Skepthical review: *Guided Super-Resolution Denoising of Thermal Sunyaev-Zel'dovich Maps using a Conditional Diffusion Model*

Summary

This manuscript presents a two-stage deep-learning pipeline to reconstruct high-resolution (≈ 1 arcmin) thermal Sunyaev–Zel'dovich (tSZ) Compton-y maps from simulated multi-frequency microwave observations (SO-like; Sec. 2.1). Stage 1 is a U-Net–style Super-Resolution Denoising Autoencoder (SR-DAE) with gated cross-attention that uses the high-frequency, CIB-dominated channels as spatial guidance; it is trained with a composite loss combining an L_1 image term, a power-spectrum consistency term, and a penalty discouraging correlation of residuals with CIB (Sec. 2.3). Stage 2 wraps this backbone into a Conditional Diffusion Model (CDM) to sample from the conditional distribution $p(\mathbf{tSZ} \mid \mathbf{observed})$, producing ensembles whose mean is used as a point estimate and whose variance is interpreted as pixel-level uncertainty (Sec. 2.4–2.5). Results on FLAMINGO-based simulations suggest improved small-scale structure recovery compared to cILC and Wiener filtering (Sec. 3.1), plus robustness tests (Sec. 3.2) and uncertainty calibration checks (Sec. 3.3.1). The overall direction—non-linear, conditional reconstruction with uncertainty quantification in the presence of non-Gaussian CIB—is timely and potentially impactful. However, several core elements are currently under-specified (data model realism, SR-DAE/CDM definitions, loss math, training details), and some internal inconsistencies (dataset accounting, null-test mean, inconsistent MSE reporting) plus limited quantitative summaries weaken confidence in the claims and make the work difficult to reproduce or fairly benchmark (Sec. 2.1–2.5, Sec. 3.1–3.3).

Strengths

- Compelling reframing of the problem: leveraging CIB-dominated high-frequency channels as spatial guidance for tSZ super-resolution/denoising rather than treating them only as contaminants (Introduction; Sec. 2.1–2.3).
- Physically motivated objective design: combining pixel-domain fidelity with spectral-domain constraints and an explicit residual–CIB decorrelation penalty targets known failure modes of linear component-separation methods (Sec. 2.3).
- Two-stage design is conceptually coherent: a deterministic SR-DAE for high-fidelity reconstructions and a conditional diffusion model to represent uncertainty and sample diversity (Sec. 2.4–2.5).
- Validation breadth is strong in spirit: comparisons to cILC/WF, robustness checks (OOD massive clusters, high-noise, null tests), interpretability via Integrated Gradients, and uncertainty calibration via PIT (Sec. 3.1–3.3).

- Use of scientifically relevant diagnostics (transfer function, residual power spectra, reconstruction gain, and a scaling-relation-style test) demonstrates intent to connect ML performance to downstream astrophysical utility (Sec. 2.5; Sec. 3.3.2–3.3.4).
- Several figures (notably Fig. 1 and parts of Fig. 3) are effective at visually contrasting linear vs learned reconstructions and at illustrating robustness/interpretability claims.

Major issues

1. **Simulation/observation model is not specified at the level needed to assess realism, generalization, or reproducibility (Sec. 2.1–2.2).** Key ambiguities include: the instrument model for each frequency (e.g., 545/857 GHz are Planck/HFI-like rather than Simons Observatory), beams/bandpasses/noise levels per band, pixel size and patch area, whether noise is white/correlated and homogeneous/inhomogeneous, and which sky components are included beyond tSZ+CIB+instrumental noise (e.g., primary CMB, kSZ, radio sources, Galactic dust). As written, it is hard to judge whether the task reflects real component-separation difficulty or is simplified in ways that could overstate performance.

   Recommendation: Expand Sec. 2.1–2.2 with a concrete data-model table: for each channel list central frequency, bandpass assumption, beam (FWHM or full beam transfer), pixelization, and noise level/model. Explicitly state whether 545/857 GHz are assumed as external Planck-like priors in an SO-era analysis, and how they are beam-matched and calibrated. Enumerate all simulated components (tSZ, CIB, CMB, kSZ, radio/IR point sources, Galactic dust/synchrotron, etc.), and if some are omitted, justify why and discuss implications. If feasible, add at least one ablation including the primary CMB (especially relevant at 90/150 GHz) to demonstrate robustness to a dominant real-sky contaminant.

