Skepthical review: *Scattering transform synthesis of correlated foregrounds: benchmarking against diffusion models on FLAMINGO*

Summary

This manuscript studies generative modelling of *correlated* extragalactic foreground maps from FLAMINGO, focusing on the joint tSZ+CIB field (and extensions to tSZ+CIB+kSZ and multi-frequency CIB). The core technical contribution is a **generator-agnostic calibration/post-processing operator** that alternates (i) an $N \times N$ **per-ℓ-bin Fourier-space whitening/re-colouring** (implemented via matrix square-roots/Cholesky; Eq. (12)) to match the full auto- and cross-power spectra, with (ii) **pixel histogram (rank/CDF) matching** (paired and ensemble variants) to enforce 1-point marginals. The paper benchmarks this operator when applied to three upstream generators—Scattering Transform / ScatCov microcanonical synthesis, a jointly trained DDPM, and a Gaussian random-field baseline—and finds that after calibration they become effectively indistinguishable on a wide battery of diagnostics (Secs. 4–5, 7.3). A second, conceptually important contribution is a "non-by-construction" ladder (Sec. 6) that quantifies what multi-channel ScatCov can recover *without* truth-based projections: e.g. the tSZ×CIB pixel cross-correlation saturates around ∼60% (with soft C_ℓ rescaling improving some 2-pt agreement but leaving tail/peak limitations). The manuscript is technically rich and empirically broad; the main opportunities for improvement are (i) sharper framing of what is *guaranteed* vs *learned* (and what that implies about discriminative power of diagnostics), (ii) clearer treatment of paired histogram conditioning and patch-to-patch variability, and (iii) more explicit numerical/statistical details around Eq. (12), uncertainty quantification, and recommended "which recipe for which science goal" guidance.

Strengths

- The $N \times N$ per-ℓ-bin covariance matching in Fourier space (Eq. (12)) is a simple, elegant, and scalable generalisation of single-field amplitude rescaling to correlated multi-component foreground vectors (Secs. 4.1.1, 5.3–5.4).
- The paper provides an unusually comprehensive evaluation suite on FLAMINGO (power spectra and cross-spectra, pixel cross-r, 1-point CDFs including deep tails, Minkowski functionals, selected 3-/4-point and gradient statistics, cluster radial profiles, peak counts and peak 2-pt functions, plus a downstream ILC test; Secs. 4–7).
- The falsification test showing that a Gaussian random field plus calibration can match many downstream diagnostics is scientifically valuable: it forces correct attribution of performance to the calibration operator rather than the generator (Secs. 5.5–5.6).
- The "non-by-construction" analysis (Sec. 6) is conceptually important and practically useful: it quantifies an apparent expressive ceiling of multi-channel ScatCov for tSZ×CIB correlation (∼60%) and provides a near-zero-training-cost alternative to trained diffusion baselines in some regimes.
- Clear discussion (especially Sec. 7.5) of supervised (DDPM) vs unsupervised (microcanonical/ST) vs semi-supervised (calibration requiring ensemble C_ℓ/CDF targets) dependence, with thoughtful remarks about simulation bias and real-sky deployment.

- The extensions that target specific known failures—band-pass histogram + Cholesky alternation for scale-resolved skewness/kurtosis and a peak-aware dispersion step for deep-peak deficits—are pragmatic, generator-agnostic, and empirically promising (Secs. 7.2–7.4).

Major issues

1. **The manuscript's central logical separation—what is enforced by construction vs what is genuinely learned—is crucial but not made maximally explicit at the point where the reader first encounters the main results.** As a consequence, several impressive-looking agreements in Secs. 4–5 (auto-/cross-C_ℓ, 1-point CDFs; and potentially derived summaries) risk being interpreted as generator performance rather than calibration guarantees, and it becomes hard to identify which diagnostics actually test generator expressivity (e.g. morphology/topology, rare-event spatial structure, disjoint-band higher-order couplings).

   Recommendation: Add, immediately after Sec. 4.1.1, a concise table (or boxed bullets) mapping every diagnostic used in Secs. 4–7 into: (i) **strictly enforced** (specify which step: Eq. (12) C_ℓ-match; final histogram match; BP extension; dispersion), (ii) **partially controlled/indirectly affected**, (iii) **not enforced** (therefore informative about the generator/remaining degrees of freedom). Then, when presenting results in Secs. 4.1–4.3, 5.1, 5.5–5.6, and 7.2–7.4, label panels/text as "by construction" vs "not by construction" so the reader can immediately see what is being validated vs what is being learned.

2. **The paired pixel histogram (rank/CDF) match conditions each generated sample on the exact empirical 1-point distribution of the paired truth patch.** While the paper argues (via near-zero pixelwise correlation) that this is not simply copying, the conditioning is still very strong and can (a) be perceived as information leakage for benchmarking, and (b) artificially shrink patch-to-patch variability in 1-point-derived summaries (skewness/kurtosis/minima) relative to what an unconditional generator should reproduce—especially relevant for covariance estimation and SBI use-cases.

   Recommendation: Strengthen Secs. 4–5 and Sec. 6 by explicitly distinguishing **paired mode** (per-patch CDF is imposed) from **ensemble mode** (pooled CDF target). Add one compact figure/table showing the **distribution across patches** of key 1-point summaries (e.g. patch mean, variance, skewness, kurtosis, min/max, deep-tail quantiles) for: truth vs paired-mode outputs vs ensemble-mode outputs. In text, clarify for which scientific tasks paired-mode is appropriate (e.g. controlled stress-tests) and where it is not (e.g. unconditional ensemble sampling for uncertainty quantification).

3. **The paper repeatedly claims that the alternating procedure "joint Cholesky C_ℓ-match + histogram match" recovers both the full $N \times N$ C_ℓ matrices and the full 1-point CDFs 'by construction' (Abstract; Sec. 4.1.1; Sec. 5.6).** As written, this is mathematically overstated: a histogram/rank match generally changes C_ℓ, and a Fourier recolouring generally changes the pixel histogram. Simultaneous satisfaction is an *empirical* outcome of alternation, and also depends on the **final operator order**.

   Recommendation: In Sec. 4.1.1, precisely specify the operator ordering and what constraints are exact at the final output (e.g. "final step is histogram match → CDF exact, C_ℓ within tolerance", or vice versa). If claiming simultaneous exactness, provide a justification (e.g. alter-

nating projections under stated assumptions) or weaken the claim to "to numerical tolerance after K alternations" and report typical residual errors in both constraints after the final iteration.

4. **Eq. (12) is central, but key numerical/implementation details are currently too implicit for robust reproduction and for interpreting the $N > 2$ results (Secs. 4.1.1, 5.3–5.4).** In particular: how $C(\ell)$ is estimated (binning, apodisation/windowing), how matrix square roots/inverses are computed per ℓ-bin, how near-singular or non-PSD estimates are handled, and how the Fourier-space transform is applied while preserving real-valuedness (Hermitian symmetry) and avoiding edge/periodicity artefacts.

   Recommendation: Add an implementation subsection (end of Sec. 4.1.1 or an appendix) specifying: (i) ℓ-binning (flat-sky $|\boldsymbol{k}|$, bin edges, min modes/bin), (ii) window/apodisation used when estimating $C(\ell)$ and when applying Eq. (12) (and consistency between them), (iii) square-root convention (Cholesky vs symmetric eigensqrt/SVD) and regularisation (eigenvalue clipping/diagonal loading), (iv) how transforms are applied to $\boldsymbol{k}$ and $-\boldsymbol{k}$ to preserve Hermitian symmetry and real maps, and (v) typical numerical residuals in matched bandpowers for $N = 2/3/4$.

