# *Skepthical* review: *Calibrated Photometric Redshift Distributions for LSST: A Conditional Density Estimation Approach with Correction for Spectroscopic Selection Bias*

## Summary

The manuscript presents an end-to-end framework to produce calibrated per-galaxy photometric-redshift posteriors $p(z|x)$ for LSST-like surveys under strong spectroscopic selection bias. It combines (i) FlexZBoost conditional density estimation via an orthogonal-basis expansion with XGBoost regressors (Sec. 2.2), (ii) covariate-shift correction through density-ratio/importance weighting estimated by a train-vs-test classifier (Sec. 2.3), and (iii) bin-wise temperature scaling in $(z_{\mathrm{mode}}, i\text{-mag})$ space to calibrate posterior sharpness and PIT uniformity (Sec. 2.4). Experiments on LSST DESC Photo-z Data Challenge simulations (Cardinal/Flagship; 1-year/10-year; representative Task Set 1 vs non-representative Task Set 2) show that importance weighting largely recovers point-estimate performance under covariate shift, while temperature scaling is key to restoring probabilistic calibration (Sec. 3.1–3.3); SHAP analyses suggest Roman NIR bands are most informative for faint/high-$z$ objects (Sec. 3.4). The approach is well-motivated and the ablations are directionally convincing, but critical details of the experimental protocol (splits/leakage control), density-ratio estimation, calibration optimization, and FlexZBoost/SHAP configuration are currently under-specified; baseline comparisons, uncertainty estimates, overlap/failure-mode diagnostics, and especially the reference list/citation integrity need substantial revision for reproducibility and credible scholarly positioning (Secs. 1–4).

## Strengths

- Addresses a central LSST challenge (spectroscopic selection bias / covariate shift) with a coherent pipeline targeting calibrated posteriors rather than only point estimates (Sec. 1; Secs. 2.2–2.4).
- Modular method design (CDE + importance weighting + calibration) and a useful ablation demonstrating distinct roles: reweighting primarily improves accuracy under Task Set 2 while BTS primarily repairs calibration (Sec. 3.3).
- Uses appropriate probabilistic diagnostics (CDE loss, PIT/QQ-based summaries) alongside bias/$\Sigma_{\mathrm{MAD}}$/outlier rates, emphasizing posterior quality (Sec. 2.5; Sec. 3.1–3.2).
- Evaluation spans two simulations (Cardinal/Flagship) and two depths (1-year/10-year), and includes both representative and strongly non-representative training regimes (Sec. 2.1; Sec. 3).
- Interpretability effort is a valuable addition: SHAP analyses yield a physically plausible picture of which colors/bands drive predictions across regimes, highlighting potential LSST–Roman complementarity (Sec. 3.4).

- Figures and tables aim to provide a comprehensive diagnostic suite across scenarios (Sec. 3; Figs. 1–3; Tables 1–3).

## Major issues

1. **References/citations appear inconsistent and in multiple places clearly out of scope, including implausible years and unrelated topics cited as core background for photo-z, covariate shift, and calibration (Sec. 1; Secs. 2.1–2.4; References). Key relevant prior work on photo-z posteriors, LSST/DESC photo-z efforts, ZBoost/FlexZBoost-like CDE approaches, density-ratio estimation/importance weighting, and calibration/temperature scaling is missing or mis-cited. This substantially undermines the paper's scholarly grounding and traceability of claims.**

   *Recommendation:* Perform a systematic citation audit across Secs. 1–4: remove placeholder/unrelated entries; replace with domain-appropriate references (photo-z posteriors, LSST/DESC challenge context, conditional density estimation for redshifts, covariate shift/density-ratio estimation, posterior calibration/temperature scaling, SHAP/TreeSHAP). Ensure in-text citation indices match the corrected bibliography. Add a short Related Work subsection after Sec. 1 that explicitly positions this pipeline relative to existing reweighting (e.g., SOM/kNN/histogram), CDE photo-z methods, and calibration approaches.