2. **SR-DAE architecture and training procedure are under-specified, preventing reproduction and making it difficult to attribute gains to the proposed design choices (Sec. 2.1–2.3).** The text mentions a U-Net with gated cross-attention, but does not provide layer-wise details (depth, channels, kernels/strides), where attention is inserted, the precise gated cross-attention formulation, normalization/activation choices, patch size, and the full training recipe (optimizer, LR schedule, batch size, augmentations, regularization, early stopping).

   Recommendation: In Sec. 2.3 (or an Appendix), add a reproducibility-focused specification: (i) input/output tensor shapes (patch size, pixel resolution), preprocessing/standardization per channel, and any beam-matching; (ii) a layer-by-layer architecture table for the SR-DAE including where and how gated cross-attention is applied (equations or pseudocode); (iii) the full training recipe (optimizer, LR schedule, batch size, weight decay, augmentations, gradient clipping, early stopping/selection criterion). If space is limited, provide a public configuration file/model card and cite it.

3. **The composite loss is central but not mathematically defined in a reproducible way (Sec. 2.3; Eq. (1)).** In particular, L_{spec} and L_{corr} lack explicit formulas, normalization, binning/ℓ-range, masking/apodization/beam treatment, and how multi-channel CIB information enters. The weights λ_1 and λ_2 are not clearly stated/tuned. This blocks verification of what the model is actually optimizing and whether reported spectral fidelity is a direct consequence of the loss design.

   Recommendation: Extend Sec. 2.3 to explicitly define L_{spec} and L_{corr}: specify the power-spectrum estimator (flat-sky Fourier vs spherical harmonics; 2D vs binned 1D C_ℓ), binning and ℓ-range, any ℓ-dependent weighting (linear vs log), window/apodization/mask handling, and whether beam deconvolution is applied. For L_{corr}, define the residual r, which CIB map(s) are used, and whether correlation is computed in pixel space (e.g., Pearson r), Fourier space, or in radial bins around clusters; state whether maps are mean-subtracted and how channels are aggregated. Report λ_1, λ_2 and how selected (validation sweep/heuristic), and add a brief sensitivity/ablation showing the effect of removing each term on (i) residual–CIB correlation and (ii) small-scale power recovery.

4. **The Conditional Diffusion Model is described only at a high level, limiting assessment of the claimed posterior sampling and uncertainty quantification (Sec. 2.4–2.5).** Critical missing details include: forward noising equation and time horizon T, exact noise/variance schedule, whether the network predicts $\epsilon/x_0/v$, the explicit conditional training loss, how conditioning on the observed multi-frequency maps and diffusion time is injected (concatenation/FiLM/cross-attention), what is frozen vs fine-tuned from SR-DAE, and sampling settings beyond "50-step DDIM" and "10 samples."

   Recommendation: Augment Sec. 2.4 with a complete CDM specification: write down the forward process $q(y_t \mid y_0)$, the reverse parameterization, the schedule (β_t/α_t, T), the prediction target ($\epsilon/x_0/v$) and full conditional loss with consistent notation (e.g., y =tSZ, x =observed). Describe the conditioning pathway and time embeddings and whether SR-DAE weights are reused/frozen. For sampling, state DDIM parameters, number of steps, and initialization. Add an Appendix plot showing convergence of (i) PIT calibration and (ii) at least one reconstruction metric vs number of diffusion steps and vs number of samples.

5. **CDM quantitative behavior raises a major interpretability concern: the reported MSE of the CDM ensemble mean is dramatically worse than the deterministic SR-DAE (e.g., 9.5667 vs 1.339 in scaled units; Sec. 3.3), which is not straightforward to reconcile with the interpretation of the CDM mean as a posterior mean estimate.** This could indicate weak conditioning, a normalization/evaluation mismatch, too-few sampling steps, or that samples add high-frequency structure that is plausible but uncorrelated with the specific truth realization—undermining claims about point-estimate fidelity.