5. **The manuscript's 'big picture' framing sometimes still reads like a generator benchmark (ST vs DDPM), whereas the dominant scientific lever is the calibration mapping and the equivalence class it induces (ST/DDPM/Gaussian becoming similar after constraints).** This is a strong and somewhat surprising conclusion whose implications are underdeveloped: it could mean (a) Gaussian phases + enforced 1-pt/2-pt are sufficient for many targets *in this setting*, or (b) the chosen diagnostics are insufficiently discriminative once those constraints are enforced.

   Recommendation: Tighten the framing in Sec. 1, Secs. 5.5–5.6, and Sec. 7.5–7.6 to explicitly state the main object as the **calibration operator** and discuss what it implies about diagnostic discriminativity under enforced 1-pt/2-pt constraints. Add a short subsection: "When do advanced generators add value beyond Gaussian+calibration?" (conditional sampling, non-stationarity/masks, additional tracers, rare-event spatial patterns, survey systematics). Consider modestly revising title/abstract emphasis accordingly.

6. **Claims of "indistinguishability" and generator-agnostic behaviour are sometimes made with small-N patch counts (often $N \approx 20$; sometimes $N \approx 5$ for 3- and 4-channel experiments in Secs. 5.3–5.4) and without uniform uncertainty visualisation across figures (many plots in Secs. 4–7).** Given strong intra-patch spatial correlations, the effective number of independent samples is not obvious, and some tests risk overconfidence.

   Recommendation: Adopt a consistent uncertainty protocol across the headline diagnostics (Secs. 5.5–5.6, 7.2–7.4): bootstrap over patches and report mean±std and/or 95% CIs, and add error bands/error bars in the most important figures. Clearly mark $N = 5$ results as feasibility demonstrations unless expanded. Where possible, increase N for the key multi-channel/multi-frequency claims or explicitly qualify conclusions as "within current sampling noise."

7. **The paper introduces multiple variants (soft C_ℓ rescale; joint match in paired vs ensemble mode; BP+Cholesky extension; dispersion step; non-by-construction ensemble pipeline in Sec. 6), but the recommended use of each for concrete science**

goals is scattered and easy to miss. The non-by-construction pipeline is described as "deployable" (Sec. 6.1.0.3), but its limitations (tails/peaks, residual cross-r, scale-resolved higher moments) need clearer front-and-centre guidance for practitioners.

Recommendation: Add a "recipe selection / decision table" near the end of Sec. 6 or in Sec. 8 mapping each pipeline variant to application scenarios (2-pt-only covariance estimation; tail/cluster-sensitive analyses; compsep/foreground marginalisation; SBI training; multi-frequency coherent simulations; real-sky constraints). For each, list required inputs (paired truth vs ensemble targets), what is guaranteed vs approximate, typical residuals (cross-r, tail quantiles, S_3/K_4, peak counts), and rough compute cost.

8. **The very high ScatCov coefficient correlations after calibration (including for Gaussian+recipe; Sec. 5.5–5.6) are striking but not fully interpreted.** As presented, a reader may conclude "Gaussian phases are enough," whereas an alternative (and likely) interpretation is that ScatCov correlations become weakly discriminative once the 1-point CDF and $N \times N$ C_ℓ constraints are enforced, at least for the specific ScatCov configuration.

   Recommendation: Add an ablation/diagnostic in Sec. 5.5–5.6: report ScatCov agreement broken down by coefficient order/block (e.g. S_1 vs higher-order; cross-channel blocks only; or excluding coefficients most directly tied to the power spectrum), or after regressing out C_ℓ/CDF effects. Explicitly state whether ScatCov remains informative under the enforced constraints, and in which coefficient subsets it still differentiates generators (if any).

9. **The discussion of simulation bias and real-sky applicability in Sec. 7.5 is valuable but somewhat one-sided: it positions microcanonical/ST matching as unbiased and the calibration/DDPM pathways as inherently simulation-biased, without equally emphasising mitigation options (multi-simulation calibration; anchoring targets to data; hybrid/weak supervision; partial-statistic matching) and without clearly stating which science questions remain well-served even with some simulation dependence.**

   Recommendation: Expand Sec. 7.5–7.6 to include a balanced mitigation roadmap: which calibration targets could be estimated from data (e.g. power spectra/cross-spectra, selected 1-point constraints), how to combine multiple simulation suites to reduce model dependence, and what residual biases matter for representative downstream analyses. Clarify when semi-supervised calibration is appropriate vs when strictly unsupervised synthesis is preferable.

Minor issues

1. Pearson correlation notation is inconsistent: Eq. (5) uses $r(a,b) = \langle ab\rangle / \sqrt{\langle a^2\rangle\langle b^2\rangle}$ but text refers to "demeaned Pearson" and later tables show nonzero means (e.g. Table 4). Without explicit demeaning, Eq. (5) is not Pearson correlation in general.

   Recommendation: Either redefine Eq. (5) with mean subtraction ($a \to a - \langle a\rangle$, $b \to b - \langle b\rangle$) or explicitly state that maps are demeaned before computing r everywhere, and ensure tables/figure captions use consistent terminology.

2. "Phase-preserving" claims are overstated in Sec. 3.5 and around Eq. (12): soft C_ℓ rescaling preserves the complex argument per Fourier coefficient but still filters the field, and the $N \times N$ joint transform generally changes per-channel complex phases when the mixing matrix is

non-diagonal.

Recommendation: Replace with precise statements: soft rescale preserves Fourier phases but changes amplitudes (a filter); Eq. (12) preserves $\boldsymbol{k}$-mode locality (no mode coupling across $\boldsymbol{k}$) but not per-channel phases unless the transform is diagonal. Update any downstream claims relying on stronger invariance (Minkowski/extrema/edges).

3. Pixel scale appears inconsistent: $\mathbf{5}^\circ \times \mathbf{5}^\circ$ patches at $\mathbf{256} \times \mathbf{256}$ imply $\sim \mathbf{1.17}$ arcmin/pixel, not 5 arcmin/pixel.

   Recommendation: Clarify whether there is additional downsampling/smoothing, or correct the stated pixel scale, and ensure consistency across Sec. 2 and figure captions.

4. Data/preprocessing choices needed for replication are incomplete (Sec. 2; also relevant to Sec. 3.5): patch selection (train/val/test indices), mean subtraction conventions, windowing/apodisation for spectra and morphology, beam/smoothing, boundary conditions, and unit conversions (especially kSZ and multi-frequency CIB).

   Recommendation: Add a concise preprocessing checklist/table in Sec. 2 (or an appendix): how patches are selected, what window is used (e.g. Tukey parameters), whether maps are mean-subtracted per patch, any beam smoothing/deconvolution, boundary handling for ST/ScatCov, and explicit unit conversions for each component.

5. The 3-channel (tSZ+CIB+kSZ) and 4-channel multi-frequency CIB experiments (Secs. 5.3–5.4) focus mainly on spectra and cross-$\boldsymbol{r}$, leaving unclear whether non-Gaussian/morphological features of the additional channels are preserved (especially kSZ).

