GENSYMAUDIT: ANCHOR-RELATIVE AUDITING OF DISTRIBUTIONAL SYMMETRY IN GENERATIVE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative models are increasingly deployed through fixed sampling and conditioning pipelines, yet structural properties of their output distributions are often inferred from the training data or ancestor model rather than measured on the deployed artifact itself. We study this problem through distributional symmetry and introduce GENSYMAUDIT, a sample-only framework for auditing whether a deployed generative artifact changes the element-wise symmetry profile of the real distribution it is intended to model. This problem cannot be addressed by testing generated samples against perfect symmetry: real data may already contain substantial directional structure, while the deployed distribution additionally depends on the model checkpoint, conditioning protocol, sampler, guidance, and sampling budget. GenSymAudit constructs condition-matched real and generated pools, measures transformation-specific separability using grouped classifier two-sample tests, and assigns anchor-relative verdicts using negative controls, calibrated uncertainty, multiplicity correction, and quality gates. Both fMoW and Git-10M exhibit strong departures from perfect D_4 invariance. Relative to these real-data anchors, DiffusionSat exhibits three statistically resolved amplifications, dominated by vertical reflection ($+18.11$ points), while Text2Earth shows smaller bidirectional changes, attenuating horizontal reflection (-2.98 points) and amplifying vertical reflection ($+3.62$ points). Controlled few-step experiments further show that likelihood-derived symmetry certificates can become numerically unreliable while independent sample-space measurements continue to detect asymmetry. These findings support the broader conclusion that structural properties should be measured on the complete artifact actually deployed rather than inferred from its reference data, ancestor model, or nominal sampling configuration.

1 INTRODUCTION

Generative models are increasingly deployed as complete artifacts defined not only by model weights, but also by conditioning, sampling, guidance, and inference budget. Yet structural properties of their output distributions are often inferred from the training data or model design rather than measured on the deployed system itself. Such inference can be unreliable: visually plausible generations do not necessarily imply distributional fidelity, and recent studies have documented systematic discrepancies between real and generated data (Bohacek et al., 2026). This raises a basic question: does a deployed generative artifact preserve the structural properties of its reference distribution?

We study this question through distributional symmetry. Symmetry is not merely a geometric property: transformation invariance underlies common data-augmentation strategies, equivariant learning, and assumptions about directional generalization (Cohen & Welling, 2016; Mitton & Murray-Smith, 2021; Lawrence et al., 2026). A generator that selectively amplifies or suppresses directional structure can therefore alter the effective training distribution seen by downstream models even when its samples remain visually plausible. We view such changes as a form of structural distributional distortion. Symmetry is particularly amenable to auditing because the relevant transformations are known and can be applied directly to samples.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

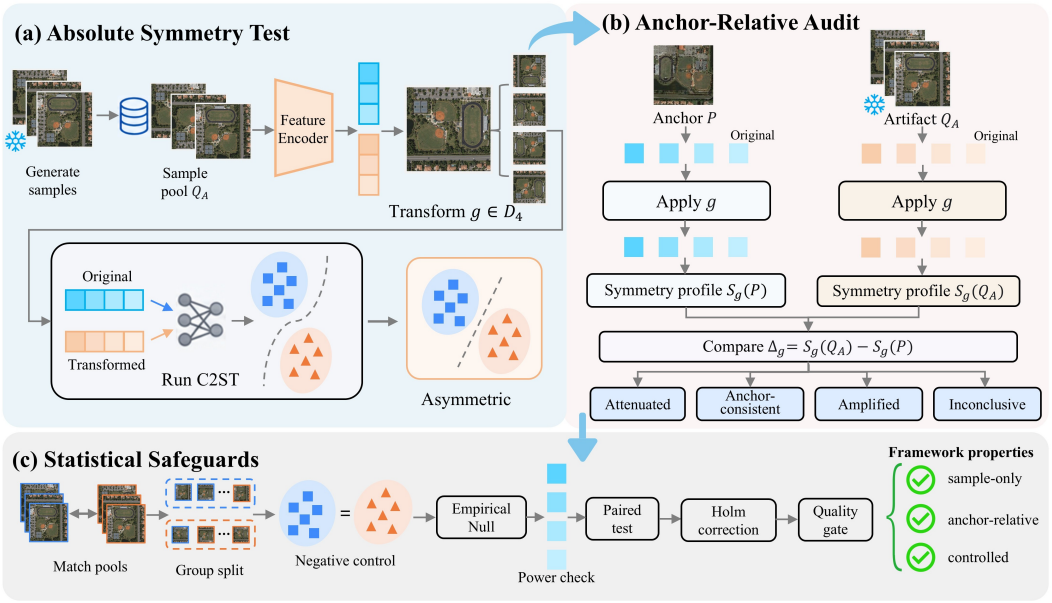


Figure 1: **Overview of GENSYMAUDIT.** Absolute testing detects asymmetry in Q_A but cannot distinguish inherited structure from artifact-associated change. GENSYMAUDIT compares the element-wise symmetry profiles of matched anchor P and artifact Q_A, with statistical controls and quality gating supporting reliable anchor-relative verdicts.

Existing approaches, however, address a different question. Equivariant architectures and generative processes enforce prescribed symmetries by construction (Cohen & Welling, 2016; Köhler et al., 2020; Xu et al., 2022; Hoogeboom et al., 2022; Zhang et al., 2025), while statistical symmetry tests ask whether a single observed distribution is invariant (Chiu & Bloem-Reddy, 2023; Soleymani et al., 2025). Neither directly determines whether a deployed generator has changed the symmetry structure of the data it is intended to model. This distinction matters because real data themselves need not be symmetric: directional statistics may arise from scene layout, illumination, acquisition geometry, geographic sampling, or canonical orientation (Lawrence et al., 2026; Lippmann et al., 2026). Consequently, asymmetry in generated samples may simply be inherited from the reference data.

This leads to two distinct questions: is the generated distribution symmetric? and does it preserve the symmetry profile of its reference distribution? For generative fidelity, we argue that the second is the relevant one. The distinction is especially important under modern deployment pipelines, where few-step sampling, consistency models, distillation, guidance, and conditioning can alter the realized output distribution (Song et al., 2023; Yin et al., 2024b). The appropriate object of study is therefore the complete deployed artifact rather than the underlying checkpoint alone. Moreover, group-averaged measures can hide localized changes, such as preserving rotational structure while selectively amplifying one reflection (Luo et al., 2023; Kim et al., 2024; Yin et al., 2024b).

Motivated by these observations, we introduce GENSYMAUDIT, a sample-only framework for artifact-level, anchor-relative symmetry auditing. Given a finite transformation group, GenSymAudit measures the element-wise symmetry profiles of matched real and generated distributions and compares them relative to a real-data anchor rather than to perfect symmetry. We use grouped classifier two-sample tests as transformation-specific separability measurements, together with negative controls, sensitivity calibration, paired inference, multiplicity correction, and quality checks. This separates directional structure inherited from the data from structure amplified, attenuated, or introduced by deployment. Figure 1 summarizes the framework.

We evaluate GenSymAudit on two released remote-sensing generators. For DiffusionSat (Khanna et al., 2024), the fMoW(Christie et al., 2018) reference distribution is strongly asymmetric under D_4, and the deployed generator exhibits three resolved amplifications relative to this anchor, dominated

by vertical reflection ($+18.11$ points). Text2Earth (Liu et al., 2025) shows smaller but bidirectional changes relative to Git-10M: horizontal reflection is attenuated (-2.98 points), whereas vertical reflection is amplified ($+3.62$ points). These results show why perfect invariance is an inappropriate universal target and why symmetry preservation must be assessed relative to the reference distribution.

Controlled experiments further show that likelihood-derived symmetry measurements can become numerically unreliable under aggressive few-step integration even when independent sample-space measurements remain stable. Released ImageNet-64 consistency models additionally show that artifacts with the same sampling budget can exhibit different symmetry profiles depending on deployment provenance. Together, these findings support auditing the distribution produced by the artifact actually deployed, rather than inferring its structural behavior from reference data, model lineage, or sampling budget alone.

Our contributions are:

- We formulate artifact-level, anchor-relative symmetry auditing, which asks whether a deployed generator changes the element-wise symmetry profile of a matched real-data reference rather than whether its outputs are perfectly invariant.
- We introduce GenSymAudit, a sample-only auditing framework with statistical controls and calibration for producing interpretable anchor-relative symmetry verdicts.
- We audit released remote-sensing generators and controlled deployment settings, showing that substantial asymmetry can be inherited from real data while generator-associated changes may be localized to individual transformations.

2 Related Work

Generative-model evaluation and auditing. Generative-model evaluation has increasingly moved beyond aggregate perceptual or feature-space metrics toward sample-based and property-specific diagnostics(Heusel et al., 2017; Kynkäänniemi et al., 2019). Common metrics can depend strongly on the chosen representation and may miss particular distributional errors or human judgments (Stein et al., 2023; Jiralerspong et al., 2023; Ospanov et al., 2024; Veeramacheneni et al., 2025), motivating more targeted evaluations (Asim et al., 2025; Salvy et al., 2026). Closest to our motivation, Bohacek et al. (2026) compare real and generated concept distributions to identify semantic concepts that are suppressed or exaggerated by image generators. Modern deployment further complicates evaluation because sampling budget, guidance, conditioning, and distillation can change the realized output distribution (Liu et al., 2023; Luo et al., 2023; Kim et al., 2024; Yin et al., 2024b;a; Lee et al., 2025). GenSymAudit likewise evaluates the deployed artifact relative to reference data, but focuses on whether a transformation-defined structural profile is preserved rather than on overall perceptual or semantic fidelity.

Testing and diagnosing distributional symmetry. A complementary literature asks whether an observed distribution is invariant under specified transformations. Non-parametric, robust, and conditional procedures provide statistical tests of distributional invariance (Chiu & Bloem-Reddy, 2023; Soleymani et al., 2025), while classifier two-sample tests offer a general sample-based instrument for detecting distributional differences (Gretton et al., 2012; Lopez-Paz & Oquab, 2017; Lippmann et al., 2026). (Lawrence et al., 2026; Lippmann et al., 2026) use related ideas to diagnose symmetry breaking and study its consequences for augmentation and equivariant learning. These methods primarily answer an absolute question: whether a single distribution satisfies a symmetry. GenSymAudit instead compares the element-wise symmetry profiles of a reference distribution P and a deployed distribution $Q_{\mathcal{A}}$, treating asymmetry already present in P as an anchor rather than as generator error.

Equivariant and symmetry-constrained generative models. Symmetry has also been imposed as a modeling prior. Group-equivariant neural networks constrain representations to transform predictably under prescribed actions (Cohen & Welling, 2016), and equivariant diffusion, flow, molecular, crystal, and material-generation methods extend these ideas to generative processes (Köhler et al., 2020; Xu et al., 2022; Hoogeboom et al., 2022; Lin et al., 2024). These approaches address

how symmetry should be enforced during model construction (Cornet et al., 2024; Zhang et al., 2025; Kreit et al., 2026). GenSymAudit is complementary: it assumes no architectural symmetry guarantee and audits the realized output distribution post hoc.

Existing evaluation methods do not explicitly test preservation of a reference-dependent structural profile; symmetry tests ask whether a single distribution is invariant; and equivariant models enforce symmetry by construction. None directly asks whether deployment changes the symmetry structure already present in the data. GenSymAudit formulates this missing problem as artifact-level, anchor-relative auditing, with released remote-sensing generators such as DiffusionSat and Text2Earth providing real-world case studies (Khanna et al., 2024; Liu et al., 2025; Rombach et al., 2022).

3 METHOD

Problem formulation and deployed artifact. We consider the post-hoc problem of auditing whether a deployed generative model preserves the structural symmetry of the distribution it is intended to model. Let P denote a real-data reference distribution over $\mathcal{X}$ and let $Q_{\mathcal{A}}$ denote the distribution produced by a deployed generative artifact $\mathcal{A}$. Importantly, $\mathcal{A}$ refers to the complete deployment configuration rather than the network weights alone: it includes the model parameters, sampler, sampling budget, guidance, and conditioning protocol. Two artifacts derived from the same model family can therefore induce different output distributions even when they share an architecture or sampling budget. Our audit is consequently defined on $Q_{\mathcal{A}}$ itself.