Recommendation: In Sec. 3.3 (and/or Appendix), provide a targeted diagnosis: (i) report MAE/RMSE/MSE for individual CDM samples, the CDM sample mean, and the CDM sample median, alongside SR-DAE, all computed under a clearly defined normalization; (ii) add ℓ-space coherence or correlation diagnostics (e.g., coherence $(C_\ell^{\rm cross})^2/(C_\ell^{\rm rec}C_\ell^{\rm true})$ and also $C_\ell^{\rm rec}/C_\ell^{\rm true}$, not only the transfer $T(\ell) = C_\ell^{\rm cross}/C_\ell^{\rm true}$) to show whether CDM mean preserves true structure or injects excess power; (iii) check whether CDM mean approaches SR-DAE output under any limiting setting (e.g., fewer steps, different conditioning strength), which would help interpret what diffusion is adding; and (iv) clarify how the diffusion model is intended to be used scientifically if its point estimate is worse—e.g., is SR-DAE the recommended point estimator and CDM used only for uncertainty/ensembles?

6. **Baseline methods (cILC and WF) are under-described and may not reflect best-practice competitive implementations, complicating the strength/fairness of the comparison (Sec. 2.2; Sec. 3.1).** For cILC, it is unclear whether weights are global or ℓ/needlet-dependent, how SEDs are obtained, and how beams/noise are handled. For WF, it is unclear whether filtering is per-band or joint multi-frequency, how signal/noise spectra are estimated (from truth, from observed maps, from simulations), and whether cross-spectra are used to mitigate noise bias.

   Recommendation: In Sec. 2.2, specify the exact cILC and WF pipelines (domain: map/harmonic/needlet; ℓ-binning; beam handling; covariance estimation; constraints; any regularization). Clarify whether the baselines are intentionally "simple" references, and if so, state this explicitly and discuss how stronger variants (e.g., needlet ILC/NILC/GNILC) could change results. Ideally, add one more competitive linear baseline (e.g., needlet/constrained NILC) or demonstrate that your cILC/WF hyperparameters are tuned on validation data to near-optimal performance under the same simulation assumptions.

7. **Quantitative reporting is currently insufficiently systematic, and several internal inconsistencies reduce confidence (Sec. 2.5; Sec. 3.1–3.3; Fig. 3).** Examples: MSE values are reported in "scaled units" without a precise definition/mapping to Compton-y; SR-DAE/DAE test-set MSE is reported inconsistently (0.9458 vs 1.339); the null-test is described as zero-centered but reports mean $\approx$ $\mathbf{0.3466}$ with $\sigma \approx \mathbf{0.1054}$; and residual/transfer/gain plots are mostly interpreted qualitatively without representative numeric values and uncertainties.

   Recommendation: Add a compact quantitative summary in Sec. 3.1–3.3: (i) a table of pixel-space metrics (MAE/RMSE/MSE, Pearson r) for cILC, WF, SR-DAE, and CDM mean on the main test, OOD, and high-noise splits, with error bars across patches; (ii) numeric summaries of transfer/residual metrics at representative multipoles (e.g., $\ell \approx \mathbf{1000/3000/5000}$) with scatter; (iii) explicitly define the scaling/normalization used for "scaled units" and provide a conversion back to physical y units

(or an interpretable normalization such as error normalized by $\mathbf{std}(y_{\mathrm{true}})$); and (iv) fix and reconcile the inconsistent MSE and null-test statements by clearly stating dataset, model variant, and normalization used in each figure/table.

8. **Dataset splitting, OOD definition, and leakage controls are not described with enough precision, and there is an apparent split arithmetic inconsistency (Sec. 2.1; Sec. 3.2).** The stated train/val/test counts (1066/228/229) sum to 1523, leaving no patches for a separate "top 5%" OOD subset ($\sim$ **76** patches). Additionally, if patches overlap spatially or are drawn from the same underlying realization/lightcone region, train/test leakage could inflate performance. The OOD definition (top 5% by peak tSZ) may also not cover other realistic domain shifts (CIB SED variation, different foreground mix, calibration/beam errors).