   Recommendation: Add a lightweight non-Gaussian check for the additional channels (e.g. 1-point skew/kurtosis, or a Minkowski/gradient summary for kSZ and extra CIB bands) or state explicitly if they are close to Gaussian at this patch scale and confirm that the recipe does not introduce artefacts.

6. The BP+Cholesky and dispersion extensions (Secs. 7.2–7.4) introduce several hyperparameters (number of BP bins, alternation count, ν_{disp}, ϵ) with limited robustness guidance.

   Recommendation: Provide a short sensitivity/robustness sweep (appendix is fine) showing how $\boldsymbol{S_3}(\ell)/\boldsymbol{K_4}(\ell)$ errors and peak-count closure change under reasonable hyperparameter variations, and recommend default ranges.

7. The DDPM baseline description (Sec. 3.6 and mentions in Sec. 7.6) lacks basic training diagnostics and a compact quantitative comparison between v18 and v21 (pre- and post-recipe), making it harder to judge baseline strength and generality of the "generator-agnostic" conclusion.

   Recommendation: Add minimal training curves/validation metrics (appendix) and a small table comparing v18 vs v21 on a few key metrics (auto-/cross-$\boldsymbol{C}_\ell$ ratios, pixel cross-$\boldsymbol{r}$, $\boldsymbol{M}_1$ peak) before and after calibration.

8. Figure interpretability is hindered by inconsistent acronyms/labels (JM/BP/non-BC), missing units, and limited in-caption methodological detail (binning, normalisation, masking), and many plots lack uncertainty visualisation (noted across multiple figures in Secs. 4–7).

Recommendation: Add a glossary table near Sec. 3/4 defining all pipeline variants and acronyms; standardise figure legends/units/panel labels; and expand key captions to be self-contained (include N, binning, windows).

Very minor issues

1. Various typographical/copyediting issues and corrupted labels appear across equations/tables/figures (e.g. spacing artifacts near Eq. (8), corrupted acronyms in some tables, inconsistent symbol formatting S_2 vs S_2, $K4$ vs K_4).

   Recommendation: Proofread from the LaTeX source, standardise notation, and fix corrupted labels (e.g. tSZ, C_ℓ, cross-r) for publication polish.

2. Matrix-square-root terminology is sometimes ambiguous: the text says "Cholesky C_ℓ-match" but Eq. (12) is written with $C^{1/2}$ and $C^{-1/2}$, which are not unique without a convention.

   Recommendation: Explicitly define the convention (e.g. $C^{1/2} := \mathbf{chol}(C)$ or symmetric eigen-sqrt) so Eq. (12) is unambiguous.

3. A few mathematical/notation details could be tightened: Eq. (1) suppresses orientation indices though $\psi_{j\ell}$ is used; Eq. (2) weights w_q can diverge if reference coefficients are near zero; Sec. 6 states ScatCov synthesis and soft C_ℓ rescale 'commute'.

   Recommendation: Make suppressed indices explicit (or state the suppression clearly), add ϵ-regularisation for w_q, and rephrase the 'commute' statement to a correct weaker claim (e.g. "can be applied post hoc without re-optimising").

4. Some references/citations and cross-references appear incomplete or inconsistently formatted (including the companion 'compsep' paper and several ML citations), and there is an imprecise statement about the number of independent spectra for N channels.

   Recommendation: Audit bibliography completeness/formatting and cross-references; clarify the unique spectra count (N autos + $N(N-1)/2$ crosses, or $N(N+1)/2$ including autos).

Key statements and references

- ✖ **The FLAMINGO simulation suite used as ground truth in this work is a $\sim 10^8$ CPU-hour hydrodynamical computation, yet even at that cost it provides only $\mathcal{O}(10^3)$ patches at 5° scale, far below the sample size required for percent-level covariance estimation on tSZ×CIB cross-spectra or for simulation-based inference training on full-sky patches (Schaye et al. 2023).**
- *Reference(s):* Schaye et al. (2023)
- *Justification:* Schaye et al. (2023) analyzes idealized cloud–wind interactions across hydrodynamics solvers and does not discuss the FLAMINGO simulation suite, its CPU-hour cost, the number of 5° patches, or requirements for tSZ×CIB covariance estimation or simulation-based inference. Thus the statement is not supported by the attached paper.

- ⚠ **The scattering transform provides a translation-invariant, phase-sensitive statistical description of non-Gaussian random fields through successive wavelet modulus averages, as introduced by Mallat (2012) and Bruna & Mallat (2013), and has**

been used to construct generative models that preserve higher-order moments beyond power spectra.

- *Reference(s):* Mallat, 2012, Bruna and Mallat, 2013
- *Justification:* Mallat (2012) defines the scattering transform as successive wavelet modulus operations followed by averaging, yielding translation-invariant representations and proving Lipschitz stability to deformations. It also shows that expected scattering coefficients of stationary (non-Gaussian) processes depend on higher-order moments and can discriminate processes with identical power spectra (Mallat, 2012, Sec. 4; abstract). Bruna and Mallat (2013) reiterate these properties, show texture discrimination using scattering, and employ a PCA-based generative classifier, but do not construct generative models that preserve higher-order moments; moreover, the transform explicitly removes phase via the modulus, not phase-sensitive (Mallat, 2012, Sec. 2.2; Bruna and Mallat, 2013, Sec. 2–3). Thus the claim is only partially supported.

- **✖ Allys et al. (2020) pioneered the use of scattering-transform-based microcanonical synthesis for astronomical foregrounds, showing that LBFGS-optimised synthesis of dust maps matching first- and second-order scattering coefficients (S_1 and S_2) produces statistically indistinguishable synthetic fields, and emphasising that such unsupervised synthesis avoids inheriting simulation bias because it conditions only on the observed realisation.**

- *Reference(s):* Allys et al. (2020)
- *Justification:* Allys et al. (2020) develop Wavelet Phase Harmonics (WPH) statistics and use a microcanonical maximum-entropy synthesis optimized with L-BFGS-B for projected large-scale structure density fields, not astronomical dust foregrounds or the scattering transform S_1 /S_2 (see Sec. II and IV). While they mention related scattering-transform work (e.g., Cheng et al. (2020) and other refs.), this paper does not pioneer scattering-based synthesis, does not match first- and second-order scattering coefficients, and does not claim avoidance of simulation bias by conditioning only on an observed realization—in practice they target WPH statistics estimated from a set of 30 simulation maps (Sec. IV.B). Thus the statement is not supported by Allys et al. (2020).

- **✖ Prabhu et al. (2025) extended scattering-transform work to a denoising diffusion probabilistic model (DDPM) baseline for correlated tSZ×CIB foregrounds, demonstrating that a score-based diffusion model trained on Planck data recovers auto-spectra but degrades cross-component correlations, with their key metric $r_{\mathrm{tSZ}\times\mathrm{CIB}}$ reaching 0.91 for ST synthesis versus 0.71 for DDPM (their Table 3).**

- *Reference(s):* Prabhu et al. (2025)
- *Justification:* Prabhu et al. (2025) train a DDPM on Agora simulations of correlated CIB and tSZ at 150 GHz, not on Planck data, and they do not discuss or extend scattering-transform (ST) methods. They report that the DDPM reproduces both auto- and cross-power spectra within sample variance (see §4.3, Fig. 4), rather than degrading cross-component correlations. There is no Table 3 or metric $r_{\mathrm{tSZ}\times\mathrm{CIB}}$ with values 0.91 (ST) vs 0.71 (DDPM). Hence the statement is not supported by Prabhu et al. (2025).

