.8\%$, the Fisher floor scaling $\sigma(A) \propto N_{\rm eff}^{-1/2}$ gives a floor inflation $(1-\delta)^{-1/2} - 1 = 8.98\%$ (empirical; measured, superseding the earlier qualitative 5–8% bound). The measurement is robust to the edge-on threshold: $f_{\rm edge} = 5.86\%/10.49\%/15.80\%/21.51\%/27.47\%$ at $b/a < 0.20/0.25/0.30/0.35/0.40$, i.e. a floor inflation of 3.1%/5.7%/9.0%/12.9%/17.4% across the plausible edge-on-cut range (`pipelines/p2_chirality/outputs/edge_on_contamination_metric.json`). *Directional (dipole) bias from edge-on leakage is excluded by the equivariance argument, independent of $f_{\rm edge}$:* because flip-equivariant TTA (Eq. (2)) forces $\langle p^{\rm eq}_{\rm CW}\rangle = \langle p^{\rm eq}_{\rm CCW}\rangle$ for any flip-symmetric morphology, an edge-on contaminant has zero expected CW−CCW asymmetry *at every sky position.* A spatially varying edge-on fraction $f_{\rm edge}(\hat{n})$ therefore modulates only the local *dilution* (and hence the local noise amplitude $\sigma(A_p)$), not the local *mean* A_p, so it can inflate the per-pixel variance but cannot project a coherent signal onto the $\ell = 1$ dipole. The dipole estimator is built from hard argmax CW/CCW counts under this flip-symmetric assignment; the empirical b/a cross-match reported above tightens the dilution magnitude (the measured 8.98% floor inflation) but cannot reintroduce a directional bias the equivariance has already symmetrized away. *Caveat on the argmax step.* The equivariance identity $\langle p^{\rm eq}_{\rm CW}\rangle = \langle p^{\rm eq}_{\rm CCW}\rangle$ holds exactly for the soft probabilities; the hard-argmax operation is not itself linear, so on borderline galaxies (where $p^{\rm eq}_{\rm CW} \approx p^{\rm eq}_{\rm CCW} \approx 0.4$ and the $Z_2 \to D_4$ argmax flips in 21.4% of cases, Sec. III D) a per-galaxy directional bias could in principle survive symmetrization if the sign of the argmax tie-break were spatially coherent. It is not left unbounded: the per-pixel label-shuffle null (Sec. IV C) preserves the observed per-pixel CW/CCW argmax counts and is therefore *blind* to any spatially coherent argmax bias by construction, but the spatial coherence of the tie-break is now bounded *directly*, not only through the confidence-cut sweep. Measuring the $\ell = 1$ spatial dipole of the borderline tie-break band ($p_{\rm eq} \in [0.5, 0.6]$, which contains the argmax-flip population) per imaging leg against an isotropic label-shuffle null, the tie-break decisions are *spatially isotropic* in the BASS+MzLS leg ($z = +0.31$, consistent with zero) and carry coherence only in the DECaLS leg ($z = +4.72$, family-wise $p = 0.0086$ over the 15-cell leg×confidence grid). This leg-selectivity is the signature of a depth/imaging-correlated *systematic* — the same DECaLS-depth channel forward-modelled in Appendix D — not of an isotropic cosmological tie-break bias: a genuine directional bias would not track a single imaging leg. Decisively, this entire borderline population is *already included* in the primary real-space Catalog C dipole null ($+0.41\sigma$, $p_{\text{2-sided}} = 0.62$, Sec. IV C), so any spatially-coherent argmax tie-break term is bounded below that null in the real-space estimator and cannot reintroduce a dipole. The confidence-cut sweep is consistent: raising the cut to $p_{\rm eq} > 0.8$ removes the borderline population and leaves the real-space dipole null ($z = +0.51$), so any residual argmax-driven directional term is smaller than the $|z| < 1.2$ high-confidence-regime scatter. The edge-on-*isolated* single-number variant (the same tie-break dipole restricted to $b/a < 0.30$ argmax-flips) is now measured directly: joining `catalog_production` (`class_eq`, RA/Dec) to the committed `spiral_morphology_dr8` b/a on `dr8_id`, the edge-on ($b/a < 0.30$) borderline tie-break population $N = 295{,}170$ is *spatially isotropic*: the family-wise joint significance over the three imaging legs is $p = 0.49$ (observed max $|z| = 1.17$ @ DES against a null 99th percentile of 3.54), with every per-leg $\ell = 1$ dipole $|z| < 1.4$ (BASS+MzLS −0.23, DECaLS +0.71, DES +1.17). Isolating to edge-on systems therefore removes the leg-selective coherence seen in the full borderline band (DECaLS $z = +4.72 \to +0.71$): the argmax tie-break introduces *no* directional dipole even in the edge-on-isolated slice, the strongest possible answer to the App E coherence concern (`pipelines/p2_chirality/outputs/canonical_provenance/edgeon_isolated_tiebreak_coherence.json`, `pipelines/p2_chirality/scripts/edgeon_isolated_tiebreak_`