   Recommendation: In Sec. 2.1 and Sec. 3.2, fix the dataset accounting by explicitly listing: total patch count, OOD count, and the remaining train/val/test counts (with integers that add up). State whether patches overlap and how you prevent spatial leakage (e.g., split by independent sky areas/lightcone segments or by halo IDs). Provide summary statistics (mass, redshift, peak y) for each split. If feasible, add a second OOD axis (e.g., altered CIB SED/noise/beam perturbation) or at least discuss that peak-y OOD does not capture full real-data domain shift.

9. **Scientific validation via scaling relation is currently ambiguous in its physical interpretation (Sec. 2.5; Sec. 3.3.2).** The text refers to a Y_{SZ}–"mass proxy" relation but also uses "peak tSZ signal" as a proxy, which is not a standard mass proxy in real analyses and is sensitive to beam/noise. The definition of Y_{SZ} (aperture, centering, background subtraction, beam correction) is not sufficiently specified, making it hard to interpret "tighter/less biased" claims.

   Recommendation: Clarify Sec. 2.5 and Sec. 3.3.2 by (i) explicitly defining the x-axis quantity (true halo mass M_{500} from FLAMINGO vs peak y vs another proxy) and renaming accordingly (e.g., Y_{SZ}–y_{peak} if that is what is used); (ii) defining how Y_{SZ} is computed (aperture radius, centering, integration method, background subtraction, beam handling); and (iii) reporting fitted slope/normalization/scatter/bias with uncertainties (bootstrap) for truth, cILC, SR-DAE, and CDM.

10. **Limitations and failure modes of a simulation-trained, generative reconstruction approach are not discussed explicitly enough, despite being central for real-data applicability (Sec. 3.2–3.3; Sec. 4).** Key concerns include domain shift (single simulation suite/feedback model), learned tSZ–CIB correlations that might not match reality, sensitivity to unmodeled foregrounds and calibration/beam systematics, and the possibility of hallucinated small-scale structure in low-SNR regions (especially relevant for diffusion sampling).

    Recommendation: Strengthen Sec. 4 (or end of Sec. 3.3) with a focused limitations section: explicitly discuss domain shift risks (FLAMINGO baryonic physics, CIB modeling), potential biases in pressure profiles/scaling relations, sensitivity to missing

components and instrument systematics, and hallucination risks. Outline concrete mitigations (training across multiple simulations/feedback models; foreground/model perturbation during training; domain adaptation; cross-validation with external observables such as X-ray or weak-lensing; conservative masking/SNR-based usage rules for diffusion samples).

Minor issues

1. Figure 1 presentation could inadvertently over-rely on qualitative inspection: potential differences in color limits, lack of explicit beam/effective resolution indication, limited spatial-scale annotation, and absence of residual (recon−truth) panels; showing a single patch risks cherry-picking (Fig. 1; Sec. 3.1).

   Recommendation: For Fig. 1, enforce identical color limits across reconstructions and state this in the caption; indicate effective beam FWHM with an inset circle and specify whether panels are beam-matched; add a scale bar/axis ticks and patch size; include residual panels with symmetric diverging colormap; and add a supplementary figure with multiple randomly selected patches (or a montage) to demonstrate typical behavior.

2. Figure 3 contains clarity/consistency issues: the null-test mean shown appears inconsistent with the "zero-centered" claim; MSE bars/histograms lack sample sizes and uncertainty; saliency maps may not be normalized consistently across panels; and "scaled units" are not defined on axes (Fig. 3; Sec. 3.2–3.3).

   Recommendation: Revise Fig. 3 to (i) clearly define the null-test statistic and units, mark mean/median with uncertainties, and correct the "zero-centered" wording if needed; (ii) add error bars/box/violin plots and sample counts for MSE comparisons; (iii) state Integrated Gradients normalization and enforce consistent colorbar scaling where comparisons are intended; and (iv) define the scaling used for all "scaled unit" axes in the caption or a referenced metrics subsection.

3. PIT and uncertainty evaluation are not described at sufficient procedural detail, especially given only 10 diffusion samples (Sec. 2.5; Sec. 3.3.1). With small sample sizes, PIT resolution is limited and details (pixelwise vs regionwise PIT, masking, SNR stratification) materially affect interpretation.