- **✖ The ScatCov operator used here follows Prabhu et al. (2025, Section 2.3) by computing cross-coefficient statistics between second-order scattering moments, enabling phase-sensitive correlation tracking across components beyond what single-channel scattering coefficients provide.**
- *Reference(s):* Prabhu et al. (2025, Section 2.3)
- *Justification:* Prabhu et al. (2025, Section 2.3) does not describe a ScatCov operator or scattering transforms. The paper focuses on DDPMs and evaluates power spectra, pixel histograms, Minkowski functionals, and bispectrum/trispectrum (see Sections 3.1–3.2). There is no Section 2.3 detailing cross-coefficient statistics between second-order scattering moments or phase-sensitive correlation tracking. Therefore, the statement is not supported by the attached paper.
- **✖ Following Allys et al. (2020, Eq. 3), the ScatCov synthesis loss weights each coefficient by the inverse of its reference amplitude, $w_q = 1/|\mathrm{ScatCov}q(x)|$, to equalise the contributions of S1–S4 terms during LBFGS optimisation.}**
- *Reference(s):* Allys et al. (2020, Eq. 3)
- *Justification:* Allys et al. (2020, Eq. 3) defines the covariance of wavelet transforms $C_{\xi_1,\xi_2}(\tau)$ and does not discuss any synthesis loss weighting scheme. In the synthesis section, the loss $L = d(\Phi(\rho), \Phi(\tilde{\rho}))$ is referenced to external works for its exact form, with no mention of weights $w_q = 1/|\mathrm{ScatCov}q(x)|$ or equalising S1–S4 terms during L-BFGS optimisation. Therefore, the stated weighting is not supported by Allys et al. (2020, Eq. 3).}
- **✖ Spectral-matched Gaussian initialisation, in which the starting field is drawn from a Gaussian random field with angular power spectrum C_ℓ matched to the reference patch, yields 5–10× faster convergence of scattering-transform synthesis compared to white-noise initialisation, consistent with the benchmarks reported by Allys et al. (2020, Table 1).**
- *Reference(s):* Allys et al. (2020, Table 1)
- *Justification:* Allys et al. (2020, Table 1) lists fiducial cosmological parameter values, not synthesis benchmarks. The paper does not compare spectral-matched Gaussian vs white-noise initializations or report 5–10× convergence speedups for scattering-transform (or WPH) synthesis. Their syntheses start from Gaussian white noise (Sec. IV.B), and no convergence-time comparisons are provided.
- **✖ The joint Cholesky C_ℓ-matching plus paired pixel-histogram matching procedure used here follows the standard combination employed in previous synthesis pipelines that aim to recover both 1-point and 2-point structure, as in Allys et al. (2020) and Cheng & Ménard (2020), but generalises it to multi-component ($N \times N$) cross-spectral covariance matrices.**
- *Reference(s):* Allys et al. (2020), Cheng & Ménard 2020
- *Justification:* Allys et al. (2020) performs statistical syntheses via microcanonical maximum-entropy matching of Wavelet Phase Harmonics; it does not use a Cholesky C_ℓ-matching or paired pixel-histogram procedure, nor discuss generalizing such a method to $N \times N$ cross-

spectral covariances. The attached Cheng & Ménard 2020 document (about GRB synchrotron polarization) is unrelated and contains no such synthesis pipeline. Thus the claimed method and its generalization are not supported by the attached papers.

- ⚠ **Prior microcanonical SC-matching work (Allys et al. 2020; Mousset et al. 2024) showed that maximum-entropy synthesis constrained only by scattering (or wavelet phase harmonic) statistics can recover power spectra, pixel PDFs, and Minkowski functionals to within sampling noise without any explicit post-processing, in contrast to the explicit calibration route taken in this paper.**

- *Reference(s):* Allys et al. (2020), Mousset et al. (2024)
- *Justification:* Allys et al. (2020) used a microcanonical maximum-entropy synthesis constrained by WPH statistics (plus a small set of added low-pass scaling moments) and showed strong quantitative agreement with target fields: mean power spectrum within $\sim$5%, its standard deviation and correlations within $\sim$10%, pixel PDFs closely matched over several orders of magnitude, and Minkowski functionals within $\leq$0.5%, without any post-processing. Mousset et al. (2024) extended scattering-based generative models to the sphere and likewise showed good reproduction of PDFs, power spectra and Minkowski functionals from SC constraints alone, though they note small residual oscillations in power spectra and do not claim agreement "within sampling noise." Thus, the claim of recovery "to within sampling noise" solely from scattering/WPH constraints is stronger than what is explicitly demonstrated, making the statement only partially supported.

- ⚠ **Mousset et al. (2024) and Allys et al. (2020) emphasise that microcanonical SC-matching synthesis is genuinely unsupervised—requiring only a single target field or ensemble and no paired simulation/truth labels—whereas the DDPM of Prabhu et al. (2025) is supervised and inherits any bias in its training simulations, and the calibration recipe presented here is semi-supervised because it depends on a fiducial truth ensemble for the C_ℓ matrix and 1-point CDF; consequently, on real-sky data the SC-matching route is the only one of the three that cannot bake simulation bias into the generator by construction.**

- *Reference(s):* Mousset et al. (2024), Allys et al. (2020), Prabhu et al. (2025)
- *Justification:* Supported: Mousset et al. (2024) explicitly build microcanonical SC-based generative models from a single target full-sky map and emphasise no need for large training datasets ("our method holds in the limit case of a single data realisation"; Sect. 3.1, Conclusions). Allys et al. (2020) use a microcanonical maximum-entropy model constrained by WPH statistics estimated from a small ensemble and note it could be trained directly on observational data (Sec. IV, Conclusions), i.e., no paired labels. Supported in part: Prabhu et al. (2025) train a DDPM on Agora simulations and explicitly caution that any generative model's fidelity depends on, and can inherit biases from, the training simulations (Sec. 5.2). Not supported: the claim that the DDPM is "supervised" is not stated—DDPM training here is generative (no labels). Also, no "semi-supervised calibration recipe" depending on a fiducial C_ℓ matrix and 1-point CDF is presented; the only post-hoc adjustment is a global variance rescaling (Sec. 3.2). Therefore the collective statement is only partially supported.

Mathematical consistency audit

This section audits **symbolic/analytic** mathematical consistency (algebra, derivations, dimensional/unit checks, definition consistency).

Maths relevance: substantial

The paper contains a moderate set of central mathematical constructions: (i) definitions of scattering/ScatCov coefficients and a weighted L_2 synthesis loss; (ii) a per-ℓ-bin Fourier-domain rescaling ('soft C_ℓ match'); (iii) a multi-channel per-ℓ-bin covariance whitening/recolouring transform (Eq. 12) claimed to enforce all auto- and cross-spectra; and (iv) several analytic interpretations (e.g., $-1/\sqrt{2}$ residual-correlation floor). The core linear-algebra of Eq. (12) is consistent, but several 'by construction' and 'phase-preserving' claims are overstated or depend on missing convergence/ordering details for the alternating histogram/spectrum projections.