Let G be a finite group acting on $\mathcal{X}$, with identity element e, and let $g_{\#}Q$ denote the push-forward of Q under $g \in G$. A distribution is G-invariant when $Q = g_{\#}Q$ for every $g \in G$. In our remote-sensing experiments, $G = D_4$, whose rotations and reflections act exactly on a square pixel grid. The formulation is not specific to D_4 and applies to other finite groups whenever the corresponding actions are well defined.

Anchor-relative symmetry profile. Rather than reducing symmetry to one group-averaged scalar, we characterize each group element separately. Given a distribution discrepancy $d(\cdot, \cdot)$, we define the symmetry profile of Q as

$$\mathbf{S}_d(Q; G) = \left[d(Q, g_{\#}Q) \right]_{g \in G \setminus \{e\}} . \tag{1}$$

Each component measures the distributional discrepancy induced by a particular group action and therefore retains the directional structure of symmetry breaking. For a discrepancy that separates distributions, $\mathbf{S}_d(Q; G) = \mathbf{0}$ exactly when Q is G-invariant.

The reference distribution itself, however, need not be invariant. In remote sensing, for example, acquisition geometry, illumination, geographic layout, and canonical scene orientations can all induce directional statistics. Judging a generator against zero asymmetry would therefore confound structure already present in the data with structure introduced by generation. We instead use the real-data distribution as an anchor and define, for each g,

$$\Delta_{d,g}(Q_{\mathcal{A}} \mid P) = d(Q_{\mathcal{A}}, g_{\#}Q_{\mathcal{A}}) - d(P, g_{\#}P) . \tag{2}$$

Thus, the object of interest is not whether the generated distribution is absolutely symmetric, but whether deployment changes the symmetry profile of its reference distribution. Positive shifts indicate stronger asymmetry relative to the anchor, while negative shifts indicate attenuation. We refer to statistically resolved positive and negative shifts as *amplified* and *attenuated*, respectively. A shift that lies within a calibrated equivalence region is *anchor-consistent*; if the anchor is consistent with the symmetry null but the artifact is not, the defect is *invented*. Cases that cannot be resolved at the available statistical power are reported as *inconclusive*. These labels describe distributional differences and do not imply a causal mechanism.

Sample-only estimation. GenSymAudit operationalizes Eq. equation 2 using only samples from P and $Q_{\mathcal{A}}$. Our primary estimator is a classifier two-sample test (C2ST) (Karras et al., 2022; Friedman, 2003; Lopez-Paz & Oquab, 2017). For each $g \in G \setminus \{e\}$, we form two disjoint sample pools from the same distribution, transform only one pool by g, and train a binary classifier to distinguish

$$x \sim Q \qquad \text{from} \qquad x^g \sim g_{\#}Q. \tag{3}$$

The pools are disjoint before transformation so that the classifier cannot solve the task through sample correspondence. Under a Bayes-optimal classifier, classification accuracy satisfies

$$a_g^*(Q) = \frac{1}{2} + \frac{1}{2}\,\mathrm{TV}(Q, g_{\#}Q)\,, \tag{4}$$

but a finite learned classifier need not attain this optimum. We therefore treat held-out C2ST accuracy as an empirical separability statistic rather than as a direct numerical estimate of total variation.

The same measurement protocol is applied independently to the real-data anchor and the deployed artifact, yielding $\widehat{S}_g(P)$ and $\widehat{S}_g(Q_{\mathcal{A}})$. Their difference,

$$\widehat{\Delta}_g = \widehat{S}_g(Q_{\mathcal{A}}) - \widehat{S}_g(P), \tag{5}$$

is the empirical counterpart of Eq. equation 2. Uncertainty is estimated by resampling the independent grouping units of the dataset while preserving the original conditioning and stratification structure. The resulting confidence interval is compared with a pre-calibrated equivalence region(Schuirmann, 1987) to determine the element-wise verdict. When several group elements are audited simultaneously, multiplicity is controlled across the corresponding tests(Holm, 1979). We additionally use a non-classifier two-sample discrepancy, such as Energy Distance(Székely & Rizzo, 2013), as an independent check on the direction of the principal sample-space findings rather than as a separate criterion for selecting positive results.

Audit validity. A separability score is interpreted only if the measurement pipeline itself passes a common set of controls. First, a negative control applies the complete C2ST procedure to two independent i.i.d. subsets from the same pool; any systematic deviation from chance indicates leakage or an implementation artifact. Repeating this control across observation seeds provides an empirical null for the corresponding audit arm. Second, controlled asymmetries of increasing strength are injected into otherwise matched samples to determine the smallest perturbation that the complete pipeline can reliably detect. This power calibration determines the practical resolution of the audit and prevents a non-significant anchor–artifact difference from being interpreted as evidence of equality.

We also preserve the dependence structure of the data. Observations belonging to the same scene, geographic instance, or conditioning unit remain in the same split, and stratification is used when conditioning variables could otherwise provide a shortcut. Anchor and generated arms use matched pool sizes because finite-sample C2ST accuracy depends on the amount of training data available to the discriminator. Finally, symmetry conclusions are reported only in sampling regimes where output quality remains comparable to the reference regime. To avoid introducing a preferred orientation through this criterion itself, a base quality statistic ψ can be symmetrized as

$$\psi_G(x) = \frac{1}{|G|}\sum_{g\in G}\psi(g{\cdot}x). \tag{6}$$

Sampling configurations outside the anchor-calibrated quality band are treated as degraded outputs rather than as evidence of increased or decreased symmetry. Exact confidence construction, equivalence margins, multiplicity correction, and control protocols are provided in the appendix.

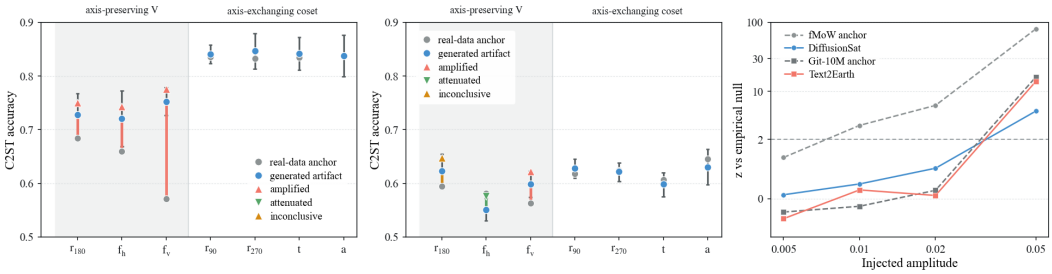
Optional white-box diagnostics. The sample-only procedure above defines the core of GenSymAudit. When additional model access is available, we retain two white-box quantities only as auxiliary diagnostics. For models exposing a tractable density q, prior work defines the group-averaged density

$$\bar{q}(x) = \frac{1}{|G|}\sum_{g\in G} q(g{\cdot}x) \tag{7}$$

and the symmetry defect

$$D_{\mathrm{sym}}(q) = D_{\mathrm{KL}}(q \,\|\, \bar{q}), \tag{8}$$

which is non-negative and vanishes for a G-invariant distribution. We do not introduce this certificate. Our use of it is conditional on whether it can be evaluated reliably for the deployed sampler. In probability-flow models, coarse numerical integration may cause the reverse likelihood computation



(a) DiffusionSat vs fMoW anchor (b) Text2Earth vs Git-10M anchor (c) Power calibration

Figure 2: **Anchor-relative audit results and power calibration.** Paired C2ST profiles compare DiffusionSat with the fMoW anchor and Text2Earth with the Git-10M anchor across D_4, while injected asymmetries calibrate detection sensitivity against the empirical null.

to cease being a faithful inverse of the forward sampling trajectory (Song et al., 2021). We therefore admit D_{sym} only when a forward–reverse round-trip diagnostic passes at the audited sampling budget; otherwise the verdict is determined exclusively from sample space.

When a model exposes a score field but no tractable likelihood, we instead use the normalized score-equivariance residual

$$R_G = \frac{\mathbb{E}_{x,g}\left[\|s_\theta(g{\cdot}x) - g{\cdot}s_\theta(x)\|_2^2\right]}{\mathbb{E}_x\left[\|s_\theta(x)\|_2^2\right]}. \tag{9}$$

R_G measures equivariance of the learned score on the evaluated support, but it is not a distributional divergence and is therefore not compared numerically with D_{sym} or with the sample-only symmetry profile.

Overall, GenSymAudit freezes a generator in its actual deployment configuration, constructs matched real and generated sample pools, estimates their element-wise symmetry profiles under the same measurement protocol, and interprets only the anchor-relative differences that survive the statistical and quality controls. The resulting audit requires only samples; likelihood- and score-based quantities serve as optional diagnostics rather than as prerequisites for the final verdict.

4 Experiments

4.1 Experimental Setup

Released generators and real-data anchors. Our primary experiments audit two publicly released remote-sensing generators. We evaluate DiffusionSat (Khanna et al., 2024) against a matched fMoW anchor and Text2Earth (Liu et al., 2025) against Git-10M. Both audits use matched pools of $n = 28{,}870$ real and generated images at 256×256 resolution and evaluate the seven non-identity elements of D_4. For fMoW, all acquisitions and generated draws associated with the same geographic instance remain grouped throughout splitting and evaluation, with additional stratification over the 62 scene categories. Git-10M contains independent tiles and is stratified by the six-level zoom prefix used in the released Text2Earth conditioning.

Audit protocol. We use the grouped C2ST statistic defined in Sec. 3 as the primary sample-only measurement. Anchor and artifact arms use identical pool sizes and the same classifier architecture, preprocessing, grouped splitting, and cross-validation protocol. All main remote-sensing results below use the final nine-seed evaluation. Uncertainty combines the between-seed measurement component with an instance-level paired bootstrap, and simultaneous testing over the seven non-identity elements is controlled using the Holm procedure. Anchor-consistent verdicts are established through equivalence testing with margin $\delta = 0.05$, rather than through failure to reject a difference. Negative controls, empirical nulls, injected-asymmetry calibration, and a sample-quality gate must pass before a symmetry reading is interpreted. The initial three-seed screen and full statistical protocol are retained only as protocol documentation in Appendix A.

Table 1: **Complete anchor-relative symmetry audit of DiffusionSat and Text2Earth.** Final nine-seed verdicts are reported for all seven non-identity D_4 elements.

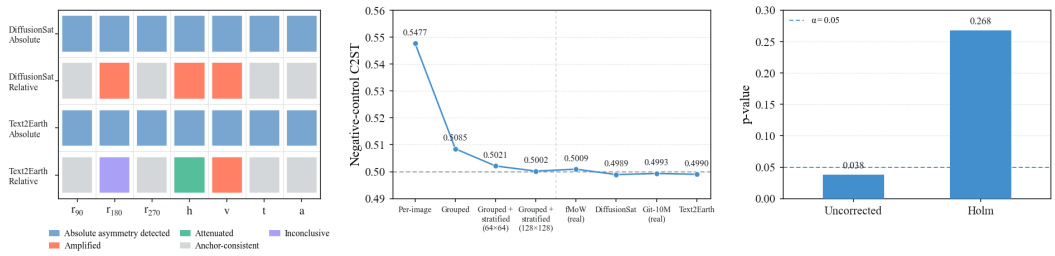
D_4 element	Anchor Acc	Artifact Acc	Change	Verdict
DiffusionSat / fMoW				
Rotation 90°	83.64	84.08	+0.43	Anchor-consistent
Rotation 180°	68.36	**72.75**	**+4.39**	**Amplified**
Rotation 270°	83.30	84.69	+1.39	Anchor-consistent
Horizontal flip	65.90	**71.99**	**+6.10**	**Amplified**
Vertical flip	57.10	**75.20**	**+18.11**	**Amplified**
Transpose	83.53	84.15	+0.62	Anchor-consistent
Anti-transpose	83.85	83.76	−0.08	Anchor-consistent
Text2Earth / Git-10M				
Rotation 90°	61.81	62.73	+0.92	Anchor-consistent
Rotation 180°	59.43	62.21	+2.79	*Inconclusive*
Rotation 270°	62.42	62.07	−0.35	Anchor-consistent
Horizontal flip	58.02	**55.04**	**−2.98**	**Attenuated**
Vertical flip	56.21	**59.83**	**+3.62**	**Amplified**
Transpose	60.41	59.76	−0.65	Anchor-consistent
Anti-transpose	64.47	62.97	−1.51	Anchor-consistent

Validation and downstream evaluation. We additionally use permutation-symmetric DW-N systems to test the reliability of likelihood-derived measurements under controlled few-step sampling. Practical relevance is evaluated by training identical ResNet-18 classifiers on matched real or generated remote-sensing imagery and measuring their accuracy and orientation dependence on held-out real data. Additional deployment and sensitivity analyses, including ImageNet-64 consistency models, production NFE sweeps, classifier-capacity checks, acquisition-metadata analysis, and MatterGen(Zeni et al., 2025), are reported in the appendix.