   Recommendation: In Sec. 3.3.1, specify whether PIT is computed per pixel or aggregated regions, whether a mask is applied (e.g., around clusters), whether PIT is stratified by SNR/tSZ amplitude, and whether the empirical CDF from samples is used directly (recommended) vs assuming Gaussianity from mean/variance. Add a sensitivity check of PIT vs number of samples (e.g., 10/30/100) and vs DDIM step count in an Appendix.

4. Integrated Gradients interpretability analysis is potentially anecdotal without methodological details and aggregation across many patches (Sec. 3.2). Baseline choice, number of interpolation steps, and stability across baselines are not specified.

   Recommendation: In Sec. 3.2, state the IG baseline, interpolation steps, and any smoothing. Add a simple population-level summary (e.g., average absolute attribution per channel over N test patches, possibly binned by SNR or cluster-centric radius) to support claims about adaptive channel use beyond a few examples.

5. Instrument naming is inconsistent (SO vs Planck high-frequency channels), which can confuse readers about the assumed data availability and beams/noise (Sec. 2.1; figure captions).

   Recommendation: Standardize terminology throughout: if 545/857 GHz are Planck-like external maps used alongside SO bands, state this explicitly in Sec. 2.1 and captions; otherwise remove Planck references and describe the full configuration as an SO-like (or SO+Planck) hybrid with a clear table.

6. A manuscript metadata/presentation concern: the author/affiliation line "Anthropic, Gemini & OpenAI servers. Planet Earth." (as noted) reads like placeholder text and is not appropriate for a scientific submission.

   Recommendation: Replace placeholder affiliations with standard institutional affiliations and acknowledgments consistent with the venue's policies.

Very minor issues

1. Notation and wording inconsistencies (e.g., "Sunyaev-Zel'dovich" apostrophes, "arcminute/arcmin" variants, capitalization of SR-DAE/CDM, stray heading markers like a leading '#') reduce polish (Sec. 1–4).

   Recommendation: Proofread and standardize notation/style across the manuscript; ensure section headings and figure references follow a consistent LaTeX/venue style.

2. Probabilistic notation is informal in places (e.g., $p(\mathbf{tSZ} \mid \mathbf{Observed})$ without introducing symbols for observed maps and targets), which makes Sec. 2.4–2.5 harder to parse precisely.

   Recommendation: Introduce consistent symbols (e.g., $\boldsymbol{x}$ for observed multi-frequency maps, $\boldsymbol{y}$ for tSZ) and use $p(\boldsymbol{y} \mid \boldsymbol{x})$ throughout; align this notation with the diffusion-model equations you add.

3. Power-spectrum notation C_ℓ is used without explicitly stating flat-sky vs full-sky convention, binning, and mean-subtraction/windowing assumptions, which matters for L_{spec} and transfer-function definitions (Sec. 2.3; Sec. 3.1).

Recommendation: Add a brief clarification (main text or Appendix) of the spectrum estimation convention (flat-sky Fourier vs spherical harmonics), binning, and any window/mask/apodization/beam corrections.

Mathematical consistency audit

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

Maths relevance: light

The paper contains only a small number of explicit equations (a composite loss, and a power-spectrum transfer function). Most mathematical content is described verbally (cILC constraints, Wiener filter construction, diffusion training objective, PIT calibration) without explicit formulas, which limits the depth of a symbolic audit.

Checked items

1. ⚠ **Composite loss structure** (Eq. (1), Sec. 2.3, p.4)
    - **Claim:** Training loss is a linear combination of an L_1 pixel loss, a spectral loss, and a residual–CIB correlation penalty: $L = L_{L1} + \lambda_1 L_{\text{spec}} + \lambda_2 L_{\text{corr}}$.
    - **Checks:** notation consistency, dimensional/unit consistency (qualitative), well-posedness of objective
    - **Verdict:** UNCERTAIN; confidence: medium; impact: critical
    - **Assumptions/inputs:** L_{L1}, L_{spec}, L_{corr} are scalar losses computed on the same training example/batch, λ_1 and λ_2 are scalar nonnegative weights
    - **Notes:** The linear-combination form is algebraically fine, but L_{spec} and L_{corr} are not defined mathematically, so it is impossible to check normalization, units/scales, or whether the combined objective is consistent (e.g., whether L_{corr} is bounded/scale-invariant and how multi-channel CIB inputs are handled).