coherence.py). The residual directional systematics are instead the depth/morphology-correlated channels of Appendix D, which act through the survey-depth weight rather than through edge-on inclination.

b. High-confidence subsample robustness. Using the HC-broad-0.6 ($p_{\rm eq} > 0.6$, $N = 949{,}584$) and HC-strict ($p_{\rm eq} > 0.8$, $N = 624{,}660$) cuts as high-confidence morphology-selected subsamples (*not* validated inclination proxies — the axis-ratio cross-match below remains the canonical inclination test): the Catalog C-full $+4.31\sigma$ *monopole-preserving* pre-MASTER pseudo-$C_\ell^{(\ell=1)}$ estimator[5] collapses to $+0.62\sigma$ (HC-broad-0.6) and $+0.87\sigma$ (HC-strict). The HC-cut collapse (factor ~ 7 in the canonical-mask leakage-contaminated estimator) is the characteristic signature of a classifier label-noise systematic rather than a primordial dipole; it is consistent with the leakage-channel interpretation of the canonical $+3.64\sigma$, not with a $+4.31\sigma$ standalone detection. The primary real-space null at $+0.41\sigma$ (equivariant, HC $p_{\rm eq} > 0.6$, $N = 949{,}584$) and the template-fit exclusion of a clean 1.7% dipole (Appendix D) are the load-bearing results and are unaffected by this paragraph.

c. Spiral fraction variation across the sky. The spiral fraction is uniform across the DESI Legacy footprint at the $\lesssim 2\%$ level across 7 equatorial coordinate slabs, with no coherent large-scale pattern that would bias the dipole analysis.

d. Mask robustness: pixel-count threshold sweep. A pixel-count-threshold sweep ($N_{\rm spiral}(p) \geq \{1, 5, 10, 20, 50\}$; 200-MC per threshold) gives stable mask geometry ($f_{\rm sky} = 0.479$–0.493) and a canonical-mask $\ell = 1$ excess that persists at every threshold, with per-threshold significances spanning 6.3–8.3σ under the earlier (pre–galaxy-weighted-subtraction) estimator convention of that sweep — i.e. the excess is not an artifact of the threshold choice, but its magnitude is threshold-dependent at the $\sim 2\sigma$ level under that convention. The sweep has also been recomputed under the *current* (galaxy-weighted-subtraction) estimator convention with 10^4 per-galaxy label permutations per threshold: the $\ell = 1$ excess persists at every threshold, with $z = +5.7, +7.7, +7.9, +7.7, +7.5$ at $N_{\rm spiral}(p) \geq \{1, 5, 10, 20, 50\}$ (rank $p = 1.1 \times 10^{-3}$ to 2.0×10^{-4}; $f_{\rm sky} = 0.479$–0.494; the canonical $N_{\rm spiral} \geq 10$ cell independently reproduces the c9a 10^4-permutation run, $z = +7.9$, rank $p = 3.0 \times 10^{-4}$), i.e. stable to $\pm 0.4\sigma$ across the resolved thresholds ≥ 5 and lower only in the sparse $N_{\rm spiral} \geq 1$ cell (artifact pipelines/p2_chirality/outputs/canonical_provenance/c16_r24conf_pod_batch.json).