Checked items

1. ✔ **Scattering coefficient definitions** (Eq. (1), Sec. 3.1, p.3)
    - **Claim:** Defines first and second scattering-like moments using a wavelet $\psi_{j\ell}$: $S_{1,j} = \langle |x \star \psi_{j\ell}| \rangle_r$ and $S_{2,j} = \langle |x \star \psi_{j\ell}|^2 \rangle_r$.
    - **Checks:** notation consistency, definition consistency
    - **Verdict:** PASS; confidence: medium; impact: minor
    - **Assumptions/inputs:** $x(r)$ is an image/field; $\psi_{j\ell}$ is a wavelet at scale j and orientation ℓ, $\langle\cdot\rangle_r$ denotes spatial averaging over r.
    - **Notes:** Formulas are syntactically consistent, but the index notation suppresses ℓ (orientation) in $S_{1,j}$ and $S_{2,j}$ despite explicit dependence via $\psi_{j\ell}$. This is a clarity/notation issue rather than an algebraic error.
2. ⚠ **Weighted ScatCov loss well-posedness** (Eq. (2), Sec. 3.2, p.3)
    - **Claim:** Uses an inverse-amplitude weighted squared error: $L_{\mathrm{SC}}(s)=\sum_q w_q (\mathrm{ScatCov}q(s)-\mathrm{ScatCov}_q(x))^2 with w_q=1/|\mathrm{ScatCov}q(x)|$.}
    - **Checks:** dimensional/units sanity, well-posedness (division by zero), definition consistency
    - **Verdict:** UNCERTAIN; confidence: medium; impact: moderate
    - **Assumptions/inputs:** All reference coefficients $\mathrm{ScatCov}q(x)$ are nonzero or are regularised in implementation.}
    - **Notes:** As written, w_q can diverge when $\mathrm{ScatCov}q(x)=0 or be extremely large for tiny coefficients; the paper does not specify an \epsilon$-floor or masking. This omission blocks a strict analytic claim that Eq. (2) defines a finite objective for all references.}
3. ⚠ **Spectral-matched initialisation** (Eq. (3), Sec. 3.2, p.3)
    - **Claim:** Initial field s_0 is generated by shaping Gaussian noise in Fourier space: $s_0 = \mathcal{F}^{-1}(\mathcal{F}(\epsilon) \odot \sqrt{P_{\mathrm{ref}}(\ell)})$.
    - **Checks:** algebra sanity, definition sufficiency
    - **Verdict:** UNCERTAIN; confidence: low; impact: minor

- **Assumptions/inputs:** $P_{\rm ref}(\ell)$ is the desired (2D) Fourier power as a function of ℓ, FFT conventions and the mapping between ℓ and discrete $\boldsymbol{k}$ are fixed.
- **Notes:** The structure 'multiply Fourier modes by $\sqrt{P}$' is correct in principle, but $P_{\rm ref}(\ell)$ is only specified up to proportionality ($P_{\rm ref}(\ell) \propto \ell(\ell+1)C_\ell/2\pi$) and the FFT/$C_\ell$ normalisation mapping is not stated, so an exact symbolic verification is not possible from the PDF alone.

4. ✔ **Joint loss decomposition** (Eq. (4), Sec. 3.3, p.3)
   - **Claim:** Defines $L_{\rm joint}$ as sum of per-channel ScatCov losses plus cross-correlation and sign penalties.
   - **Checks:** definition consistency
   - **Verdict:** PASS; confidence: high; impact: minor
   - **Assumptions/inputs:** $L_{\rm SC}$ is defined as in Eq. (2), $\lambda_{\rm cross}$, $\lambda_{\rm sign}$ are nonnegative.
   - **Notes:** No algebra to check beyond consistent composition.

5. ✖ **Correlation coefficient definition vs 'demeaned Pearson'** (Eq. (5) and text, Sec. 3.3, p.3)
   - **Claim:** Defines $r(a,b) = \langle ab\rangle / \sqrt{\langle a^2\rangle\langle b^2\rangle}$ and calls it the 'demeaned Pearson coefficient' appropriate for zero-mean fields.
   - **Checks:** definition consistency, notation consistency
   - **Verdict:** FAIL; confidence: high; impact: moderate
   - **Assumptions/inputs:** Either a, b are implicitly demeaned or fields are assumed zero-mean.
   - **Notes:** Eq. (5) omits mean subtraction. It equals Pearson correlation only if $\langle a\rangle = \langle b\rangle = 0$ (or if a, b are implicitly demeaned, which is not stated in the equation). Later the paper reports nonzero means (e.g., Table 4), making the 'demeaned Pearson' description inconsistent unless an unstated preprocessing step is applied.

6. ✔ **Sign penalty definition** (Eq. (6), Sec. 3.3, p.3)
   - **Claim:** Penalises positive tSZ excursions at 150 GHz using $L_{\rm sign} = ||\,\mathbf{max}(s_{\rm tSZ}, 0)||^2$.
   - **Checks:** definition consistency
   - **Verdict:** PASS; confidence: medium; impact: minor
   - **Assumptions/inputs:** **max** is applied elementwise in pixel space, $||\cdot||$ denotes an L_2 norm over pixels.
   - **Notes:** Mathematically consistent as a nonnegative penalty; the PDF rendering shows a malformed norm-squared symbol, but the intent is clear.

7. ✔ **Multi-channel operator mapping** (Eq. (7), Sec. 3.4, p.3)
   - **Claim:** Defines a multi-channel ScatCov operator $\mathbf{\Phi}^{(2)} : \mathbb{R}^{2\times H\times W} \to \mathbb{R}^D$.
   - **Checks:** type consistency
   - **Verdict:** PASS; confidence: high; impact: minor
   - **Assumptions/inputs:** $\mathbf{\Phi}^{(2)}$ produces a flattened coefficient vector including cross-channel terms.
   - **Notes:** Dimensional mapping is consistent with the described construction.