2. ✖ **L1 term interpretation vs stated estimator** (Sec. 2.3 (L_{L1} definition) and Sec. 3.3 (posterior median statement), pp.4 and 8)
    - **Claim:** L_{L1} enforces pixel-level accuracy; later the deterministic DAE output is described as the posterior median and also said to be optimal for MSE.
    - **Checks:** loss–estimator consistency, statistical optimality sanity-check
    - **Verdict:** FAIL; confidence: high; impact: moderate
    - **Assumptions/inputs:** L_{L1} is mean absolute error (L1), MSE refers to squared-error risk
    - **Notes:** The text conflates estimators: the conditional median aligns with $L1$ /absolute error, not with MSE. If the argument is meant to explain why the DAE has lower MSE than diffusion samples/mean, the correct statement

would involve the conditional mean for MSE, or else compare to MAE for a median-like estimator.

3. ⚠ **Spectral loss description** (Sec. 2.3, p.4)

    - **Claim:** L_{spec} penalizes differences between the power spectra of reconstructed and true maps.
    - **Checks:** definition completeness, normalization/constraints
    - **Verdict:** UNCERTAIN; confidence: low; impact: critical
    - **Assumptions/inputs:** A power spectrum estimator is used on flat-sky patches, Some norm is applied to spectral differences
    - **Notes:** No explicit formula is given for how spectra are computed (auto vs cross, binning, weighting by ℓ, log vs linear, normalization by C_{true}, etc.). Without this, one cannot verify internal consistency (e.g., whether L_{spec} is scale-invariant or dominated by large-scale power) or compatibility with Eq. (2)'s C_ℓ conventions.

4. ⚠ **Residual–CIB orthogonality penalty** (Sec. 2.3, p.4)

    - **Claim:** L_{corr} is a normalized cross-correlation penalty between reconstruction residual and input CIB maps, forcing orthogonality to CIB.
    - **Checks:** definition completeness, symbol/quantity consistency
    - **Verdict:** UNCERTAIN; confidence: medium; impact: moderate
    - **Assumptions/inputs:** Residual $r = \hat{y} - y_{\text{true}}$ (implied but not explicitly defined), Multiple high-frequency channels contain CIB information
    - **Notes:** The residual is not formally defined, nor is the cross-correlation measure (Pearson correlation? normalized dot product? per-pixel covariance?), nor how multiple CIB channels are combined. Also unclear whether maps are mean-subtracted before computing correlation, which materially changes the penalty.

5. ✔ **Power spectrum transfer function definition** (Eq. (2), Sec. 2.5, p.5)

    - **Claim:** Transfer function is $T(\ell) = C_\ell^{\text{cross}}/C_\ell^{\text{true}}$, where C_ℓ^{cross} is cross-spectrum of reconstruction and truth and C_ℓ^{true} is truth auto-spectrum; $T \approx 1$ indicates unbiased recovery of power.
    - **Checks:** algebra/sanity limiting cases, dimensional/unit consistency, definition consistency
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** $C_\ell^{\text{true}} \neq 0$ in the ℓ-range of interest, C_ℓ^{cross} is computed with consistent normalization as C_ℓ^{true}
    - **Notes:** The ratio is dimensionless and has correct limiting behavior: if $\hat{y} = y_{\text{true}}$ then $C_\ell^{\text{cross}} = C_\ell^{\text{true}}$ and $T(\ell) = 1$. If $\hat{y} = \alpha y_{\text{true}} + n$ with n uncorrelated with y_{true} then $T(\ell) = \alpha$, consistent with interpreting T as an amplitude/re-

sponse. The statement 'unbiased recovery of power' is broadly consistent with this definition (though it measures response, not necessarily equality of auto-power in presence of correlated noise).