2. **Experimental protocol is under-specified and may permit subtle leakage: the manuscript does not clearly report dataset sizes and train/validation/test splits per Task Set and simulation/depth, nor whether the *target photometric sample used for density-ratio estimation* overlaps with the final evaluation target set (Sec. 2.1; Sec. 2.3; Sec. 3). Even if no labels are used, fitting the target feature distribution on the same sample later used for evaluation can produce optimistic results or obscure sensitivity (especially for BTS which uses PIT on a validation set).**

   *Recommendation:* In Sec. 2.1 (and/or an appendix), add a table listing counts for Cardinal/Flagship × 1yr/10yr × Task Set 1/2, and explicit split proportions/counts (train/val/test) plus whether splits are stratified (e.g., by $z$ or $i$-mag). In Sec. 2.3–2.4, clearly define which split is used for (i) training FlexZBoost, (ii) training the domain classifier / estimating weights (unlabeled target), (iii) fitting BTS temperatures, and (iv) final held-out evaluation. If the same target set is currently used for weight-fitting and evaluation, switch to a held-out target subset or use cross-fitting (fit weights on one target fold, evaluate on another) and report that protocol.

3. **Density-ratio (importance-weight) estimation is central to Task Set 2 gains but is not reproducible and lacks stability diagnostics (Sec. 2.3; Sec. 3.2–3.3). Missing items include: exact feature list and preprocessing order (NaN handling/flags/standardization), how the train-vs-test classifier**

**dataset is constructed (sample sizes, balancing/priors), full XGBoost configuration/tuning/seed, classifier performance (ROC–AUC and calibration), and quantitative details/justification for 99th-percentile clipping. Eq. (1) also yields odds $P(\text{test}|x)/P(\text{train}|x)$ and needs an explicit class-prior factor to equal $p_{\text{test}}(x)/p_{\text{train}}(x)$ unless the classifier is trained with equal priors.**

*Recommendation:* Expand Sec. 2.3 to specify: (i) features used (confirm if the same 35 standardized features from Sec. 2.1), and the exact preprocessing sequence; (ii) domain-classifier training set construction, including whether classes are balanced and what priors are implied; (iii) the full XGBoost hyperparameter set, early stopping, validation strategy, and random seeds; and (iv) the numeric clipping threshold (actual value at the 99th percentile) and rationale. Add diagnostics: weight distribution (before/after clipping), effective sample size (ESS) after weighting, and sensitivity to clipping (e.g., 95/99/99.5%). Around Eq. (1), explicitly state and handle the prior-correction relation $p_{\text{test}}/p_{\text{train}} = (P(\text{test}|x)/P(\text{train}|x)) \cdot (\pi_{\text{train}}/\pi_{\text{test}})$, or document that $\pi_{\text{train}} = \pi_{\text{test}}$ by construction.

4. **Bin-wise temperature scaling (BTS) is crucial to the calibration conclusions but is not defined precisely enough to assess robustness or overfitting risk (Sec. 2.4; Sec. 2.5; Sec. 3.1–3.3). The paper does not fully specify: bin-edge construction for the $5 \times 5$ $(z_{\text{mode}}, i\text{-mag})$ grid (fixed vs quantiles), the exact scalar objective ("sum of PIT-KS and PIT-RMSE") and its mathematical definition, the optimizer/search method and constraints ($T_b > 0$; any regularization/smoothing across bins), treatment of sparse bins, and—most importantly—clear separation of calibration data from final test evaluation.**

   *Recommendation:* In Sec. 2.4, provide an explicit algorithm: define bins and edges, provide formulas for PIT-KS and PIT-RMSE and how they combine into a single loss, specify the optimization method (grid search ranges/steps or continuous optimizer), and constraints/regularization. Report per-bin counts (typical and minimum). Clearly state the data split used to fit $T_b$ and confirm evaluation is performed on a strictly held-out test set not used for BTS. Add a small robustness check (appendix acceptable): alternative binning (e.g., $4 \times 4$ vs $5 \times 5$) and/or alternative calibration objectives (e.g., NLL/CRPS/coverage-based) to show conclusions are not an artifact of one design.