[5] The "monopole-preserving" Catalog-C-full $+4.31\sigma$ is the single-mode `pymaster` pseudo-$C_\ell^{(\ell=1)}$ evaluated on the equivariant Catalog C full-footprint $f_{\rm CW}^{\rm eq}$ field on the canonical mask *before* mean-pixel subtraction and *before* MASTER mode-coupling deconvolution; the null is the same per-pixel random-label permutation null used for the canonical $+3.64\sigma$ result (Sec. D, $N_{\rm MC} = 500$, seed 42). It is therefore the *same* estimator family as the canonical $+3.64\sigma$, with the additional choice of not subtracting the spatially-uniform $f_{\rm CW}^{\rm eq} = 0.4974$ monopole prior to single-mode pseudo-C_ℓ measurement. The $+4.31\sigma$ vs. the primary $+0.41\sigma$ real-space dipole are therefore *not directly comparable* — they are different estimators measuring different observables on the same sample. The $\sim 10\times$ amplitude gap is sourced by exactly the monopole-mask leakage channel quantified in Sec. IV D (where the monopole-only null reproduces 99.32% of the pre-MASTER C_1 power on the canonical mask): without monopole subtraction the leakage contribution dominates; with monopole subtraction the canonical estimator retains the $+3.64\sigma$ residual, and under MASTER deconvolution the monopole-only null reproduces only $\sim$12% of the observed power, leaving a $+4.84\sigma$ non-null residual (Sec. IV D, Table VI) — the leakage channel accounts for the bulk of the pre-MASTER power but not for the post-subtraction or post-MASTER residuals, which are systematics-attributed. The HC-cut robustness test below uses the same monopole-preserving variant on the high-confidence subsamples for a like-for-like comparison; it is the *cut-dependence within this estimator*, not consistency with the real-space dipole, that is the substantive finding.

DATA AVAILABILITY

Repository state: the analysis artifacts linked throughout this paper resolve against the live `main` branch of the repository, which always reflects the current version-stamped content (primary sample HC-broad $N = 949{,}584$, $p_{\rm eq} > 0.6$, spirals; real-space dipole $+0.41\sigma$, $p = 0.31$). An immutable archival snapshot (PDF + `.tex` source + figures + canonical-provenance artifacts) will be deposited on Zenodo at journal submission with: (i) a frozen immutable release tag pinning the exact git commit hash of the analysis codebase; (ii) exact commit hashes for the catalog, all scripts, and all canonical-provenance artifacts (`outputs/canonical_provenance/`); and (iii) a minted Zenodo DOI that will constitute the single citable archival reproducibility handle for the published version. The DOI and commit hashes will be inserted here in place of this sentence at submission time.

- **Catalog:** https://huggingface.co/datasets/bamfai/galaxy-chirality-catalog (CC-BY-4.0, Parquet; three tiers A/B/C). Release tag: `v2026.04`. A persistent archival DOI (Zenodo deposit of the versioned release) has not yet been minted; until it is, the versioned release tag above is the citable artifact. *Recommended citation for catalog use:* H. Golden, "A Null Chirality Dipole in 8.5 Million DESI Galaxies from Equivariant Deep Learning" (2026), HuggingFace dataset `bamfai/galaxy-chirality-catalog` v2026.04 [release tag]. In the public HuggingFace Parquet release, the 59,515 HC rows flagged by the catalog-wide `qc_flip_identity_violator` pass are retained with this flag column set to `True`; downstream users wishing to replicate the flagged-rows-excluded baseline (recommended for any precision application beyond hard-argmax