6. ⚠ **Reconstruction gain ratio definition (verbal)** (Sec. 2.5 bullet 'Reconstruction Gain', p.5; Sec. 3.3.4, p.9)
    - **Claim:** Gain ratio is the ratio of residual power spectrum of cILC reconstruction to that of the deep learning model as a function of ℓ.
    - **Checks:** dimensional/unit consistency, definition completeness
    - **Verdict:** UNCERTAIN; confidence: medium; impact: minor
    - **Assumptions/inputs:** Residual power spectrum means power spectrum of (recon − truth), Both methods' residual spectra are computed with the same estimator
    - **Notes:** The ratio is dimensionless and conceptually fine, but no explicit equation is provided: it is unclear whether the 'residual power spectrum' is $C_{\mathrm{res}}(\ell) = C_{(\hat{y}-y_{\mathrm{true}})}(\ell)$ or something else (e.g., auto-power of residual maps, binned). Without a formula, consistency with other harmonic-space quantities cannot be verified.

7. ✖ **Null-test 'zero-centered' description** (Sec. 3.2, p.7-8; Fig. 3 caption p.11)
    - **Claim:** Noise-only inputs produce an output distribution that is 'zero-centered Gaussian' / 'centered around zero,' while the text also reports a nonzero mean value.
    - **Checks:** internal statement consistency
    - **Verdict:** FAIL; confidence: medium; impact: minor
    - **Assumptions/inputs:** Zero-centered means mean approximately $\mathbf{0}$ in the same scaling used to report the mean
    - **Notes:** The descriptive claim ('zero-centered') conflicts with the simultaneous report of a nonzero mean in the same passage. Even without validating the numeric value, the wording is internally inconsistent unless an additional scaling/de-normalization distinction is stated (not present).

8. ⚠ **Conditional distribution notation and training objective for CDM** (Sec. 2.4, p.4)
    - **Claim:** CDM learns full conditional distribution $p(\mathbf{tSZ} \mid \mathbf{Observed})$ via a noise prediction task with a linear noise schedule; DDIM sampler used for inference.
    - **Checks:** definition completeness, notation consistency
    - **Verdict:** UNCERTAIN; confidence: high; impact: critical
    - **Assumptions/inputs:** A standard diffusion forward process and a conditional denoiser are used

- **Notes:** No equations are provided for the forward noising process, the conditional denoising model, or the loss minimized in training. Therefore the mathematical claim that the model learns $p(\mathbf{tSZ} \mid \mathbf{Observed})$ cannot be audited for internal consistency.

Limitations

- The provided content contains only two explicit numbered equations; most key mathematical objects (cILC weights, Wiener filter, spectral and correlation losses, diffusion objective) are described verbally without formulas, preventing full verification.
- Some potentially relevant mathematical definitions may be embedded in figures or omitted from the parsed text; this audit is limited to the text and visible equation content supplied.
- No explicit definitions are given for power spectrum estimation conventions on flat-sky patches (Fourier normalization, binning, windowing), limiting the ability to check harmonic-space consistency beyond basic sanity checks.

Numerical results audit

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

8 numeric/consistency checks were run: 5 PASS and 3 FAIL. The main failures concern (i) dataset split totals vs a stated **5%** held-out subset, (ii) a null-test mean described as "centered around zero" despite being $\sim$ **3.29** standard deviations from zero, and (iii) conflicting test-set MSE values for the deterministic SR-DAE/DAE across sections.

Checked items

1. ✖ **C1_dataset_split_totals** (p.3, Section 2.1 "Simulations and dataset")
    - **Claim:** "The full dataset comprises **1523** flat-sky patches… isolate the top **5%**… The remaining patches are then split into a training set (**1066**), a validation set (**228**), and a standard test set (**229**)."
    - **Checks:** parts_vs_total_with_percentage_subset
    - **Verdict:** FAIL
    - **Notes:** Train+val+test=**1523**, implying OOD=**0**, but **5%** of **1523** is $\sim$ **76.15** (floor/round **76**, ceil **77**). Closest rounding still differs by **76** patches (beyond $\pm$**1**).
2. ✔ **C2_frequency_band_count** (p.3, Section 2.1 "Simulations and dataset")
    - **Claim:** "Each data sample consists of a set of six multi-frequency maps corresponding to the SO frequency bands at **90**, **150**, **217**, **353**, **545**, and **857** GHz."
    - **Checks:** count_matches_list_length
    - **Verdict:** PASS

- **Notes:** Six frequencies are listed (**90**, **150**, **217**, **353**, **545**, **857**), matching the stated count of six.