8. ✔ **Soft C_ℓ rescale amplitude factor** (Eq. (8), Sec. 3.5, p.4)
    - **Claim:** Defines $\alpha_b = \sqrt{\langle|\hat{t}k/^2|rangle}/\langle|\hat{g}k/^2|rangle_k.}} and applies $\hat{g}'_k=\alpha_b(k) \hat{g
    - **Checks:** algebra correctness, normalisation logic
    - **Verdict:** PASS; confidence: high; impact: moderate
    - **Assumptions/inputs:** Bins b partition Fourier modes, denominators are nonzero (or regularised).
    - **Notes:** This scaling enforces equality of average $|\hat{g}'|^2$ to average $|\hat{t}|^2$ within each bin by construction.
9. ✖ **Soft C_ℓ rescale 'spatial properties preserved' claim** (Text after Eq. (8), Sec. 3.5, p.4)
    - **Claim:** Because Fourier phases are preserved exactly, 'every spatial coherence property' (including extreme-pixel positions, edges, Minkowski M_1/M_2) is preserved by the rescale.
    - **Checks:** logical implication check
    - **Verdict:** FAIL; confidence: high; impact: moderate
    - **Assumptions/inputs:** Phase preservation is taken to imply invariance of spatial/topological features.
    - **Notes:** Preserving Fourier phases does not, in general, preserve extrema locations, edges, or Minkowski functionals when amplitudes are changed as a function of ℓ (this is a nontrivial filtering operation). The statement should be weakened/qualified.
10. ✔ **Cosine schedule definition** (Eq. (9), Sec. 3.6, p.4)
    - **Claim:** Defines $\bar{\alpha}_t$ via a cosine schedule $\bar{\alpha}_t = f(t)^2/f(0)^2$ with $f(t) = \cos\left(\frac{(t/T)+s}{1+s} \cdot \frac{\pi}{2}\right)$.
    - **Checks:** algebra sanity, range sanity
    - **Verdict:** PASS; confidence: medium; impact: minor
    - **Assumptions/inputs:** $t \in [0, T]$.
    - **Notes:** The formula is self-consistent and yields $\bar{\alpha}_0 = 1$. No internal inconsistency found.
11. ✔ **DDPM reverse sampling step and coefficient discussion** (Eq. (10) and §3.6.1, p.4)
    - **Claim:** Reverse update uses coefficient $\beta_t/\sqrt{1-\bar{\alpha}t}on|epsilon\theta;priorincorrectuseof\sqrt{\beta_t/(1-\bar{\alpha}_t)} causesanerrorfactor1/\sqrt{\beta_t}$.
    - **Checks:** algebra between shown steps
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** $\sigma_t = \sqrt{\beta_t}$ as stated.
    - **Notes:** The ratio of incorrect to correct coefficient is $(\sqrt{\beta_t})/(\beta_t) = 1/\sqrt{\beta_t}$, matching the text.
12. ✔ **DDPM training objective** (Eq. (11), Sec. 3.6.2, p.4)
    - **Claim:** Noise-prediction MSE objective: $\mathbb{E}||\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon, t)||^2$.
    - **Checks:** notation consistency

- **Verdict:** PASS; confidence: high; impact: minor
- **Assumptions/inputs:** ϵ is standard normal noise.
- **Notes:** Self-consistent as a denoising objective; no internal algebra to verify.

13. ✔ **Cross-spectral covariance definition** (§4.1.1, p.5)

    - **Claim:** Defines per-bin 2×2 covariance $C_{ab}(\ell)=\langle \mathrm{Re}(\hat{F}_a \hat{F}_b^*)\rangle$.
    - **Checks:** definition consistency
    - **Verdict:** PASS; confidence: medium; impact: minor
    - **Assumptions/inputs:** All maps are real-valued so Fourier coefficients satisfy Hermitian symmetry, the statistic of interest is the real cross-spectrum.
    - **Notes:** Using $\mathbf{Re}(\cdot)$ yields a real-symmetric covariance consistent with later use of real matrix square roots. The paper's subsequent 'matches C_{ab}' claims are with respect to this definition.

14. ✔ **Joint Cholesky C_ℓ-match enforces target covariance** (Eq. (12), Sec. 4.1.1, p.5)

    - **Claim:** Applying $\hat{F}_{\rm new}(k)=C(\ell)^{1/2}C_{\rm gen}(\ell)^{-1/2}\, \hat{F}_{\rm gen}(k)$ enforces matching of C per bin.}, $C_{\mathrm{CIB},\ell}$, and $C_{\mathrm{tSZ} \times \mathrm{CIB},\ell}$
    - **Checks:** linear algebra consistency, constraint/normalisation check
    - **Verdict:** PASS; confidence: high; impact: critical
    - **Assumptions/inputs:** $C_{\mathrm{gen}}(\ell)$ is invertible (SPD) and the same sample covariance used for whitening, matrix square roots satisfy $C^{1/2}C^{1/2} = C$ (or, for Cholesky, $C^{1/2}(C^{1/2})^T = C$), the transform is applied uniformly to all modes in the bin.
    - **Notes:** For the paper's real-part covariance, $C' = AC_{\mathrm{gen}}A^T$ with $A = C_{\mathrm{ref}}^{1/2}C_{\mathrm{gen}}^{-1/2}$ gives $C' = C_{\mathrm{ref}}$ exactly. This is the main mathematically solid 'by construction' part (for 2-point spectra).

15. ✖ **Eq. (12) described as 'phase-preserving'** (Abstract; §4.1.1, p.5; §5.2, p.11)

    - **Claim:** The joint $N \times N$ Cholesky C_ℓ-match is described as phase-preserving and preserving generator phase information.
    - **Checks:** logical implication check, property verification
    - **Verdict:** FAIL; confidence: high; impact: moderate
    - **Assumptions/inputs:** 'Phase-preserving' is intended per-channel in Fourier space.
    - **Notes:** For $N > 1$, $\hat{F}_{\rm new}=A\, \hat{F}$ with a generally non-diagonal real matrix A changes the complex argument of each channel's Fourier coefficient in general. Only the N=1 (or diagonal A) case preserves per-channel Fourier phases. A weaker true statement is that it does not mix different k-modes (no convolution across k), only mixes channels at fixed k$.

16. ⚠ **Pixel cross-correlation follows from matching spectra** (§4.1.1 and §4.2, pp.5–6)

    - **Claim:** Matching per-bin auto- and cross-spectra implies the pixel-level correlation $r_{\mathrm{tSZ}\times\mathrm{CIB}}$ matches truth (as a band integral).
    - **Checks:** definition consistency, missing-assumption identification

- **Verdict:** UNCERTAIN; confidence: medium; impact: moderate
- **Assumptions/inputs:** Fields are mean-zero or r is computed on demeaned fields, pixel covariance equals an integral/sum over the cross-spectrum across the matched band and windowing is consistent.
- **Notes:** The implication holds cleanly for demeaned fields (covariance determined by cross-spectrum) and if the same spectral weighting/windowing is used. Because Eq. (5) is not explicitly demeaned and later means are nonzero, the connection between matched cross-spectrum and reported 'Pearson r' is not fully verifiable from the PDF alone.

17. ✔ **Histogram match implies matching all 1-point statistics** (§4.1.1, p.5)
    - **Claim:** Rank-preserving paired histogram match forces gen pixel CDF to equal truth CDF, thus matching all 1-point statistics (mean, variance, skewness, kurtosis, extrema).
    - **Checks:** probability/CDF logic
    - **Verdict:** PASS; confidence: high; impact: moderate
    - **Assumptions/inputs:** Histogram match is performed on the full pixel set per channel, without ties/pathologies.
    - **Notes:** If the final output has undergone the rank-replacement step, its empirical CDF equals the target empirical CDF exactly, hence all empirical 1-point moments and order statistics match exactly.

18. ⚠ **Alternating histogram match and Cholesky gives simultaneous exact 1- and 2-point matching** (§4.1.1, p.5; 'by construction' claims in Abstract and §5.6)
    - **Claim:** Alternating histogram matching and Eq. (12) 'locks' both the full pixel CDF and the per-bin spectra/cross-spectra simultaneously.
    - **Checks:** logical completeness, missing derivation/proof identification
    - **Verdict:** UNCERTAIN; confidence: high; impact: critical
    - **Assumptions/inputs:** The alternating-projection procedure converges to a fixed point satisfying both constraints, the final-step ordering is specified.
    - **Notes:** Each operation generally breaks the constraint enforced by the other (histogram match changes spectra; Cholesky recolouring changes the histogram). Without (i) a specified final operator order and (ii) a convergence/fixed-point argument, the claim 'exact by construction' for both simultaneously cannot be verified analytically from the PDF.