4.2 Anchor-Relative Symmetry Audits of Released Generators

We first examine whether deployed generators preserve the element-wise symmetry structure of their reference distributions. The relevant target is not perfect D_4 invariance, since the real-data distributions themselves contain substantial directional structure. Figure 2 shows the complete seven-element C2ST profiles for the two generator–anchor pairs. For fMoW, anchor C2ST ranges from 57.10 under vertical reflection to 83.85 under anti-diagonal reflection, while Git-10M ranges from 56.21 to 64.47, with 50 corresponding to chance. Both are therefore substantially non-invariant relative to the empirical null. This observation is central to the audit: judging generated samples only against perfect symmetry would attribute structure already present in the reference data to the generator itself. GenSymAudit instead evaluates the artifact-relative change, allowing inherited directional structure to be separated from deployment-associated changes.

Relative to the fMoW anchor, DiffusionSat exhibits three statistically resolved amplifications. The strongest occurs under vertical reflection, where C2ST accuracy increases from 57.10 to 75.20, corresponding to a change of +18.11 points, followed by horizontal reflection (65.90 → 71.99, +6.10 points) and 180° rotation (68.36 → 72.75, +4.39 points); the remaining four elements are anchor-consistent under the calibrated equivalence criterion. Text2Earth shows weaker but bidirectional changes relative to Git-10M: horizontal reflection decreases from 58.02 to 55.04 (−2.98 points), vertical reflection increases from 56.21 to 59.83 (+3.62 points), and 180° rotation changes from 59.43 to 62.21 but remains inconclusive, while the remaining four elements are anchor-consistent. Overall, eight of the fourteen artifact–transformation pairs are anchor-consistent, five exhibit resolved amplification or attenuation, and one remains inconclusive. Table 1 reports the complete final verdicts for all fourteen pairs; the corresponding statistical intervals, adjusted significance tests, and equivalence details are provided in the appendix.



(a) Real-data anchor (b) Split leakage control and final null validation (c) Multiplicity correction

Figure 3: **Ablation and validation of the GENSYMAUDIT safeguards.** Real-data anchoring prevents inherited asymmetry from being attributed to the artifact, stricter split controls return negative-control C2ST to chance, and the three-seed screen demonstrates the effect of Holm correction.

The resolved changes are notably localized rather than distributed uniformly across D_4. All five statistically resolved shifts lie within the axis-preserving subgroup $V = \{e, r_{180}, \mathrm{flip_h}, \mathrm{flip_v}\}$, whereas the axis-exchanging transformations remain anchor-consistent. Directional acquisition cues, illumination geometry, and cardinal scene layout are plausible contributors to this pattern, but the available evidence does not support a unique causal explanation. In fMoW, approximately 86% of acquisitions share a southeast sun-azimuth quadrant; however, stratifying the anchor into matched long- and short-shadow subsets leaves vertical-reflection C2ST at 53.0 ± 0.9 and 51.8 ± 0.6, respectively, indicating that shadow strength alone cannot explain the observed asymmetry. The high baseline separability of the axis-exchanging transformations may also reduce discriminator headroom for detecting additional artifact-induced shifts. Moreover, the asymmetry persists across quality-valid sampling budgets in our NFE analysis, making a purely low-step numerical explanation insufficient. We therefore interpret the subgroup concentration as a reproducible empirical pattern while avoiding a stronger mechanistic claim.

4.3 ABLATION OF AUDIT SAFEGUARDS

We next examine three safeguards required for reliable artifact-level conclusions. Figure 3 summarizes the effects of real-data anchoring, split-leakage control, and multiplicity correction. Power calibration is reported separately in Figure 2 and determines the empirical resolution used to set the equivalence margin $\delta = 0.05$.

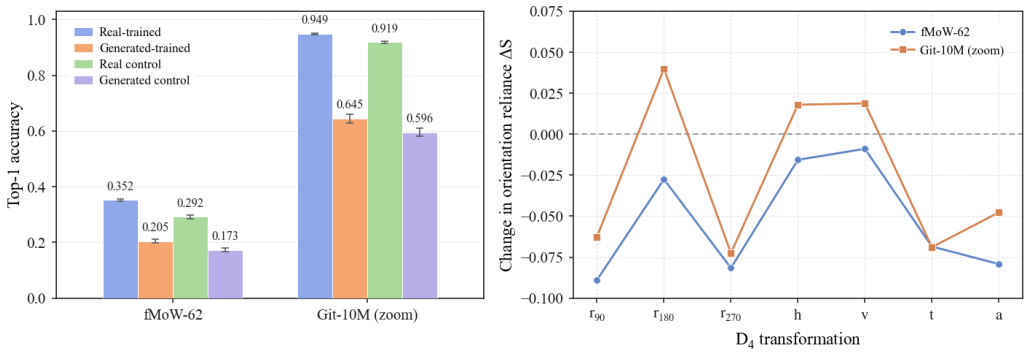
Without a real-data anchor, inherited directional structure is conflated with artifact-associated change: both DiffusionSat and Text2Earth would be reported as asymmetric under all seven non-identity D_4 elements, rather than exhibiting the localized three and two shifts identified by the full audit. Split construction is similarly consequential. A naive per-image split yields a negative-control C2ST of 0.5477, whereas grouping repeated observations reduces it to 0.5085 and additional stratification returns it to 0.5002 at the primary resolution. The negative control therefore exposes leakage that would otherwise contaminate the empirical null.

Multiplicity correction also changes the statistical interpretation. In the initial three-seed screen, the Text2Earth horizontal-reflection comparison gives a nominal $p = 0.038$, but its Holm-adjusted value increases to 0.268 when the seven transformations are treated as one family. This screening example isolates the effect of multiplicity correction and does not replace the final nine-seed verdict. Separately, injected-asymmetry calibration places the empirical detection floor at approximately 0.003–0.042, motivating the conservative equivalence margin $\delta = 0.05$ rather than treating a non-significant difference as evidence of preservation.

Together, real-data anchoring, leakage control, multiplicity correction, and calibrated equivalence testing address distinct sources of audit error and prevent unsupported symmetry verdicts.

4.4 DOWNSTREAM UTILITY AND DIRECTIONAL ASSOCIATION

We finally examine whether the directional structure identified by GenSymAudit is reflected in models trained on generated data. For each corpus, the same randomly initialized ResNet-18 is trained



(a) Downstream task utility

(b) Directional structure across D_4

Figure 4: **Downstream utility and directional structure.** Top-1 accuracy compares classifiers trained on real and generated imagery, including reduced-size controls. Element-wise changes in orientation reliance ΔS summarize the axis-preserving versus axis-exchanging pattern across D_4 on fMoW-62 and Git-10M.

on matched real or generated imagery and evaluated on the same held-out real test set. We consider fMoW-62 scene classification and Git-10M zoom-level classification, using five training seeds for every arm. Reduced-size controls based on disjoint half-splits are included to assess training-set-size and sample-replacement effects. Rotation and reflection augmentation are disabled because they would suppress the orientation dependence being measured.

Training on generated imagery produces a substantial utility gap. Top-1 accuracy decreases from 0.3522 ± 0.0039 to 0.2049 ± 0.0061 on fMoW-62 and from 0.9491 ± 0.0021 to 0.6453 ± 0.0160 on Git-10M, corresponding to relative reductions of 41.8% and 32.0%, respectively (Figure 4(a)); the same ordering persists in the reduced-size controls. These gaps should not be attributed solely to symmetry because real and generated imagery may differ along many other dimensions. At the individual-element level, two-sided exact Spearman tests show an association between the audit shift and change in orientation reliance on fMoW ($\rho = 0.82$, unadjusted $p = 0.034$), but not on Git-10M ($\rho = 0.36$, $p = 0.44$). This dataset-dependent result is exploratory and does not establish causality.

A clearer pattern appears at subgroup resolution. The three non-identity axis-preserving transformations—180° rotation, horizontal reflection, and vertical reflection—occupy the top three ranks of the downstream change in both corpora, while the four axis-exchanging transformations occupy the bottom four (Figure 4(b)). This separation gives permutation $p = 0.029$ for each corpus and Fisher-combined $p = 0.007$, although the post-hoc subgroup analysis remains exploratory. The direction of change is also task-dependent: the fMoW generated-trained classifier becomes less orientation-reliant overall, whereas the clearest increases on Git-10M occur within the axis-preserving subgroup. Thus, GenSymAudit identifies a structural partition reflected across both tasks, but the experiment establishes neither an element-wise causal relationship nor that symmetry explains the overall synthetic-data utility gap. Complete measurements and seed-level statistics are reported in the appendix.

5 Conclusion

We introduced GenSymAudit, a sample-only framework for auditing the distributional symmetry of deployed generative artifacts. Our results show that directional asymmetry may already exist in reference data, making real-data anchors essential for separating inherited structure from artifact-associated shifts. Relative to their anchors, DiffusionSat exhibits three resolved amplifications dominated by vertical reflection, whereas Text2Earth shows weaker horizontal attenuation and vertical amplification. Controlled studies further reveal that likelihood-based measurements can fail in few-step regimes and that similarly configured consistency models can exhibit distinct symmetry profiles. Symmetry should therefore be audited on the deployed artifact itself rather than inferred from its data, ancestry, or sampling configuration.

REPRODUCIBILITY STATEMENT

To support reproducibility, the supplementary material includes the implementation, evaluation scripts, and experimental configurations, as well as the experimental results for the models and datasets used in this paper. Please refer to the appendix for detailed experimental settings and procedures.

AI USE STATEMENT

Large language models (LLMs) were used during parts of the implementation and manuscript preparation, including language polishing, paragraph shortening, and transition refinement. All outputs were reviewed and verified by the authors.

REFERENCES

Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele, and Jan Eric Lenssen. Met3r: Measuring multi-view consistency in generated images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6034–6044, 2025.

Maty Bohacek, Thomas Fel, Maneesh Agrawala, and Ekdeep Singh Lubana. Uncovering conceptual blindspots in generative image models using sparse autoencoders. In *The Fourteenth International Conference on Learning Representations*, 2026.

Kenny Chiu and Benjamin Bloem-Reddy. Non-parametric hypothesis tests for distributional group symmetry. *arXiv preprint arXiv:2307.15834*, 2023.

Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6172–6180, 2018.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2990–2999. PMLR, 2016.

François Cornet, Grigory Bartosh, Mikkel N. Schmidt, and Christian A. Naesseth. Equivariant neural diffusion for molecule generation. *Advances in Neural Information Processing Systems*, 37:49429–49460, 2024.

Jerome H. Friedman. On multivariate goodness-of-fit and two-sample testing. In *Proceedings of PHYSTAT 2003: Statistical Problems in Particle Physics, Astrophysics and Cosmology*, pp. 311–313, 2003.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The journal of machine learning research*, 13:723–773, 2012.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Emiel Hoogeboom, Víctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 8867–8887. PMLR, 2022.