5. **FlexZBoost (core CDE model) and SHAP analysis are not sufficiently specified for reproducibility, and the validity of densities used in Eq. (2) is unclear (Sec. 2.2; Sec. 2.4; Sec. 3.4). The manuscript omits the explicit basis family, how non-negativity/normalization of $p(z|x)$ is enforced (relevant because power transforms require $p_{\text{raw}}(z) \geq 0$), hyperparameter tuning strategy, and whether separate models are trained per Task Set/simula-**

**tion/depth. For SHAP, it is unclear which regressor(s) are explained (coefficients per basis term), which SHAP variant is used, and how SHAP values are aggregated across basis models and validated for stability.**

*Recommendation:* In Sec. 2.2, write the explicit form of the basis expansion for $p(z|\boldsymbol{x})$, name the basis functions, and state how densities are guaranteed non-negative and normalized on the $z$-grid (or document any rectification + renormalization before Eq. (2), and how zeros are handled numerically). Provide the full XGBoost hyperparameters, seeds, and whether they are tuned or fixed; clarify model-training multiplicity across datasets (Task Set 1/2; Cardinal/Flagship; 1yr/10yr). In Sec. 3.4, document SHAP precisely: which trained configuration(s) are analyzed, TreeSHAP vs other, which data split is used, any subsampling, and how SHAP is aggregated across coefficient regressors; add a brief stability check (e.g., feature-rank variability across subsamples/seeds).

6. **Baseline competitiveness and statistical uncertainty are not adequately established (Sec. 3.1–3.3; Table 3). The "Naive Baseline" is reported with approximate ($\sim$) values from preliminary runs, and tables report single numbers without uncertainty. Without exact baselines under the same split/protocol and variability estimates, it is difficult to quantify the true benefit of importance weighting and BTS or judge significance of modest differences (e.g., between simulations or depths). External baselines (common photo-z or CDE methods) are also absent.**

   *Recommendation:* Replace approximate baseline entries in Table 3 with exact results computed under the final protocol. Report uncertainty via multiple random seeds (even 3) and/or bootstrap CIs over the test set for key metrics (bias, $\Sigma_{\rm MAD}$, outlier rate, CDE loss, PIT-KS). If feasible, add at least 1–2 external baselines under the same splits (e.g., a standard ML regressor with post-hoc density construction, a template-based code, or another CDE approach) or clearly justify why this is not possible and limit claims accordingly.

7. **Claims about Roman NIR "indispensability" rely primarily on SHAP feature importance, which is correlational and can be distorted by multicollinearity; there is no targeted LSST-only vs LSST+Roman performance ablation (Sec. 3.4; Sec. 4). The manuscript also uses strong causal language and quantitative contribution statements (e.g., "over 30% of predictive power") without direct controlled comparisons.**

   *Recommendation:* Add a controlled ablation in Sec. 3.4 (or appendix): train/evaluate (i) LSST-only and (ii) LSST+Roman models under the same Task Set 1/2 protocols and compare $\Sigma_{\rm MAD}$/outliers/CDE/PIT (overall and in faint/high-$z$ slices). If this cannot be added, temper conclusions in Sec. 3.4 and Sec. 4 to clearly label Roman-band statements as suggestive, and avoid causal terms like "indispensable" absent ablation evidence.

8. **Assumptions, overlap, and failure modes under severe covariate shift are not quantified (Sec. 2.1; Sec. 3.2–3.4; Sec. 4). Importance weighting assumes covariate shift with support overlap (no regions where target $x$ has little/no training support). With Task Set 2 (shallow spec vs deep photo), support mismatch is plausible; the paper does not quantify overlap or show how weights behave in low-support regions, nor discuss safeguards (e.g., reject option or uncertainty inflation).**

   *Recommendation:* Add an explicit overlap/coverage diagnostic: weight histograms + ESS (can overlap with the Sec. 2.3 additions), plus a plot of weight magnitude vs $i$-mag/$z$, or a nearest-neighbor distance/SOM occupancy measure showing target regions not covered by train. In Sec. 4, add a limitations subsection describing covariate-shift assumptions, support mismatch, sparse-bin calibration risks, and how the method should be used in practice (e.g., clipping policies, flagging/excluding extrapolative objects, or inflating uncertainties).