19. ✔ **Residual-correlation algebraic floor** (§4.2.1, p.6)
    - **Claim:** If gen is independent of truth with equal variance, then $\mathbf{corr}(\mathbf{gen}-\mathbf{truth},\mathbf{truth}) = -1/\sqrt{2}$.
    - **Checks:** algebra between shown steps
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** $\mathbf{corr}(\mathbf{gen},\mathbf{truth}) = 0$ and $\mathbf{Var}(\mathbf{gen}) = \mathbf{Var}(\mathbf{truth})$.
    - **Notes:** Correct: $\mathbf{Cov}(\mathbf{res},\mathbf{truth}) = -\mathbf{Var}(\mathbf{truth}),\ \mathbf{Var}(\mathbf{res}) = 2\mathbf{Var}(\mathbf{truth}) \Rightarrow \mathbf{corr} = -1/\sqrt{2}$.

20. ✔ **Minkowski functional definitions and M_0 matching argument** (§4.4, p.9)

- **Claim:** Defines $M_0(u) = \langle \mathbb{I}[x > u] \rangle$, $M_1(u) = \langle |\nabla x| \mathbb{I}[x > u] \rangle$, $M_2(u) = \langle \chi \rangle$, and argues M_0 overlays truth exactly after histogram match.
- **Checks:** definition consistency, logical implication check
- **Verdict:** PASS; confidence: medium; impact: minor
- **Assumptions/inputs:** Threshold u is in absolute units (or fixed in σ units) and the histogram match is to the truth distribution in those units.
- **Notes:** Given exact CDF matching to truth in the same units, the area fraction above any threshold u matches truth, so $M_0(u)$ matches. (This does not imply M_1/M_2 match, consistent with the paper's discussion.)

21. ✔ **Triple loss function with sign constraint** (Eq. (13), Sec. 5.3, p.12)

    - **Claim:** Extends joint loss to three components with two correlation penalties and a positivity-of-mean penalty for CIB: $w_{\text{sign}} \cdot \mathbf{max}(-\bar{x}_{\text{CIB}}, 0)^2$.
    - **Checks:** definition consistency
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** $\bar{x}_{\text{CIB}}$ is the patch mean; **max** applies to a scalar.
    - **Notes:** Penalty is mathematically consistent and enforces nonnegative mean when weighted strongly.

22. ✔ $N \times N$ **generalisation of Eq. (12)** (§5.3 and §5.4, pp.12–13)

    - **Claim:** Eq. (12) extends to $\mathbf{3 \times 3}$ and $\mathbf{4 \times 4}$ by using $C_{\text{ref}}(\ell)^{1/2} C_{\text{gen}}(\ell)^{-1/2}$ with C matrices of matching dimension.
    - **Checks:** linear algebra generalisation check
    - **Verdict:** PASS; confidence: high; impact: moderate
    - **Assumptions/inputs:** $C_{\text{gen}}(\ell)$ is SPD/invertible for each bin, a valid matrix square root is used.
    - **Notes:** The covariance-matching argument generalises directly to any N given invertibility and a consistent square-root definition.

Limitations

- Audit is restricted to the provided PDF text/images; no code or supplementary material was consulted.
- Several key claims are empirical ('numerical precision', 'iteration converges in 6 cycles'). This audit cannot validate those numerically and only assesses whether they are guaranteed by the stated mathematics.
- Some formulas (e.g., FFT/C_ℓ normalisations, ScatCov coefficient definitions, handling of zero coefficients) are underspecified in the PDF, preventing full analytic verification.

Numerical results audit

This section audits **numerical/empirical** consistency: reported metrics, experimental design, baseline comparisons, statistical evidence, leakage risks, and reproducibility.

Out of 23 candidate numeric checks, 15 passed, 2 failed, and 6 were uncertain (mostly due to missing tabulated inputs for the intended recomputations). The two failures concern (i) a unit/pixel-scale consistency check for patch size vs pixel resolution, and (ii) a rounding/precision claim about matching to

four decimal places for a median ratio example. Other recomputed percentages/ratios (held-out recovery percentages, skew/kurtosis ratios, algebraic constant $-1/\sqrt{2}$, and various percent recoveries) were numerically consistent within stated tolerances.

Checked items

1. ✔ **C01_train_aug_count** (p.3, end of §2 DATA)
    - **Claim:** 8-fold augmentation yielding 1600 training pairs from $N_{\text{train}} = 200$ paired patches.
    - **Checks:** multiplication
    - **Verdict:** PASS
    - **Notes:** Exact integer arithmetic.
2. ✖ **C02_patch_pixels_area_consistency** (p.2, §2 DATA)
    - **Claim:** $5^\circ \times 5^\circ$ patches at 256×256 pixels correspond to 5 arcmin pixel scale.
    - **Checks:** unit_consistency
    - **Verdict:** FAIL
    - **Notes:** $60 \times 5/256 = 1.1719$ arcmin; inconsistent with 5 arcmin unless additional downsampling/definition. Flag for check.
3. ✔ **C03_patch8_percentile_from_1523** (p.3, §2 DATA (Patch 8 description))
    - **Claim:** Patch 8 sits in the top 3.8% on deepest decrement and the top 10.8% on σ (in $N_{\text{patch}} = 1523$).
    - **Checks:** percentage_to_count
    - **Verdict:** PASS
    - **Notes:** $0.038 \times 1523 \approx 58$, $0.108 \times 1523 \approx 165$.
4. ✔ **C04_bandavg_from_bins_ST_tSZ** (p.6, Table 2)
    - **Claim:** Band-average [500,6000] for ST+jm tSZ is 1.088 ± 0.040, given per-band values 1.128, 1.115, 1.062 over contiguous bins.
    - **Checks:** weighted_average_by_interval
    - **Verdict:** PASS
    - **Notes:** Width-weighted average across [500,1500],[1500,3000],[3000,6000] gives 1.08845, consistent with 1.088 (noting averaging-definition caveat).
5. ⚠ **C05_bandavg_from_bins_ST_CIB** (p.6, Table 2)
    - **Claim:** Band-average [500,6000] for ST+jm CIB is 1.010 ± 0.014, given per-band values 0.996, 1.022, 1.009.
    - **Checks:** weighted_average_by_interval
    - **Verdict:** UNCERTAIN
    - **Notes:** Missing bin ranges/ratios or reported total for weighted average.
6. ⚠ **C06_bandavg_from_bins_DDPM_tSZ** (p.6, Table 2)
    - **Claim:** Band-average [500,6000] for DDPM+jm tSZ is 1.065 ± 0.123, given per-band values 1.141, 1.079, 1.034.
    - **Checks:** weighted_average_by_interval

- **Verdict:** UNCERTAIN
- **Notes:** Missing bin ranges/ratios or reported total for weighted average.