Marco Jiralerspong, Avishek Joey Bose, Ian Gemp, Chongli Qin, Yoram Bachrach, and Gauthier Gidel. Feature likelihood divergence: Evaluating the generalization of generative models using samples. *Advances in Neural Information Processing Systems*, 36:33095–33119, 2023.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 26565–26577, 2022.

Samar Khanna, Patrick Liu, Linqi Zhou, Chenlin Meng, Robin Rombach, Marshall Burke, David Lobell, and Stefano Ermon. Diffusionsat: A generative foundation model for satellite imagery. In *The Twelfth International Conference on Learning Representations*, 2024.

Dongjun Kim, Chieh-Hsin Lai, Wei-Hsiang Liao, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yutong He, Yuki Mitsufuji, and Stefano Ermon. Consistency trajectory models: Learning probability flow ode trajectory of diffusion. In *The Twelfth International Conference on Learning Representations*, 2024.

Jonas Köhler, Leon Klein, and Frank Noé. Equivariant flows: Exact likelihood generative learning for symmetric densities. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5361–5370. PMLR, 2020.

Janik Kreit, Dominic Schuh, Kim A. Nicoli, and Lena Funcke. Sesamo: Symmetry-enforcing stochastic modulation for normalizing flows. In *The Fourteenth International Conference on Learning Representations*, 2026.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems*, volume 32, pp. 3927–3936, 2019.

Hannah Lawrence, Elyssa Hofgard, Vasco Portilheiro, Yuxuan Chen, Tess Smidt, and Robin Walters. To augment or not to augment? diagnosing distributional symmetry breaking. In *The Fourteenth International Conference on Learning Representations*, 2026.

Sangyun Lee, Yilun Xu, Tomas Geffner, Giulia Fanti, Karsten Kreis, Arash Vahdat, and Weili Nie. Truncated consistency models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Peijia Lin, Pin Chen, Rui Jiao, Qing Mo, Jianhuan Cen, Wenbing Huang, Yang Liu, Dan Huang, and Yutong Lu. Equivariant diffusion for crystal structure prediction. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 29890–29913. PMLR, 2024.

Peter Lippmann, Roman Remme, Manuel Viktor Klockow, and Fred A. Hamprecht. Take note: Your molecular dataset is probably aligned. In *The Fourteenth International Conference on Learning Representations*, 2026.

Chenyang Liu, Keyan Chen, Rui Zhao, Zhengxia Zou, and Zhenwei Shi. Text2earth: Unlocking text-driven remote sensing image generation with a global-scale dataset and a foundation model. *IEEE Geoscience and Remote Sensing Magazine*, 13(3):238–259, 2025.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. In *International Conference on Learning Representations*, 2017.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.

Joshua Mitton and Roderick Murray-Smith. Rotation equivariant deforestation segmentation and driver classification. *arXiv preprint arXiv:2110.13097*, 2021.

Azim Ospanov, Jingwei Zhang, Mohammad Jalali, Xuenan Cao, Andrej Bogdanov, and Farzan Farnia. Towards a scalable reference-free evaluation of generative models. *Advances in Neural Information Processing Systems*, 37:120892–120927, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022. doi: 10.1109/CVPR52688.2022.01042.

Nicolas Salvy, Hugues Talbot, and Bertrand Thirion. Enhanced generative model evaluation with clipped density and coverage. In *The Fourteenth International Conference on Learning Representations*, 2026.

Donald J. Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6):657–680, 1987. doi: 10.1007/BF01068419.

Ashkan Soleymani, Behrooz Tahmasebi, Stefanie Jegelka, and Patrick Jaillet. A robust kernel statistical test of invariance: Detecting subtle asymmetries. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 4816–4824. PMLR, 2025.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023.

George Stein, Jesse Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L. Caterini, Eric Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. *Advances in Neural Information Processing Systems*, 36:3732–3784, 2023.

Gábor J. Székely and Maria L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013. doi: 10.1016/j.jspi.2013.03.018.

Lokesh Veeramacheneni, Moritz Wolter, Hilde Kuehne, and Juergen Gall. Fréchet wavelet distance: A domain-agnostic metric for image generation. In *The Thirteenth International Conference on Learning Representations*, 2025.

Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. Geodiff: A geometric diffusion model for molecular conformation generation. In *International Conference on Learning Representations*, 2022.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Frédo Durand, and William T. Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in Neural Information Processing Systems*, 37:47455–47487, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6613–6623, 2024b.

Claudio Zeni et al. A generative model for inorganic materials design. *Nature*, 639:624–632, 2025.

Leo Zhang, Kianoosh Ashouritaklimi, Yee Whye Teh, and Rob Cornish. Symdiff: Equivariant diffusion via stochastic symmetrisation. In *The Thirteenth International Conference on Learning Representations*, 2025.

A Remote-Sensing Audit Protocol and Complete Results

This section records the complete protocol underlying the remote-sensing audits in Sec. 4.2. We specify the released artifacts, conditioning and pool construction, grouped C2ST protocol, empirical controls, and the statistical decision rule used for the final nine-seed verdicts. We also document the conditioning scope of the DiffusionSat experiment explicitly so that the reported anchor-relative results can be interpreted under the protocol that was actually executed.

A.1 Artifacts, Conditioning, and Matched Pools

For DiffusionSat, we audit the public single-image checkpoint `checkpoint-150000` at 256×256 resolution through the released `DiffusionSatPipeline` in fp16, using the repository-default scheduler and classifier-free guidance scale 7.5. The conditioning list contains one caption–metadata record for each fMoW validation instance, constructed from the first acquisition of that instance. Captions follow the released template "a fmow satellite image of a {category} in {country}", and the seven numerical metadata fields are longitude, latitude, ground-sample distance, cloud cover, year, month, and day. We follow the normalization path used by the released training implementation, including the longitude, latitude, and year offsets expected by the metadata normalizer. The conditioning list, its order, and all per-index prior-noise seeds are frozen once and reused throughout the audit.

The fMoW anchor contains up to four acquisitions for each validation instance, yielding $n = 28{,}870$ real images. Each DiffusionSat arm contains the same number of generated samples per instance as the number of real acquisitions retained for that instance, so the two arms are matched in pool size, instance multiplicity, category composition, and the conditioning list used by the generator. The executed experiment is therefore *instance-matched but not acquisition-metadata-matched*: multiple generated draws for one instance share the conditioning record of its first acquisition, whereas the real anchor contains the corresponding later acquisitions as well. We retain this protocol because it is the one used for all reported DiffusionSat results and make the distinction explicit rather than treating the two marginal pools as exactly acquisition-matched. All acquisitions and generated draws associated with the same geographic instance remain grouped throughout splitting and inference.

Real fMoW images are center-cropped to a square and resized to 256×256 using bilinear interpolation before entering the audit. All subsequent D_4 actions are exact pixel permutations, so the transformations themselves introduce no interpolation artifacts.

For Text2Earth, we audit the released 1.3B-parameter latent-diffusion checkpoint using CLIP text conditioning, the shipped PNDM sampler at its default 50 steps, guidance scale 7.5, and 256×256 resolution. The Git-10M anchor contains 28,870 image–text pairs sampled from eleven released shards. The released text field, including its six-level zoom prefix, is passed verbatim to the generator, and one fixed prior-noise seed is assigned to each conditioning pair. Unlike fMoW, Git-10M contains independent tiles rather than repeated acquisitions of explicit geographic instances; groups are therefore defined per image and all splits are stratified by the six-level zoom prefix. Spatial overlap between distinct tiles cannot be ruled out from the released metadata; all inference is therefore conditional on the per-image grouping available in the release.

A.2 C2ST Instrument, Splitting, and Negative Controls

For each non-identity $g \in D_4$, we estimate the separability between a distribution and its transformed version with a classifier two-sample test. The primary discriminator is a small convolutional network with approximately 2M parameters, trained on 128×128 inputs using five-fold cross-validation and three epochs per fold. Every image is assigned either to the original pool or to the transformed pool, never to both, and the same classifier architecture and training procedure are used for the anchor and artifact arms. The resulting accuracy is a classifier-relative empirical separability statistic; it is not identified with the exact total-variation distance or with an architecture-independent property of the two distributions.

For fMoW, all acquisitions of the same geographic instance and all generated draws associated with that instance remain in the same split and cross-validation fold. Splits are additionally stratified over the 62 scene categories. These choices were determined empirically from the negative control. A naive per-image split yields a real-data negative-control C2ST of 0.5477, indicating that the classifier can exploit geographic-instance identity. Grouping acquisitions by instance reduces the value to 0.5085, and category stratification further reduces it to 0.5021 ± 0.0007 at 64×64 and 0.5002 at the primary 128×128 resolution.

Twelve independent negative-control re-splits give empirical nulls of 0.5009 ± 0.0027 for the fMoW anchor and 0.4989 ± 0.0033 for DiffusionSat. For the Text2Earth audit, the corresponding empirical nulls are 0.4993 ± 0.0014 for Git-10M and 0.4990 ± 0.0031 for Text2Earth, with negative controls of 0.5006 and 0.5000.

Table 2: **Negative-control validation for the fMoW audit.**

Split protocol	Negative-control C2ST
Per-image split	0.5477
Instance-grouped folds	0.5085
Grouped + stratified, 64×64	0.5021 ± 0.0007
Grouped + stratified, 128×128	$\mathbf{0.5002}$

An additional conditioning shortcut was detected during the initial Text2Earth screen: zoom strata were propagated to the Git-10M anchor but not to the generated pool. After correcting the metadata alignment and rerunning the affected cells, the three-seed rotation-90° artifact reading changed from 0.641 to 0.6136 ± 0.0217. All final Text2Earth results use the corrected zoom-stratified protocol.

A.3 Statistical Inference and Verdicts

The final remote-sensing audit uses nine observation seeds. An observation seed changes the grouped split, classifier initialization, and stochastic training trajectory while leaving the realized image pools fixed. Its variation therefore quantifies measurement uncertainty conditional on those pools rather than uncertainty over the underlying data-generating distribution.

Pool composition is handled separately through a paired instance-level bootstrap. For fMoW, the resampling unit is the geographic instance and all acquisitions and associated generated draws are resampled jointly. For Git-10M, the matched image–conditioning record is the resampling unit. The reported intervals combine the paired-pool variance and the variance of the nine-seed measurement mean in quadrature. They are then widened to the family-wise level used for the seven simultaneously audited D_4 elements; for the final nine-seed remote-sensing tables this gives the reported corrected intervals with critical value $z_{\mathrm{crit}} = 2.69$. This construction separates finite-pool variation from measurement variation, but it does not turn the released pools into random samples from an unspecified super-population.

For every element we test the relative null

$$H_{0,g}^{\mathrm{rel}}: \quad a_g(Q_{\mathcal{A}}) = a_g(P), \tag{10}$$

where a_g denotes the C2ST accuracy under transformation g. Difference-test p-values are corrected separately within each seven-element artifact family using Holm's step-down procedure.

Failure to reject the relative null is not treated as evidence of preservation. We additionally perform equivalence testing with margin

$$\delta = 0.05. \tag{11}$$

The margin is tied to the empirical resolution of the audit rather than interpreted as a universal threshold of practical importance. Injected-asymmetry calibration places the observed C2ST resolution at approximately 0.003–0.042 accuracy points across the remote-sensing arms, so $\delta = 0.05$ is deliberately no finer than the coarsest demonstrated resolution.

The final decision is sequential and mutually exclusive. If the Holm-adjusted relative-difference test is significant, the element is labeled *Amplified* when $\Delta_g > 0$ and *Attenuated* when $\Delta_g < 0$. Otherwise, if equivalence within $(-\delta, +\delta)$ is established, the element is labeled *Anchor-consistent*. If neither condition is satisfied, the result is *Inconclusive*. Thus, amplified and attenuated denote statistically resolved signed shifts, whereas anchor consistency is an equivalence conclusion rather than a non-significant difference. A resolved shift may have magnitude below δ: the directional label records statistical resolution and sign, not exceedance of a universal practical-effect threshold.