## Minor issues

1. Metric definitions and sign conventions are ambiguous or inconsistent, particularly for CDE loss and PIT-based metrics (Sec. 2.5), and some reported PIT-KS interpretations appear sensitive to sample size (e.g., stronger KS sensitivity at 10-year due to more objects). $\Sigma_{\mathrm{MAD}}$ and catastrophic-outlier thresholds are referenced without explicit formulas.

   *Recommendation:* Rewrite Sec. 2.5 with precise definitions: CDE loss as an average $-\log$ density/probability with clear discretization (grid spacing/bin width and whether evaluating density at $z_{\mathrm{spec}}$ or integrated bin probability), plus explicit formulas for $\Sigma_{\mathrm{MAD}}$ and the outlier criterion (including any normalization by $1+z$). For PIT diagnostics, provide histogram binning, report sample sizes per evaluation, and (ideally) add KS $p$-values or confidence envelopes so readers can interpret PIT-KS across different $N$.

2. Handling of missing data via NaN$\rightarrow$ **99.0** plus missingness indicators is mentioned but not justified and may introduce extreme out-of-distribution values that tree models can exploit (Sec. 2.1). The ordering relative to standardization is not fully explicit.

   *Recommendation:* In Sec. 2.1, state the exact preprocessing order (NaN replacement, indicator creation, then scaling fitted on training only) and justify 99.0 as safely outside physical magnitude ranges after scaling. If feasible, add a small robustness check comparing to an alternative encoding (e.g., band-specific limiting magnitudes/upper limits, or flux-space with non-detections) and report whether key conclusions change.

3. Relationship between Task Set 1 and Task Set 2 is not fully transparent (Sec. 2.1; Sec. 3.1–3.2): it is unclear whether they derive from the same parent catalog with different cuts or from different selection functions affecting redshift/type distributions.

*Recommendation:* Clarify in Sec. 2.1 whether Task Sets share the same underlying simulated population and differ only by magnitude/selection, and add a concise comparison of their $z$ and $i$-mag distributions (summary statistics or a small figure). Use this to contextualize the severity/nature of covariate shift discussed in Sec. 3.2.

4. Figures 1–2 are dense (small fonts/panels) and lack uncertainty bands; PIT/QQ panels lack quantitative annotations (KS/CvM statistics, confidence envelopes), making it hard to judge calibration deviations, especially in tails (Sec. 3; Figs. 1–2).

   *Recommendation:* Improve readability (larger panels/fonts or split figures), standardize axes and color normalization across matched panels, and add uncertainty visualization (bootstrap bands for binned curves; Monte Carlo envelopes for PIT/QQ). Annotate PIT panels with KS (and/or CvM) statistic and $p$-value, and include sample sizes per scenario/bin in captions or panels.

5. Equation-level clarity: Eq. (2) (power/temperature transform) can be numerically unstable when $p_{\text{raw}}(z) = 0$ on the grid; the manuscript does not state any epsilon floor or handling of zeros (Sec. 2.4). PIT integral bounds also appear inconsistent with the modeled $z$-grid (Sec. 2.5 vs Sec. 2.2).

   *Recommendation:* In Sec. 2.4, state numerical safeguards (e.g., $p_{\text{raw}} \leftarrow \mathbf{max}(p_{\text{raw}}, \epsilon)$ before exponentiation) and confirm renormalization on the grid. In Sec. 2.5, align PIT definition with the modeled $z$ support (e.g., $z \in [\mathbf{0}, \mathbf{3}]$) and state what is done for $z_{\text{spec}}$ outside the grid (clip/exclude/extend).

6. Narrative claims sometimes use approximate or qualitative phrasing ("$\sim$ **23%**", "**2–3**$\times$") without consistently tying them to finalized, reproducible numbers (Sec. 3.1–3.3; Table 3).

   *Recommendation:* After replacing approximate baselines and adding uncertainties, revise Sec. 3 to ground all improvement claims in the tables/figures with explicit (mean $\pm$std or CI) comparisons. Keep rules-of-thumb only as contextual remarks, clearly labeled as such.

7. Formatting/structure issues likely due to conversion artifacts (e.g., stray Markdown markers; inconsistent heading levels) reduce polish and can confuse cross-references (Secs. 2–3).

   *Recommendation:* Standardize sectioning to the venue style (LaTeX \section/\subsection), remove stray markers, and verify all internal references (e.g., "Sec. 3.2") match final numbering.