7. ⚠ **C07_bandavg_from_bins_DDPM_CIB** (p.6, Table 2)
   - **Claim:** Band-average [500,6000] for DDPM+jm CIB is $\mathbf{1.013 \pm 0.011}$, given per-band values 1.027, 1.013, 1.008.
   - **Checks:** weighted_average_by_interval
   - **Verdict:** UNCERTAIN
   - **Notes:** Missing bin ranges/ratios or reported total for weighted average.
8. ✔ **C08_skew_kurt_ratio_claims** (p.8, §4.3 PDF Statistics)
   - **Claim:** For tSZ, skewness ratio is 0.98 (ST) and 0.98 (DDPM); excess-kurtosis ratio is 0.96 (ST) and 0.94 (DDPM).
   - **Checks:** ratio_recompute
   - **Verdict:** PASS
   - **Notes:** Recomputed ratios: skew ST/truth=0.9803, DDPM/truth=0.9777; kurt ST/truth=0.9599, DDPM/truth=0.9393.
9. ✔ **C09_min_core_recovery_percent** (p.8, §4.3 PDF Statistics)
   - **Claim:** Deepest cluster core is recovered to within 2–10%: truth min $\mathbf{-350}$ μKCMB, ST $\mathbf{-342}$, DDPM $\mathbf{-318}$.
   - **Checks:** relative_error
   - **Verdict:** PASS
   - **Notes:** Relative errors: ST 0.02286 (2.29%), DDPM 0.09143 (9.14%), within 2–10%.
10. ✔ **C10_crossr_recovery_heldout_ST** (p.7, held-out evaluation paragraph + p.11/p.12 held-out bullets + Table 7)
    - **Claim:** Held-out pixel $r_{\mathrm{tSZ}\times\mathrm{CIB}}$: gen $\mathbf{-0.159}$ versus test-truth $\mathbf{-0.172}$ corresponds to 92% recovery.
    - **Checks:** percentage_recompute
    - **Verdict:** PASS
    - **Notes:** Recomputed recovery is 92.44%, consistent with 92% after rounding.
11. ✔ **C11_crossr_recovery_heldout_DDPM** (p.12, DDPM held-out protocol bullets + Table 7)
    - **Claim:** DDPM held-out: $r_{\mathrm{gen}} = \mathbf{-0.158}$ vs test-truth $\mathbf{-0.173}$ corresponds to 91.7% recovery.
    - **Checks:** percentage_recompute
    - **Verdict:** PASS
    - **Notes:** Recomputed recovery is 91.33%, consistent within 0.5 percentage points.
12. ✔ **C12_core_depth_recovery_ST** (p.11, held-out bullets + Table 7)
    - **Claim:** Held-out ST deepest core: gen min $\mathbf{-220}$ μKCMB vs test min $\mathbf{-337}$ corresponds to $\sim$65% recovery.
    - **Checks:** percentage_recompute

- **Verdict:** PASS
- **Notes:** Recomputed recovery is 65.28%, consistent with "∼65%".

13. ✔ **C13_core_depth_recovery_DDPM** (p.12, DDPM held-out bullets + Table 7)
    - **Claim:** DDPM held-out deepest core: gen min -330 μKCMB vs test -345 corresponds to 96% recovery.
    - **Checks:** percentage_recompute
    - **Verdict:** PASS
    - **Notes:** Recomputed recovery is 95.65%, consistent with 96% after rounding.

14. ✔ **C14_algebraic_floor_minus_1_over_sqrt2** (p.6-7, §4.2.1 (residual diagnostic) + Fig. 6 text)
    - **Claim:** Algebraic limit $r = -1/\sqrt{2} \approx -0.707$ and observed -0.706 are consistent.
    - **Checks:** constant_recompute
    - **Verdict:** PASS
    - **Notes:** $-1/\sqrt{2} = -0.70710678$; observed -0.706 differs by 0.00111, within abs tolerance.

15. ✔ **C15_sampling_coeff_error_factor** (p.4, §3.6.1 Critical sampling coefficient)
    - **Claim:** Coefficient error factor is $1/\sqrt{\beta_t} \approx 10$ for typical $\beta_t \sim 0.01$.
    - **Checks:** order_of_magnitude_recompute
    - **Verdict:** PASS
    - **Notes:** $1/\sqrt{0.01} = 10$ exactly.

16. ✔ **C16_corrunet_param_count_check** (p.4, §3.6.2 Architecture)
    - **Claim:** Model has 554,562 parameters.
    - **Checks:** self_consistency_integer_format
    - **Verdict:** PASS
    - **Notes:** Comma-formatted integer parses consistently to 554562.

17. ✖ **C17_multi_freq_median_ratio_delta** (p.13, §5.4 (text))
    - **Claim:** med(CIB217/CIB150) = 2.9354 (truth) versus 2.9370 (gen+JM), and claim of matching to four decimal places.
    - **Checks:** rounding_consistency
    - **Verdict:** FAIL
    - **Notes:** Rounded to 4 d.p., values are 2.9354 vs 2.9370 (not equal).

18. ✔ **C18_fig12_residual_dp_claim** (p.13-15, §5.4 text + Fig. 12(b) caption)
    - **Claim:** Body: cross-correlations match to four decimal places; Fig. 12(b): residual $|\Delta r|$ is 10^{-10} to 10^{-11}, orders of magnitude below 10^{-4}.
    - **Checks:** order_of_magnitude_comparison
    - **Verdict:** PASS
    - **Notes:** 10^{-10} and 10^{-11} are 10^{-6} and 10^{-7} of 10^{-4}, respectively.

19. ✔ **C19_crossr_ladder_percentages** (p.16-18, §6 + Fig. 15)

- **Claim:** Cross-r recovery ladder: truth $r = -0.163$; multi-channel gives $r = -0.086$ (53%); +soft gives $r = -0.093$ (57%).
- **Checks:** percentage_recompute
- **Verdict:** PASS
- **Notes:** Recomputed: 52.76% (≈ **53%**), 57.06% (≈ **57%**).

20. ✔ **C20_iteration_sweep_percentages** (p.16-17, §6.1 + Fig. 14 / Fig. 15 text)
    - **Claim:** Iteration sweep recovery: 32%/38%/47%/53%/56% at 200/400/800/1600/3200 steps.
    - **Checks:** monotonicity_and_bounds
    - **Verdict:** PASS
    - **Notes:** Series is within [0,100] and non-decreasing.
21. ✔ **C21_break_even_samples_calc** (p.19, §6.1.0.6)
    - **Claim:** Break-even $N_{\text{samples}} \approx 1$ h / 80 s ≈ **50** samples.
    - **Checks:** division_recompute
    - **Verdict:** PASS
    - **Notes:** $3600/80 = 45$; within loose tolerance given approximate inputs.
22. ⚠ **C22_bp_error_reduction_factors** (p.22, §7.2 (quantitative improvements))
    - **Claim:** Mean |relative error| drops from 28.6%→4.8% for $S_3(\ell)$ and 64.9%→9.2% for $K_4(\ell)$.
    - **Checks:** ratio_recompute
    - **Verdict:** UNCERTAIN
    - **Notes:** Unsupported ratio_recompute variant.
23. ⚠ **C23_peak_count_deficits_percent** (p.25, Table 17)
    - **Claim:** Peak count deficits: e.g., for $\nu = 2$ truth=710 and ST=617 corresponds to −13.1%, etc. Verify all percent deficits shown in parentheses.
    - **Checks:** percentage_recompute
    - **Verdict:** UNCERTAIN
    - **Notes:** Missing inputs for core depth recovery recompute.

Limitations

- Audit is restricted to the provided PDF text; no underlying map/spectrum data are available, so only arithmetic and simple logical/unit consistency checks are feasible.
- Figure-based numeric values not explicitly printed in the text/tables are not used (no plot digitization).
- Some 'band-averaged' quantities may use a definition (e.g., averaging over modes, log-bins, or patch-wise normalization) that differs from simple ℓ-interval width weighting; related candidates are flagged as plausibility checks rather than definitive inconsistencies.