The equivalence conclusion is sensitive to the resolution margin in the expected direction but the resolved shifts are not. Under the historical sensitivity check, reducing the margin to $\delta = 0.03$ reduces the four DiffusionSat equivalences to two, whereas increasing it to $\delta = 0.075$ also classifies Text2Earth rotation-180° as equivalent. None of the five rejected anchor-relative differences changes with δ.

Table 3: **Complete DiffusionSat/fMoW audit at the shipped configuration.**

D_4 element	fMoW	DiffusionSat	Δ	Corrected interval	Verdict
Rotation 180°	0.6836	0.7275	$\mathbf{+0.0439}$	$[\mathbf{+0.005}, \mathbf{+0.083}]$	**Amplified**
Horizontal flip	0.6590	0.7199	$\mathbf{+0.0610}$	$[\mathbf{+0.009}, \mathbf{+0.113}]$	**Amplified**
Vertical flip	0.5710	0.7520	$\mathbf{+0.1811}$	$[\mathbf{+0.156}, \mathbf{+0.207}]$	**Amplified**
Rotation 90°	0.8364	0.8408	$+0.0043$	$[-0.014, +0.022]$	Anchor-consistent
Rotation 270°	0.8330	0.8469	$+0.0139$	$[-0.020, +0.047]$	Anchor-consistent
Transpose	0.8353	0.8415	$+0.0062$	$[-0.024, +0.037]$	Anchor-consistent
Anti-transpose	0.8385	0.8376	-0.0008	$[-0.040, +0.038]$	Anchor-consistent

Table 4: **Complete Text2Earth/Git-10M audit at the shipped configuration.**

D_4 element	Git-10M	Text2Earth	Δ	Corrected interval	Verdict
Rotation 180°	0.5943	0.6221	$+0.0279$	$[-0.004, +0.060]$	*Inconclusive*
Horizontal flip	0.5802	0.5504	$\mathbf{-0.0298}$	$[\mathbf{-0.050}, \mathbf{-0.009}]$	**Attenuated**
Vertical flip	0.5621	0.5983	$\mathbf{+0.0362}$	$[\mathbf{+0.011}, \mathbf{+0.062}]$	**Amplified**
Rotation 90°	0.6181	0.6273	$+0.0092$	$[-0.009, +0.027]$	Anchor-consistent
Rotation 270°	0.6242	0.6207	-0.0035	$[-0.021, +0.014]$	Anchor-consistent
Transpose	0.6041	0.5976	-0.0065	$[-0.029, +0.016]$	Anchor-consistent
Anti-transpose	0.6447	0.6297	-0.0151	$[-0.048, +0.018]$	Anchor-consistent

A.4 Complete Nine-Seed Audit

Tables 3 and 4 report the complete nine-seed audits. These are the exact numerical results underlying the main-paper profile figure and verdict table.

The Holm-adjusted p-values for the three resolved DiffusionSat shifts are 0.0132 for rotation $180°$, 0.0099 for horizontal reflection, and $< 10^{-4}$ for vertical reflection. The other four cells satisfy the $\delta = 0.05$ equivalence criterion.

The Holm-adjusted p-values are 0.0005 for horizontal reflection, 0.0009 for vertical reflection, and 0.0957 for rotation $180°$. The latter is the only cell in the two audits that is neither significantly different nor equivalent at $\delta = 0.05$. Across the fourteen artifact–transformation pairs, eight are therefore anchor-consistent, five exhibit resolved amplification or attenuation, and one remains inconclusive.

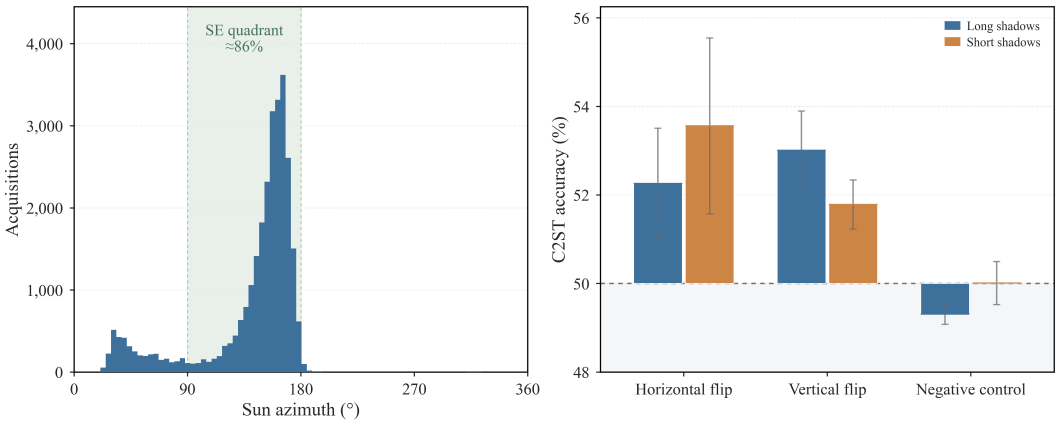
B Audit Safeguards and Sensitivity Analyses

This section provides the detailed evidence behind the safeguard ablations in Sec. 4.3 and examines several factors that affect interpretation of the remote-sensing audit. Because the safeguards address different statistical failure modes, the historical experiments do not reduce to a single common accuracy metric; we therefore report the diagnostic associated with each component rather than combining heterogeneous quantities into an artificial aggregate score.

B.1 Anchor and Split Ablations

Removing the real-data anchor changes the scientific question from an anchor-relative comparison to an absolute test of D_4 invariance. Under the latter, all seven non-identity elements are detectable for both generated distributions. The anchor-relative procedure instead isolates three DiffusionSat shifts and two Text2Earth shifts. The remaining asymmetry is not absent; it is already present in the corresponding real-data distribution and should therefore not be attributed to the generator.

Split construction produces a different failure. As shown in Table 2, a naive fMoW per-image split yields a negative-control C2ST of 0.5477. Grouping repeated acquisitions by instance reduces the shortcut to 0.5085, and category stratification returns the control to chance. Thus, the grouped and stratified protocol is required before a small element-level difference can be interpreted structurally.



(a) Sun-azimuth distribution

(b) Reflection across shadow strata

Figure 5: **Acquisition analysis of the fMoW anchor.** Approximately 86% of acquisitions fall in the southeast sun-azimuth quadrant, while reflection C2ST varies only modestly between matched long- and short-shadow strata and negative controls remain near chance.

B.2 Multiplicity and Equivalence Calibration

Multiplicity correction materially affects the initial screening stage. In the three-seed Text2Earth horizontal-reflection screen, the nominal comparison gives $p = 0.038$, whereas Holm adjustment over the seven audited elements gives $p_{\text{adj}} = 0.268$. This pilot value is not the final horizontal-reflection result; after expansion to nine seeds and paired inference, horizontal reflection is independently resolved as an attenuation with $p_{\text{adj}} = 0.0005$. The example nevertheless shows why nominal element-wise significance cannot be used during screening.

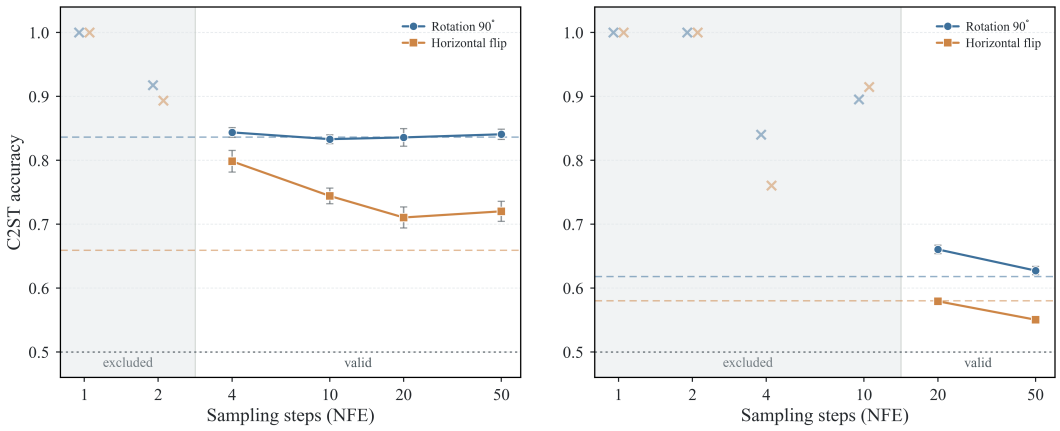
Power calibration is needed for the opposite decision. Without a calibrated equivalence margin, a non-significant difference can only be called unresolved. Injected perturbations place the C2ST resolution between approximately 0.003 and 0.042, motivating the conservative $\delta = 0.05$ equivalence margin. The eight anchor-consistent cells therefore rule out changes larger than δ in the measured C2ST profile under the current classifier, grouping, and pool protocol; they do not establish equality of the underlying distributions.

B.3 Classifier and Resolution Sensitivity

C2ST is classifier-dependent, so its absolute accuracy should not be interpreted as an architecture-independent divergence. We evaluated representative cells with a linear classifier and a wider CNN in addition to the main convolutional discriminator. The dominant DiffusionSat vertical-reflection effect is stable across capacity: the linear classifier reads $0.511 \rightarrow 0.578$, the main nine-seed discriminator $0.571 \rightarrow 0.752$, and the wider CNN $0.575 \rightarrow 0.795$. Negative controls remain between 0.496 and 0.507.

The weaker cells are more sensitive. Under the wider CNN, DiffusionSat rotation 90° changes from 0.853 to 0.809 but does not support a non-anchor-consistent interpretation under the historical sensitivity criterion. On Text2Earth, representative wide-CNN checks give approximately $0.605 \rightarrow 0.613$ for rotation and $0.522 \rightarrow 0.522$ for horizontal reflection. We therefore regard the exact magnitude of small C2ST shifts as instrument-dependent and reserve architecture-robust language for effects directly supported by the sensitivity checks. The final main-paper verdicts are the output of the prespecified primary C2ST instrument rather than claims that every discriminator family produces identical numerical separation.

A 64×64 resolution check similarly attenuates fine-scale separability while broadly preserving the profile. For Text2Earth, rotation-90° reads approximately $0.601/0.604$ for anchor/artifact and horizontal reflection approximately $0.527/0.540$. These results reinforce the interpretation of C2ST as a finite-capacity empirical instrument.



(a) DiffusionSat (fMoW)

(b) Text2Earth (Git-10M)

Figure 6: **Quality-gated NFE sweeps.**Generated-artifact C2ST is shown across sampling budgets for DiffusionSat and Text2Earth, with dashed lines denoting the corresponding real-data anchors. Faded crosses mark excluded configurations, and only quality-valid points are used to support robustness claims.

B.4 Subgroup Localization and Headroom

All five resolved remote-sensing shifts lie inside the axis-preserving Klein subgroup

$$V = \{e, r_{180}, \mathrm{flip}_{\mathrm{h}}, \mathrm{flip}_{\mathrm{v}}\}, \tag{12}$$

whereas the four axis-exchanging transformations are anchor-consistent on both checkpoints. This organization is visually striking but was identified after inspecting the audit results and should therefore be treated as exploratory rather than as a preregistered subgroup hypothesis. Moreover, D_4 contains other order-four subgroups, so a post-hoc probability calculation that treats V as the unique candidate partition would overstate the evidence.

The pattern is also confounded with discriminator headroom. In fMoW, the axis-exchanging anchor readings lie between approximately 0.833 and 0.839, whereas the three non-identity elements inside V lie between 0.571 and 0.684. Git-10M shows the same qualitative imbalance. Because C2ST accuracy is bounded above by one, a fixed underlying distributional shift may produce a smaller change in the observed statistic when the anchor is already highly separable. The subgroup concentration is therefore reported as a replicated empirical regularity that may be partially confounded by instrument headroom, not as evidence for a unique causal mechanism.

B.5 Acquisition-Metadata Analysis

The fMoW metadata provide a partial test of whether acquisition geometry contributes to the anchor's reflection asymmetry. Approximately 86% of the $28{,}870$ anchor acquisitions have sun azimuths in the southeast quadrant, making a preferred illumination direction physically plausible. To examine the role of shadow length, we divide the anchor by sun elevation into matched long- and short-shadow strata containing $n = 9{,}612$ images each.