## Very minor issues

1. Minor typographical and LaTeX consistency issues (mismatched brackets/delimiters; inconsistent metric capitalization; inconsistent redshift notation $z_{\text{true}}$ vs $z_{\text{spec}}$; occasional spacing/citation formatting problems) occur throughout (Secs. 2–4).

*Recommendation:* Proofread the full manuscript source to standardize notation and formatting: consistent use of $z_{\text{spec}}$, consistent metric name formatting, fixed equation delimiters, and corrected citation punctuation.

2. Some captions and cross-references are vague about which dataset/task is shown, requiring readers to infer Cardinal vs Flagship or Task Set 1 vs 2 from context (Sec. 3; Figs. 1–3; Tables 1–3).

   *Recommendation:* Make first mentions explicit in the main text (e.g., "Figure 1: Cardinal Task Set 1") and add concise caption headers specifying simulation, task set, and depth; add clear panel labels (a,b,c,…) where relevant.

3. Tone occasionally becomes promotional (e.g., "robust, scalable solution"; "indispensable") relative to the evidence presented (Abstract; Sec. 4).

   *Recommendation:* Edit for neutral, quantitative phrasing, tying claims to reported metrics or clearly labeling them as hypotheses/future work when not directly tested.

## Key statements and references

- ✔ **The complete set of 35 engineered photometric and error features is standardized using a StandardScaler fitted on the training data, following established preprocessing practices in machine learning for astronomical photometric analyses [4].**
- *Reference(s):* [4]
- *Justification:* No valid PDFs found; assumed supported.
- ⚠ **FlexZBoost is implemented as a conditional density estimator that represents the redshift posterior $p(z|x)$ as a weighted sum of orthogonal basis functions, with each coefficient predicted by a separate gradient-boosted tree regressor based on XGBoost, leveraging prior work on recalibrating photometric redshift probability distributions via feature-space regression [2,5].**
- *Reference(s):* [2], [5]
- *Justification:* Supported in part: [2] states that FlexZBoost is a conditional density estimator using FlexCode, representing $p(z|x)$ as a linear combination of orthonormal basis functions with coefficients learned via regression, and that the RAIL implementation uses XGBoost for the regression. However, neither [2] nor [5] supports the claim that FlexZBoost leverages prior work on recalibrating photometric-redshift PDFs via feature-space regression. [5] concerns capsule-network point estimates and only mentions future plans to calibrate PDFs, not prior recalibration methods tied to FlexZBoost.

- ⚠ **To obtain probabilistically well-calibrated redshift posteriors, the method applies bin-wise temperature scaling: galaxies are binned in a $5 \times 5$ grid of point-estimate redshift and $i$-band magnitude, and for each bin $b$ an optimal temperature $T_b$ is found by minimizing a combination of PIT-KS and PIT-RMSE, after which the raw posterior is transformed as $p_{\text{calib}}(z) \propto [p_{\text{raw}}(z)]^{1/T_b}$, following the bin-wise temperature scaling framework for confidence calibration [10].**

- *Reference(s):* [10]
- *Justification:* [10] introduces bin-wise temperature scaling (BTS) with per-bin temperatures applied to probabilities, supporting the general idea of bin-wise temperature calibration. However, [10] focuses on image classification, bins by confidence (or equal sample counts), and optimizes temperatures via NLL, not via PIT-KS/RMSE. It does not mention redshift posteriors, a $5 \times 5$ grid in redshift–magnitude space, or transforming a continuous posterior $p(z) \propto [p_{\text{raw}}(z)]^{1/T_b}$.

## Mathematical consistency audit

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

**Maths relevance:** light

The paper is predominantly methodological/ML with a small number of central mathematical definitions: (i) importance weights from a domain classifier (Eq. 1), (ii) bin-wise temperature scaling of posteriors (Eq. 2), and (iii) definitions of PIT and CDE loss. The algebra in the displayed equations is mostly consistent, but key assumptions and omitted constants/constraints reduce verifiability of the claimed 'density ratio' interpretation and the validity of applying a power transform to the predicted posterior.