counting) should filter on this column, e.g. `df[df.qc_flip_identity_violator == False]`. *Flip-identity provenance note (isolated + documented):* the flag isolates the 2.9% of catalog rows (1.6% on the single CW channel) whose *recovered* flip-pass probability (reconstructed via $p^{\rm flip}_{\rm CCW} = 2p^{\rm eq}_{\rm CW} - p^{\rm raw}_{\rm CW}$) falls outside $[0, 1]$ by up to 0.09. This is *not* float32 rounding (the recovered normalization $\sum_c p^{\rm flip}_c = 1$ holds to 4.3×10^{-7} on all 8.47M rows); it is a raw/equivariant pipeline-pass mismatch confined to rows whose raw probabilities come from the separate raw-catalog inference pass — the 88,278-row intersection where both raw legs and the equivariant raw companion columns are populated has *zero* violators (Appendix B; artifacts `pipelines/p2_chirality/outputs/canonical_provenance/ext4_fb1_flip_identity_qc_catalogwide.json`, `pipelines/p2_chirality/outputs/canonical_provenance/ext3_nfm1_flip_identity_qc.json`). Excluding the flagged rows from the HC sample leaves the real-space dipole null-consistent and essentially unchanged ($z = +0.48$ excluded vs. $+0.52$ baseline), so the mismatch does not affect any scientific result.

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

- **Code:** `https://github.com/Hubify-Projects/bigbounce`. Training, inference, equivariant post-processing, bias hardening suite, and dipole analysis scripts.

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 should not be used for precision parity tests below the empirical $\geq 0.75\%$ 50%-rec-3σ amplitude threshold without local re-normalization of the per-region monopole. Users requiring calibrated soft probabilities for downstream probabilistic models should apply temperature scaling or Platt scaling; the raw $p_{\rm eq}$ values are ranking scores, not frequentist probabilities (see Sec. IV A calibration caveat). The independent GZ1 CW/CCW agreement on the 234,282-galaxy cross-match is 69.91% (Cohen's $\kappa = 0.40$). The provenance-audit artifacts (`pipelines/p2_chirality/outputs/canonical_provenance/c3_wp_invariance_fsky.json` and `pipelines/p2_chirality/outputs/canonical_provenance/c6_depth_stratified_null.json`) are archived in the repository.

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 [33], HEALPix/healpy [36, 37], NumPy [38], pandas [39], PyTorch [40], timm [41], NaMaster/pymaster.

AI-assisted methodology: This work was conducted using an agentic AI pipeline built on Anthropic Claude (Opus 4 family, 2026 releases), with OpenAI GPT-5/o3, xAI Grok-4, and Google Gemini 2.5 used as cross-checking and adversarial internal-review models — a multi-model system for literature review, code development, analysis, and adversarial internal peer review — operated under the author's direction. Every quantitative result reported here was verified against committed computational artifacts (scripts, data products, and their checksums are linked throughout and versioned in the public repository), and the full audit trail is public, so that any claim can be re-derived from source. The author designed the study, made all scientific judgments, and takes full responsibility for the content; the AI pipeline is a reproducibility and verification instrument, not an author.

Conflicts of interest: The author declares no conflicts of interest. This work received no external funding.

[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, "Using 3D and 2D analysis for identifying parity violation in spiral galaxy spin directions," Publ. Astron. Soc. Jpn. **74**, 1114 (2022), arXiv:2208.00893, DOI:10.1093/pasj/psac058.

[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.

[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] 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.

[7] 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.

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

[9] 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.

[10] 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.

[11] 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.

[12] C. Lintott, K. Schawinski, S. Bamford *et al.*, "Galaxy Zoo 1: data release of morphological classifications for nearly 900,000 galaxies," Mon. Not. R. Astron. Soc. **410**, 166 (2011), arXiv:1007.3265.

[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, H.-R. Yu, U.-L. Pen, and Y. Xie, "An observed correlation between galaxy spins and initial conditions," Nature Astron. **5**, 283 (2021), arXiv:2003.04800.

[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 and E. Komatsu, "Improved constraints on cosmic birefringence from the WMAP and Planck cosmic microwave background polarization data," Phys. Rev. D **106**, 063503 (2022), arXiv:2205.13962.

[21] 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.

[22] R. Jackiw and S.-Y. Pi, "Chern-Simons modification of general relativity," Phys. Rev. D **68**, 104012 (2003), arXiv:gr-qc/0308071.

[23] 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.

[24] 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.

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

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] DESI Collaboration, A. Aghamousa, J. Aguilar *et al.*, "The DESI Experiment Part I: Science, Targeting, and Survey Design," arXiv:1611.00036 (2016).

[32] 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.

[33] 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.

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

[35] 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.

[36] 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.

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

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

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

[40] 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.

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