3. ✔ **C3__training__epochs__consistency** (p.4 Section 2.3; p.6 Section 3.2)
    - **Claim:** "The SR-DAE is trained for **20** epochs…" and later "Our deterministic SR-DAE model was trained for **20** epochs…"
    - **Checks:** repeated_constant_consistency
    - **Verdict:** PASS
    - **Notes:** Both sections state **20** epochs; values match exactly.

4. ✔ **C4__cdm__epochs__and__sampling__numbers** (p.4 Section 2.4; p.8 Section 3.3)
    - **Claim:** "The model is trained for **15** epochs… During inference, we use a **50**-step DDIM sampler to generate multiple (**10**) realizations…" and later "generate an ensemble of **10** tSZ map realizations…"
    - **Checks:** repeated_constant_consistency
    - **Verdict:** PASS
    - **Notes:** The number of realizations is consistently reported as **10** in both places.

5. ✔ **C5__mse__standard__vs__ood__vs__high__noise__marginality** (p.7, Section 3.2 (text describing Figure 3 bottom-right))
    - **Claim:** "MSE of **0.9458**… OOD … MSE increases to **1.8239**… high instrumental noise (**95**th percentile), the MSE shows only a marginal increase to **0.9744**."
    - **Checks:** relative_change_computation
    - **Verdict:** PASS
    - **Notes:** Ordering holds: **1.8239** (OOD) $>$ **0.9744** (high-noise) $>$ **0.9458** (standard). Computed deltas: high-noise $-$ standard $=$ **0.0286** ($\approx$ **3.02%**); OOD $-$ standard $=$ **0.8781** ($\approx$ **92.84%**).

6. ✖ **C6__null__test__mean__centered__vs__claim** (p.8, Section 3.2 (null test description))
    - **Claim:** "output… with a mean of **0.3466** and a standard deviation of **0.1054** (in scaled units), centered around zero."
    - **Checks:** claim_vs_numeric_value_sanity
    - **Verdict:** FAIL
    - **Notes:** Computed $z = |\text{mean}|/\text{std} = 0.3466/0.1054 \approx 3.288$, which conflicts with a "centered around zero" description under common interpretations.

7. ✖ **C7__mse__dae__standard__test__inconsistency** (p.7 Section 3.2 vs p.8 Section 3.3)

- **Claim:** SR-DAE test-set MSE is reported as **0.9458** (standard test set) and later "the deterministic DAE achieves a scaled MSE of **1.339** on the test set".
- **Checks:** repeated_metric_consistency
- **Verdict:** FAIL
- **Notes:** Values differ well beyond tolerance; needs clarification whether they correspond to different splits/scaling/protocols or is a numerical inconsistency.

8. ✔ **C8_cdm_vs_dae_mse_ordering** (p.8, Section 3.3)
    - **Claim:** "DAE achieves… MSE of **1.339**… CDM ensemble mean has a higher MSE of **9.5667**."
    - **Checks:** inequality_check
    - **Verdict:** PASS
    - **Notes:** Inequality holds ($\mathbf{9.5667} > \mathbf{1.339}$). Computed ratio $\mathbf{9.5667}/\mathbf{1.339} \approx \mathbf{7.145}$.

Limitations

- Only parsed text provided; figures/tables are not available as machine-readable numeric data, so claims relying on plotted curves cannot be fast-verified without pixel/value extraction (excluded by scope).
- Several qualitative phrases (e.g., "marginal increase", "centered around zero", "near unity", "significantly greater") lack explicit quantitative thresholds; proposed checks can compute implied ratios/z-scores but cannot conclusively validate wording without authors' defined criteria.
- Potential metric-definition ambiguity (e.g., two different reported SR-DAE/DAE test MSE values) cannot be resolved without additional context on scaling, dataset subset, or evaluation protocol beyond what is explicitly stated in the text.