As shown in Figure 5, vertical-reflection C2ST accuracy is 0.530 ± 0.009 in the long-shadow stratum and 0.518 ± 0.006 in the short-shadow stratum. The between-stratum difference is 0.012 on the accuracy scale, or 1.2 percentage points, and is modest relative to the reported uncertainty. Horizontal reflection shows no larger between-stratum difference, and the negative controls remain near chance. These measurements do not provide clear evidence that shadow length alone accounts for the observed reflection asymmetry.

Illumination direction, acquisition geometry, cardinal scene layout, and scene content may all contribute, and the available metadata do not identify a unique mechanism. The anchor-relative audit compares the measured symmetry profiles of the real and generated pools; it does not causally iso-

Table 5: **Two-component quality gate across the production NFE sweeps.** "Scalar pass; visual fail" denotes a point inside the scalar $\pm 25\%$ band that is excluded from generator-level claims by fixed-index inspection.

Artifact	NFE	Quality / anchor	Interpretation
DiffusionSat	50	95.0%	Valid
DiffusionSat	20	92.4%	Valid
DiffusionSat	10	87.6%	Valid
DiffusionSat	4	79.2%	Valid
DiffusionSat	2	77.5%	Scalar pass; visual fail
DiffusionSat	1	199.8%	Excluded
Text2Earth	50	82.2%	Valid
Text2Earth	20	87.7%	Valid
Text2Earth	10	125.9%	Excluded
Text2Earth	4	124.5%	Excluded
Text2Earth	2	229.8%	Excluded
Text2Earth	1	244.6%	Excluded

late or remove individual acquisition factors. In particular, the DiffusionSat comparison is instance-matched but not acquisition-metadata-matched, as detailed in Appendix A.1.

B.6 Production NFE Sweep and Quality Gate

We test whether reducing the number of function evaluations (NFE) systematically removes the measured directional structure. Such comparisons are meaningful only while the generated outputs remain usable. We therefore apply a two-component quality gate. The scalar component measures the mean absolute horizontal-neighbor difference relative to the corresponding real-data anchor, with an acceptance band of $\pm 25\%$. Configurations passing this criterion are additionally inspected using fixed sample indices to identify structural degradation not captured by the scalar statistic.

As summarized in Table 5, DiffusionSat remains valid from NFE 50 to 4. Although NFE$=2$ lies inside the scalar band, fixed-index inspection reveals directional smearing, so it is excluded by the full gate. Text2Earth remains valid only from NFE 50 to 20, as its scalar statistic leaves the acceptance band at NFE$=10$.

Within the valid DiffusionSat range, rotation-90° retains approximately 98–101% of its 50-step excess over the chance-level null, while horizontal-reflection C2ST increases rather than decreases as NFE is reduced. For Text2Earth, from NFE 50 to 20, rotation-90° changes from approximately 0.627 to 0.660 and horizontal reflection from approximately 0.550 to 0.579, with negative controls remaining near chance. Figure 6 therefore provides no evidence that reducing NFE systematically removes the measured asymmetry before output quality fails.

Figure 7 provides the visual component of the quality gate. DiffusionSat exhibits directional smearing at NFE$=2$ and becomes decoder-textured noise at NFE$=1$, where C2ST reaches 1.000. Text2Earth develops severe color and structural artifacts by NFE$=10$, while its C2ST rises toward 0.9. These extreme values measure anisotropic degradation rather than the symmetry of a usable deployed generator and are therefore excluded from the robustness analysis.

C Controlled Validation of the Likelihood-Based Measurement

This section collects the controlled experiments used to determine when the optional likelihood-derived symmetry certificate can be trusted. The experiments use permutation-symmetric DW-N systems, where model dimension, sampling budget, and defect thickness can be manipulated directly. We distinguish explicitly between the exact finite-group calculation available for S_4 and the capped orbit approximation used historically for larger permutation groups.

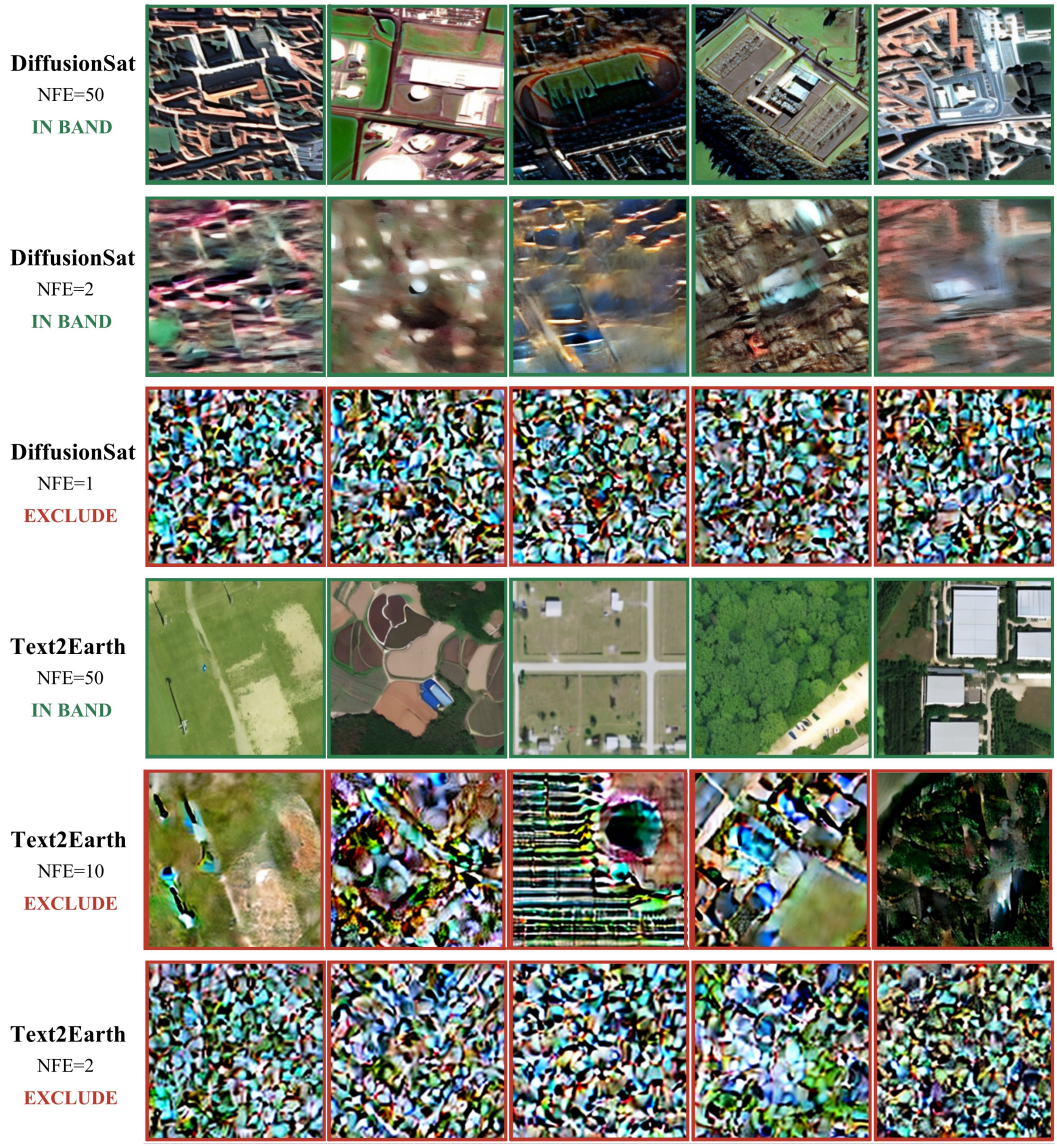


Figure 7: **Fixed-index samples supporting the quality gate.** Green **IN BAND** and red **EXCLUDE** labels denote passing and failing the scalar neighbor-difference criterion, respectively. DiffusionSat NFE$=2$ passes this scalar check but fails the full gate because of directional smearing, while DiffusionSat NFE$=1$ and Text2Earth NFE≤ 10 exhibit severe degradation and are excluded.

C.1 Orbit Approximation and Estimator Scope

For a finite group, the theoretical orbit average is an exact finite sum. The historical implementation, however, precomputes at most `max_g=24` permutation matrices. Consequently, the $N = 4$ experiment enumerates the complete group S_4, whereas for $N \geq 5$ the reported likelihood certificate uses a capped 24-element approximation to the full permutation orbit. Since

$$|S_5| = 120, \qquad |S_6| = 720, \qquad |S_7| = 5040, \qquad |S_8| = 40320, \tag{13}$$

we do not describe the larger-N values as exact-orbit measurements.

This approximation affects the absolute value of the certificate and therefore limits quantitative claims that compare certificate magnitudes across large permutation groups. It does not affect the independently implemented sample-space C2ST measurements. The Energy-Distance checks below also avoid the capped matrix set by applying independently sampled permutations from the full S_N distribution to individual samples. We therefore use the large-N certificate curves as diagnostic evidence about the behavior of the historical likelihood estimator, while the central conclusion about the one-step failure is supported jointly by the exact S_4 case, the round-trip diagnostic, C2ST, and permutation-sensitive Energy Distance.

C.2 Two-Sided Estimator Validation

On architecturally equivariant control flows using an EGNN velocity field, the exact-enumeration S_4 estimate satisfies

$$|D_{\mathrm{sym}}| < 10^{-15} \tag{14}$$

across the tested sampling budgets, demonstrating that the orbit and likelihood implementation does not introduce a measurable defect when the target dynamics are equivariant.

For a positive control, we inject the explicit permutation-breaking term

$$v(x) \leftarrow v(x) + \alpha \,\mathrm{diag}(i)\, x. \tag{15}$$

On the enumerable validation system, the certificate reads

$$0.0000, \quad -0.008, \quad 0.086, \quad 0.693 \tag{16}$$

for

$$\alpha \in \{0, 0.05, 0.2, 1\}, \tag{17}$$

while the corresponding score-equivariance residuals are

$$5.1 \times 10^{-16}, \quad 0.076, \quad 0.299, \quad 1.17. \tag{18}$$

The estimator therefore responds monotonically once the injected effect rises above its numerical floor. Since $D_{\mathrm{sym}} \geq 0$ theoretically, the small negative estimate at $\alpha = 0.05$ is interpreted as estimator noise rather than as a physical negative defect.

C.3 Round-Trip Diagnosis

Sampling integrates the probability-flow ODE from $t = 0$ to $t = 1$, while likelihood evaluation integrates the same dynamics in reverse and accumulates the divergence. A finite-step RK4 trajectory is not exactly reversible. We therefore map

$$U_0 \to X \to \widehat{U}_0$$

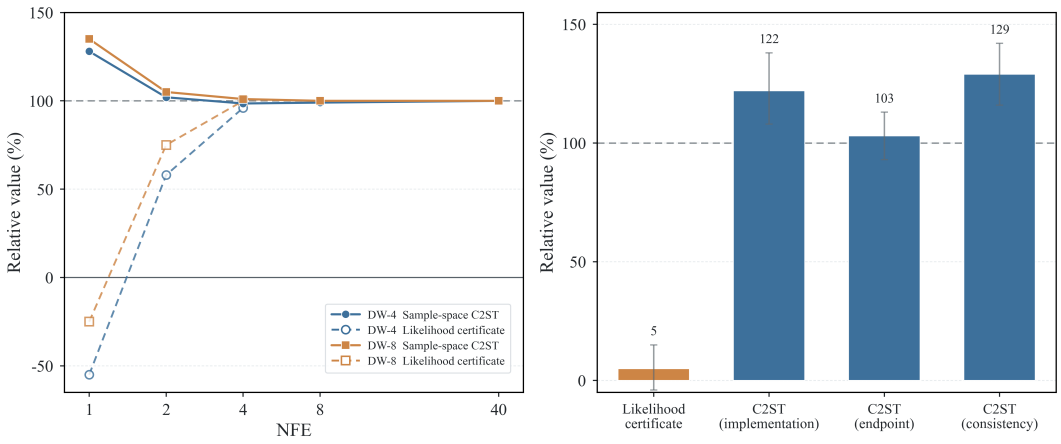
using the same number of forward and reverse steps and measure

$$\epsilon_{\mathrm{rt}} = \frac{\|\widehat{U}_0 - U_0\|_2}{\|U_0\|_2}. \tag{19}$$

The measured errors are

$$\begin{array}{cccccc} K & 1 & 2 & 4 & 8 & 40 \\ \epsilon_{\mathrm{rt}} & 0.277 & 0.0343 & 0.0045 & 0.0004 & 0.0000. \end{array} \tag{20}$$

At one step, the reverse trajectory therefore lands approximately 28% away from the state from which the forward integration started. In the same regime, the likelihood-derived certificate reaches -0.121 for $N = 4$ and -0.116 for $N = 5$, despite the theoretical quantity being non-negative. We therefore regard one-step certificate values as numerically inadmissible. NFE$= 2$ retains a substantial 3.4% round-trip error and is treated with caution, whereas from NFE$= 4$ onward the error falls below 0.005.