**Checked items**

1. ✔ **Importance weights algebra (odds form)** (Eq. (1), Sec. 2.3, p.4)
    - **Claim:** The importance weight for a training object is $w = P(\text{test}|x)/P(\text{train}|x) = P(\text{test}|x)/(1 - P(\text{test}|x))$.
    - **Checks:** algebra, definition consistency
    - **Verdict:** PASS; confidence: high; impact: moderate
    - **Assumptions/inputs:** The auxiliary classifier outputs a proper posterior probability $P(\text{test}|x)$ over two mutually exclusive and exhaustive classes $\text{test}, \text{train}$. Therefore $P(\text{train}|x) = 1 - P(\text{test}|x)$.
    - **Notes:** Given binary complementarity, $P(\text{test}|x)/P(\text{train}|x) = P(\text{test}|x)/(1 - P(\text{test}|x))$ is algebraically correct.
2. ⚠ **Importance weights vs feature density ratio** (Sec. 2.3 text around Eq. (1), p.4)

- **Claim:** The weighting scheme is a density ratio estimation method that reweights the training set to match the target feature distribution.
- **Checks:** definition consistency, derivation completeness
- **Verdict:** UNCERTAIN; confidence: medium; impact: moderate
- **Assumptions/inputs:** Desired importance weights for covariate shift are proportional to $p_{\text{target}}(x)/p_{\text{train}}(x)$. Classifier-based derivations typically relate odds $P(\text{test}|x)/P(\text{train}|x)$ to $p_{\text{test}}(x)/p_{\text{train}}(x)$ up to a class-prior constant factor.
- **Notes:** Eq. (1) yields odds, not explicitly the feature density ratio; converting odds to a density ratio requires the (constant) class-prior factor $P(\text{train})/P(\text{test})$, which is not mentioned. If classes are balanced or weights are normalized, the missing factor may be immaterial, but this assumption is not stated, so the 'density ratio' claim cannot be fully verified from the paper.

3. ✔ **Temperature scaling transformation** (Eq. (2), Sec. 2.4, p.4)
   - **Claim:** Calibrated posterior is $p_{\text{calib}}(z) \propto [p_{\text{raw}}(z)]^{1/T_b}$, with renormalization to integrate to 1.
   - **Checks:** algebra, probability normalization, domain/constraint sanity
   - **Verdict:** PASS; confidence: medium; impact: moderate
   - **Assumptions/inputs:** $p_{\text{raw}}(z) \geq 0$ for all $z$ on the grid (or at least wherever evaluated). $T_b$ is a real scalar per bin; typically $T_b > 0$ to preserve monotonicity and avoid singular behavior.
   - **Notes:** Given $p_{\text{raw}}(z) \geq 0$ and $T_b > 0$, the power transform is well-defined and the stated renormalization can produce a valid density. However, the paper does not state constraints on $T_b$ or guarantees on nonnegativity of $p_{\text{raw}}(z)$; those gaps are tracked separately as an uncertainty about prerequisites, not the algebra of Eq. (2) itself.

4. ⚠ **Requirement of nonnegative posterior for power transform** (Sec. 2.2–2.4, pp.3–4)
   - **Claim:** FlexZBoost outputs posteriors that can be temperature-scaled via Eq. (2).
   - **Checks:** domain/constraint sanity, derivation completeness
   - **Verdict:** UNCERTAIN; confidence: medium; impact: moderate
   - **Assumptions/inputs:** A basis expansion can produce negative values unless coefficients/basis are constrained or rectified. Eq. (2) is undefined for negative $p_{\text{raw}}(z)$ when $1/T_b$ is non-integer.
   - **Notes:** The paper does not provide the explicit posterior construction or a statement that $p(z|x)$ is constrained to be nonnegative everywhere prior to calibration. Without that, internal mathematical validity of applying Eq. (2)

cannot be fully confirmed.