(a) Few-step measurement sweep

(b) Teacher-relative measurements

Figure 8: **Likelihood and sample-space measurements in the few-step regime.** The likelihood-derived certificate deteriorates under aggressive integration and one-step distillation, whereas independent sample-space measurements remain near the corresponding many-step or teacher-level asymmetry.

C.4 Distillation Protocol

To test whether the apparent loss of symmetry under aggressive few-step evaluation reflects a genuine change in the sampled distribution or a failure of the likelihood-derived measurement, we construct distilled students on the DW-N systems. Teachers are non-equivariant conditional-flow-matching models trained on the corresponding DW-N distributions. A K-step student shares the teacher architecture and is optimized by endpoint matching,

$$\left\| \Phi^{K}_{\text{student}}(x_0) - \Phi^{40}_{\text{teacher}}(x_0) \right\|_2^2, \tag{21}$$

where Φ^K denotes the K-step ODE pushforward. Five independently initialized students are trained for each evaluated configuration.

For the one-step experiments used in the measurement comparison below, we use online distillation, which recomputes the teacher's 40-NFE endpoint for fresh prior draws during training. We found that repeatedly reusing a fixed pool of precomputed teacher targets can introduce a downward bias in the $K = 1$ certificate-side measurement. The online protocol is therefore treated as authoritative for the one-step analysis. Higher-step students are used only as reference checks and are not required for the main conclusion.

As a coarse fidelity control, the 1-Wasserstein distance between the student and teacher interatomic-distance distributions remains at most 0.03 across the evaluated settings. This permutation-invariant statistic rules out gross geometric collapse, but cannot determine whether atom-label asymmetry has been preserved. We therefore complement it with the permutation-sensitive sample-space measurements below.

C.5 Likelihood versus Sample-Space Measurements

We compare the likelihood-derived certificate with independent sample-space measurements under progressively coarser integration and one-step distillation. Figure 8 summarizes the NFE sweep on DW-4 and DW-8 and the teacher-relative measurements from the one-step DW-5 experiment.

Across NFE 40, 8, 4, and 2, sample-space C2ST remains close to its 40-step reference for both DW-4 and DW-8. By contrast, the likelihood-derived certificate falls to approximately 58% and 75% of its many-step value at NFE$= 2$, respectively. At NFE$= 1$, both likelihood estimates become negative despite the theoretical certificate being non-negative, whereas the sample-space measurements remain at approximately 128–135% of their reference values.

The disagreement at NFE$= 2$ alone does not establish estimator failure because the likelihood certificate and C2ST measure different distributional functionals. The negative one-step estimates provide a stronger diagnosis: the likelihood computation becomes numerically inadmissible while sample-space measurements continue to detect asymmetry in the same generated distributions.

The one-step DW-5 distillation experiment provides an independent cross-check. The likelihood-derived estimate retains only approximately 5% of the teacher defect, whereas the implementation-, endpoint-, and consistency-based C2ST measurements retain approximately 122%, 103%, and 129% of the teacher-level asymmetry. These results provide no evidence that the one-step student has become symmetric. Together with the forward–reverse round-trip diagnostic, they instead localize the discrepancy to numerical unreliability of the likelihood-derived measurement in the one-step regime.

C.6 Permutation-Sensitive Energy-Distance Checks

The interatomic-distance fidelity control above is permutation-invariant and therefore cannot determine whether atom-label asymmetry has been preserved. We consequently introduce an additional permutation-sensitive sample-space check based on Energy Distance in the orthonormal mean-free particle basis.

Let T, S, and $\bar{T}$ denote teacher, student, and teacher-orbit-average samples, respectively. We evaluate

$$r = \frac{\mathcal{E}(S, \bar{T})}{\mathcal{E}(T, \bar{T})}, \tag{22}$$

where

$$\mathcal{E}(X, Y) = 2\,\mathbb{E}\|X - Y\| - \mathbb{E}\|X - X'\| - \mathbb{E}\|Y - Y'\|. \tag{23}$$

For this sample-space measurement, the orbit-average distribution is obtained by applying an independently and uniformly sampled permutation from the full S_N distribution to each teacher sample. This procedure therefore produces a Monte Carlo sample from the full orbit average and does not use the capped 24-matrix approximation employed by the likelihood-derived certificate for larger permutation groups.

We use the U-statistic form of Energy Distance with self-distances excluded. The estimator is unbiased but can fluctuate slightly below zero when the two distributions agree. As a calibration check, two independent 2000-sample draws from the same Gaussian give -0.0009, whereas two distinct Gaussians give $+0.818$. Because the ratio above divides by $\mathcal{E}(T, \bar{T})$, we report it only when the reference value exceeds the within-teacher estimator floor by at least a factor of five.

For $N = 5$, the measured ratios are

$$r_{K=1} = 1.25 \pm 0.38, \qquad r_{K=4} = 1.18 \pm 0.38, \tag{24}$$

and for $N = 7$,

$$r_{K=1} = 1.09 \pm 0.12, \qquad r_{K=4} = 1.12 \pm 0.20. \tag{25}$$

Values near one place the student near the teacher end of the teacher–symmetrized-teacher comparison rather than near the symmetrized teacher.

We additionally measure the student's distance to its own symmetrization. For $N = 5$ at $K = 1$,

$$\mathcal{E}(S, \bar{S}) = 0.0081 \pm 0.0033, \tag{26}$$

compared with

$$\mathcal{E}(T, \bar{T}) = 0.0030 \tag{27}$$

and an estimator floor of approximately -0.0012. For $N = 7$,

$$\mathcal{E}(S, \bar{S}) = 0.0073 \pm 0.0025, \qquad \mathcal{E}(T, \bar{T}) = 0.0099. \tag{28}$$

Together, the Energy-Distance measurements, C2ST results, and forward–reverse round-trip diagnostic consistently reject the interpretation that one-step distillation has erased the sampled distribution's asymmetry. Instead, they localize the severe one-step discrepancy to the likelihood-derived measurement.

D DEPLOYMENT PROVENANCE ON IMAGENET-64

This section provides a released-checkpoint test of whether the symmetry profile of a deployed artifact can be inferred from architecture, ancestor behavior, or sampling budget alone. We use the public ImageNet-64 consistency-model family because it provides multiple artifacts with the same model architecture but different training provenance and deployment step counts. The experiment is intended as a deployment-level counterexample rather than as a broad benchmark of natural-image symmetry.

D.1 CHECKPOINTS AND AUDIT PROTOCOL

We evaluate horizontal reflection on the public ImageNet-64 consistency-model checkpoints of Song et al. (2023). The evaluated arms are real ImageNet-64 data, the EDM diffusion teacher sampled with Heun at 40 steps, one-step consistency-distillation models trained with either ℓ_2 or LPIPS objectives, a two-step deployment of the CD-LPIPS checkpoint, and a one-step consistency-training model trained from scratch. The generated checkpoints use the same released 295.9M-parameter architecture.

Each generated arm contains a matched pool of

$$n = 4 \times 10^4 \tag{29}$$

samples. Class labels follow the same fixed sequence and prior noise is deterministic for each sampling seed. Each generated image belongs either to the original pool or to the horizontally reflected pool, never both. No reflection augmentation is used in classifier training. The C2ST discriminator is evaluated with five-fold cross-validation and three training epochs per fold, using three observation seeds for the reported model readings.

Reaching the full pool size requires merging two generation seeds. We therefore test whether sampling-seed identity itself is detectable. A classifier trained to distinguish seed 42 from seed 43 obtains accuracy 0.4966, and at $n = 10^4$ the merged pool reproduces the corresponding single-seed symmetry reading to within 10^{-4}. We find no evidence that seed pooling introduces a measurable shortcut.

Negative controls replace the horizontal-reflection comparison with two independent halves of the same artifact distribution. They remain near chance for every model arm, ranging from 0.4971 to 0.5024. A coarse image-quality diagnostic also rules out trivial chance-level behavior caused by output collapse: the mean absolute horizontal-neighbour difference is approximately 13 for genuine ImageNet-64 output and approximately 85 for white noise, whereas all evaluated generated arms lie between 10.4 and 12.0 with comparable per-image variation.

D.2 POWER CALIBRATION

Sensitivity is calibrated with three controlled perturbations: a smooth intensity ramp, a high-frequency checkerboard applied to one half of the image, and an off-centre bright patch. At large amplitude the ramp is affected by pixel-range clipping, whereas in the low-amplitude regime relevant to the audit the localized patch is the most sensitive perturbation. At amplitude 0.01, for example, the patch reaches approximately $z = 3.68$ relative to the empirical null, compared with approximately $z = 0.34$ for the smooth ramp. We therefore use the localized patch as the practical sensitivity reference for this experiment.

D.3 PROVENANCE AUDIT

Table 6 reports the complete horizontal-reflection audit. Real ImageNet-64, the many-step teacher, and all distilled deployments remain indistinguishable from their negative controls. The one-step model trained from scratch is the only arm with a clear reflection defect.

Values of the final column are the non-negative empirical discriminator excess, $\max\{0, 2(\mathrm{C2ST} - 0.5)\}$, shown only as an instrument-relative effect-size aid. Below-chance finite-sample readings are therefore displayed as approximately zero rather than as negative separation.

Table 6: **Horizontal-reflection audit of released ImageNet-64 consistency models.**

Artifact	Flip C2ST	Negative control	$2(\text{C2ST} - 0.5)$
Real ImageNet-64	0.4934	–	≈ 0
EDM teacher, 40-step	0.5009	0.5022	≈ 0
CD-ℓ_2, 1-step	0.4981 ± 0.0037	0.4979	≈ 0
CD-LPIPS, 1-step	0.5001 ± 0.0011	0.5001	≈ 0
CD-LPIPS, 2-step	0.4994	0.5024	≈ 0
CT from scratch, 1-step	$\mathbf{0.559 \pm 0.020}$	0.4971	$\approx \mathbf{0.12}$

The comparison separates deployment provenance from step count. The CD-LPIPS checkpoint changes from 0.5001 at one step to 0.4994 at two steps, a difference of only 0.0007. By contrast, changing the training route while holding the one-step deployment configuration fixed changes the reading by approximately 0.060: the distilled model remains at chance, whereas the consistency-trained model reaches 0.559 ± 0.020. The real data and many-step teacher are also near chance, so the detected defect is not inherited from the reference distribution or from the teacher.

This experiment does not establish that provenance is the unique mechanism determining symmetry. It establishes the narrower deployment-facing point used in the main paper: artifacts with the same architecture and nominal sampling budget can have different structural profiles, so neither architecture nor step count alone can substitute for an audit of the artifact that is actually deployed.

E Architectural Boundary Case: MatterGen

The main experiments focus on structural properties that are not guaranteed by the deployed architecture. MatterGen provides a complementary boundary case: its geometric backbone explicitly respects global translation symmetry, whereas crystallographic space-group equivariance is not built into the model. We use this contrast to illustrate when post-hoc symmetry measurement primarily verifies an architectural guarantee and when artifact-level auditing remains informative.

We evaluate the pretrained `mattergen_base` checkpoint with its 44.5M-parameter GemNet denoiser using the score-equivariance residual

$$R_G = \frac{\mathbb{E}_{x,g}\left[\|s_\theta(g{\cdot}x) - g{\cdot}s_\theta(x)\|_2^2\right]}{\mathbb{E}_x\left[\|s_\theta(x)\|_2^2\right]}. \tag{30}$$

MatterGen does not expose a probability-flow likelihood over all of its discrete and continuous variables, so R_G is used only as a white-box equivariance diagnostic and is not compared numerically with D_{sym} or the sample-space C2ST measurements.

As an architecturally supported control, we measure global translation equivariance using 16 generated structures and 8 random torus translations per structure. For noise levels $t \leq 0.3$, the residual remains

$$R_{\text{trans}} = 1.1\text{–}2.2 \times 10^{-3}. \tag{31}$$

At substantially higher noise, the residual increases as the samples approach the prior-dominated regime; we therefore restrict the comparison below to $t \leq 0.3$.

We next evaluate crystallographic operations that are symmetries of the generated crystal structures but are not explicitly guaranteed by the architecture. Space groups and symmetry operations $\{(R, \tau)\}$ are identified using `spglib` with `symprec`=0.1. Among 128 generated structures, 114 contain non-trivial detected operations, covering 43 distinct space groups and 686 operation evaluations. The resulting residuals lie in the range

$$R_{\text{sg}} = 0.60\text{–}2.88, \tag{32}$$

which is orders of magnitude larger than the translation control. Because between-structure heterogeneity dominates the raw variance, uncertainty is aggregated at the generated-structure level; under this clustering, the measured space-group residuals lie approximately 2.7–7.7 clustered standard errors from zero. Together, the near-zero translation residual and the substantially larger residuals under the evaluated crystallographic operations illustrate the distinction between an architecturally

supported control and transformations whose equivariance is not guaranteed under the evaluated representation.

This comparison defines the intended architectural boundary of GenSymAudit. When a symmetry is guaranteed by construction, a post-hoc measurement primarily checks whether the realized implementation respects that guarantee. The more consequential setting concerns structural properties that are not jointly guaranteed by the architecture, training procedure, conditioning protocol, and deployment sampler. These are precisely the cases in which the realized artifact must be audited directly.

F Exploratory Downstream Association

This section provides the full measurements underlying the exploratory downstream analysis in Sec. 4.4. The experiment asks whether the directional structure localized by the audit is reflected in classifiers trained on generated rather than real imagery. It does not establish that the overall synthetic-data utility gap is caused by symmetry, nor does it establish an element-wise causal dose–response relationship.

F.1 Tasks and Training Protocol

We evaluate two downstream tasks. The first is 62-class scene classification on fMoW; the second predicts the six-level zoom category in Git-10M. For each corpus, a ResNet-18 is trained from random initialization on approximately $23{,}000$ images and evaluated on held-out real imagery. An ImageNet-pretrained initialization is avoided in the primary experiment because it carries its own orientation prior.

The real-trained and generated-trained arms use the same training instances at the same sample count; only the source of the image pixels differs. Inputs are 256×256. Training uses AdamW with learning rate 3×10^{-4} and cosine decay for 60 epochs with batch size 128. Five independent training seeds are used for every arm.

No rotation or reflection augmentation is used because such augmentation would directly suppress the directional dependence being measured. The only augmentations are orientation-neutral transformations such as isotropic scale, translation, and colour perturbation. Evaluation applies each $g \in D_4$ to the same held-out real test images through exact pixel permutations before normalization.

For each transformation, orientation reliance is measured as

$$S(g) = 1 - \frac{\mathrm{Acc}(g \cdot x)}{\mathrm{Acc}(x)}, \tag{33}$$

and the generated-versus-real change is

$$\Delta S(g) = S_{\mathrm{gen}}(g) - S_{\mathrm{real}}(g). \tag{34}$$

Negative ΔS indicates that the generated-trained classifier is less orientation-reliant under that transformation, while positive ΔS indicates increased reliance.

A real-versus-real control is constructed from classifiers trained on disjoint real-data halves. Additional generated half-arms use the same reduced training size, allowing the effect of data source to be separated from the effect of training-set size.

F.2 Overall Task Performance

Table 7 reports top-1 accuracy on the held-out real test set. Training on generated imagery produces a substantial reduction on both tasks: from 0.3522 to 0.2049 on fMoW-62 and from 0.9491 to 0.6453 on Git-10M zoom, corresponding to relative reductions of approximately 41.8% and 32.0%.

The half-scale comparisons show that the utility gap persists after matching the reduced training-set size. Nevertheless, these aggregate accuracy differences are not treated as symmetry effects: generated and real imagery can differ simultaneously in fidelity, coverage, texture statistics, semantic diversity, and other properties that influence downstream performance.

Table 7: **Downstream top-1 accuracy on held-out real data.** Values are mean±standard deviation over five training seeds.

Dataset	Real	Generated	Real-A	Real-B	Gen-A	Gen-B
fMoW-62	0.3522 ± 0.0039	0.2049 ± 0.0061	0.2958 ± 0.0059	0.2887 ± 0.0040	0.1707 ± 0.0057	0.1751 ± 0.0060
Git-10M zoom	0.9491 ± 0.0021	0.6453 ± 0.0160	0.9185 ± 0.0040	0.9195 ± 0.0039	0.5916 ± 0.0105	0.6002 ± 0.0158

Table 8: **Element-wise downstream orientation reliance.** Values are measured on held-out real test imagery; p is a two-sided Welch comparison across five training seeds.

Dataset	D_4 element	S_{real}	S_{gen}	ΔS	p
fMoW-62	Rotation 90°	0.2725	0.1837	−0.0889	0.0001
	V: Rotation 180°	0.1631	0.1355	**−0.0275**	0.0107
	Rotation 270°	0.2900	0.2084	−0.0816	0.0002
	V: Horizontal flip	0.0777	0.0621	**−0.0156**	0.1128
	V: Vertical flip	0.1392	0.1303	**−0.0089**	0.3641
	Transpose	0.2218	0.1533	−0.0685	0.0003
	Anti-transpose	0.3191	0.2399	−0.0792	0.0038
Git-10M zoom	Rotation 90°	0.2209	0.1584	−0.0625	0.0284
	V: Rotation 180°	0.0309	0.0709	**+0.0399**	0.0030
	Rotation 270°	0.2394	0.1669	−0.0725	0.0152
	V: Horizontal flip	0.0217	0.0397	**+0.0180**	0.0022
	V: Vertical flip	0.0372	0.0560	**+0.0188**	0.0001
	Transpose	0.2231	0.1542	−0.0689	0.0145
	Anti-transpose	0.2198	0.1722	−0.0477	0.0595

F.3 Element-Wise Orientation Reliance

Table 8 reports the complete seven-element directional response. The three non-identity elements of the axis-preserving subgroup are marked by V. The reported Welch p-values compare the five training-seed measurements and are retained as descriptive statistics rather than as the basis of the paper's subgroup-level conclusion.

The two corpora exhibit different signed responses. On fMoW, ΔS is negative for every transformation, so the generated-trained classifier is generally less orientation-reliant than the real-trained classifier. On Git-10M, the three axis-preserving transformations instead exhibit positive changes, while the four axis-exchanging transformations remain negative. This difference is one reason we avoid interpreting ΔS as a direct downstream consequence of the signed audit shift.

F.4 Real-versus-Real Controls

The disjoint real-data half-arms assess how replacing the training set changes performance without switching from real to generated imagery. The corresponding generated half-arms use the same reduced sample count. Table 7 shows that generated-trained classifiers have lower held-out accuracy at both the full and reduced training sizes, so the observed utility gap is not explained solely by the difference between these training-set sizes.

For orientation reliance, earlier diagnostics also compared the magnitude of each generated-versus-real change with a real-versus-real half-arm difference. Such ratios become unstable when the real-versus-real denominator is close to zero and are therefore not used for inference here. Moreover, a ratio based on $|\Delta S|$ contains no information about the sign of the directional response. We use the signed measurements in Table 8 to discuss increases or decreases in orientation reliance and retain the half-arms as sample-size and training-set replacement controls.

F.5 Element-Wise Association and Subgroup Pattern

We correlate the signed anchor-relative audit shift $\widehat{\Delta}_g$ with the signed downstream change $\Delta S(g)$ over the seven non-identity elements of D_4. The audit shifts are taken from Tables 3 and 4, and the

downstream changes from Table 8. Elements are aligned by transformation identity rather than by row position.

Two-sided exact permutation tests for Spearman's correlation give

$$\rho_{\mathrm{fMoW}} = 0.82, \qquad p_{\mathrm{fMoW}} = 0.034, \tag{35}$$

and

$$\rho_{\mathrm{Git}} = 0.36, \qquad p_{\mathrm{Git}} = 0.44. \tag{36}$$

Each test enumerates all $7! = 5040$ permutations of one rank vector while holding the other fixed and counts correlations whose absolute value is at least the observed value. These p-values are unadjusted across datasets. The permutation calculation describes the association among the estimated element-wise effects and does not propagate their seed- or pool-level estimation uncertainty.

The positive association reaches the nominal 0.05 level on fMoW but not on Git-10M, so it does not provide consistent cross-dataset evidence. On fMoW, all seven $\Delta S(g)$ values are negative: the positive correlation means that larger audit shifts tend to accompany less negative downstream changes, not that orientation reliance increases above the real-trained baseline. With only seven elements and unadjusted tests, the result is exploratory and does not establish a causal dose–response relationship.

A ranking pattern shared by both corpora appears at subgroup resolution. The three non-identity elements of

$$V = \{e, r_{180}, \mathrm{flip_h}, \mathrm{flip_v}\} \tag{37}$$

occupy the three largest signed ΔS values, while the four axis-exchanging transformations occupy the remaining ranks. For a fixed three-element subset and the directional alternative that it occupies the upper ranks, there are

$$\binom{7}{3} = 35 \tag{38}$$

equally likely subset assignments under rank-label exchangeability. Complete separation therefore gives the one-sided exact probabilities

$$p_{\mathrm{fMoW}} = p_{\mathrm{Git}} = \frac{1}{35} \approx 0.029. \tag{39}$$

Fisher combination of these two nominal probabilities, under independence of the corpus-level tests, gives

$$p_{\mathrm{combined}} \approx 0.007. \tag{40}$$

These subgroup probabilities are distinct from the two-sided Spearman tests above. As discussed in Appendix B.4, the partition was identified after inspecting the audit results; neither the partition nor the directional alternative is treated as preregistered. The nominal probabilities do not adjust for this selection or for other possible partitions and directions. We therefore report the shared ranking as an exploratory cross-corpus pattern, not as an independently confirmed subgroup hypothesis or a causal explanation of the utility gap.

F.6 Scope of the Downstream Evidence

The downstream experiments support a limited structural association. Training on generated imagery changes both overall task performance and orientation dependence. The element-wise association is positive and nominally significant on fMoW but not on Git-10M, whereas the axis-preserving versus axis-exchanging ranking is shared by both corpora. The signed responses nevertheless differ: fMoW exhibits negative ΔS for every transformation, while Git-10M exhibits positive changes for the three non-identity axis-preserving elements. These observations do not identify symmetry as the cause of the synthetic-data utility gap.

The two tasks also have complementary limitations. fMoW-62 reaches approximately 0.35 top-1 accuracy on held-out real data when trained from scratch at this scale. Because $S(g)$ is normalized by each classifier's clean accuracy, lower orientation reliance must not be interpreted as better absolute transformed-image performance. Git-10M zoom classification reaches approximately 0.95 top-1 accuracy on held-out real data for the real-trained classifier, but zoom prediction is a limited proxy for practical remote-sensing applications. Neither experiment evaluates detection or semantic segmentation. We therefore describe the results as exploratory downstream associations rather than as general or causal validation of the utility of the audited symmetry shifts.