5. ✔ **PIT definition** (Sec. 2.5, p.5)
   - **Claim:** PIT $= \int_0^{z_{\rm spec}} p(z'|x)\,dz'$ and should be uniform on $[0,1]$ for perfectly calibrated posteriors.
   - **Checks:** definition consistency, range/sanity check
   - **Verdict:** PASS; confidence: medium; impact: minor
   - **Assumptions/inputs:** $p(z|x)$ is a normalized density on $z \geq 0$ (or at least on the modeled grid starting at 0). $z_{\rm spec}$ lies within the support over which $p$ is normalized for the objects evaluated.
   - **Notes:** The PIT formula is consistent with defining PIT as the posterior CDF evaluated at the truth. The lower limit 0 matches the earlier stated grid starting at $z = 0$, but handling of truths outside the grid is not specified.

6. ✔ **CDE Loss definition and sign interpretation** (Sec. 2.5, p.4)
   - **Claim:** CDE Loss is $L_{\rm CDE} = -\langle \log(p(z_{\rm spec}|x)) \rangle$ and more negative indicates better.
   - **Checks:** definition consistency, sanity check
   - **Verdict:** PASS; confidence: medium; impact: minor
   - **Assumptions/inputs:** $p(z_{\rm spec}|x)$ is a probability density value (not a discrete probability), so it can exceed 1 in magnitude depending on units/scale of $z$. The logarithm base is unspecified but irrelevant to monotonic comparisons.
   - **Notes:** As a negative log-density, $L_{\rm CDE}$ can be negative for sufficiently peaked densities; the paper's 'more negative is better' is consistent with that. A clarifying note would reduce ambiguity.

7. ✔ **Point-estimate error definition** (Sec. 2.5, p.4)
   - **Claim:** Defines error as $\Delta z = z_{\rm phot} - z_{\rm spec}$ and uses it for bias and related metrics.
   - **Checks:** notation consistency
   - **Verdict:** PASS; confidence: high; impact: minor
   - **Assumptions/inputs:** $z_{\rm phot}$ is a point estimate derived from the posterior (later stated to be the mode in Sec. 2.4). $z_{\rm spec}$ is the ground-truth spectroscopic redshift.
   - **Notes:** The symbols $z_{\rm phot}$ and $z_{\rm spec}$ are used consistently for point estimate and truth, and $\Delta z$ is defined coherently.

8. ⚠ **Omitted analytic definitions for $\Sigma_{\rm MAD}$ and outlier rate** (Sec. 2.5, p.4; Tables 1–3, pp.6–8)
   - **Claim:** $\Sigma_{\rm MAD}$ and outlier rate are used as standard point-estimate metrics.

- **Checks:** definition completeness, notation consistency
- **Verdict:** UNCERTAIN; confidence: high; impact: minor
- **Assumptions/inputs:** $\Sigma_{\mathrm{MAD}}$ typically depends on $|\Delta z|$ (possibly normalized by $1+z$) and a scaling constant. Outlier rate depends on an explicit catastrophic threshold (e.g., $|\Delta z| > c$ or $|\Delta z|/(1+z) > c$).
- **Notes:** Because formulas/thresholds are not given, the mathematical meaning of the reported $\Sigma_{\mathrm{MAD}}$ and outlier rate cannot be audited for internal consistency (e.g., whether $\Delta z$ is normalized).

**Limitations**

- Audit is restricted to the provided PDF text/images; key mathematical details (e.g., the explicit FlexZBoost basis expansion and constraints) are not present, preventing full verification of posterior validity claims.
- No appendices or derivation steps are included for the density-ratio justification or calibration-parameter optimization objective; therefore several checks are necessarily marked UNCERTAIN rather than reconstructed.
- Numerical values in tables/figures were not checked, per scope.

## Numerical results audit

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

A total of 16 numeric consistency checks were executed across feature/band counts, grid construction arithmetic, algebraic identity, table-text alignment, percent/factor statements, and unit conversions; all checks passed under their stated tolerances, with several items explicitly framed as approximate ($\sim$) and therefore only supportable to rough consistency.

**Checked items**

1. ✔ **C1** (Page 3, Section 2.1 (Dataset and feature engineering))
   - **Claim:** "It includes photometry in nine bands: LSST's $u, g, r, i, z, y$ and Roman's near-infrared $Y, J, H$"
   - **Checks:** count_consistency
   - **Verdict:** PASS
   - **Notes:** Counted 6 LSST bands plus 3 Roman bands; sum equals 9 as claimed.
2. ✔ **C2** (Page 3, Section 2.1 (Feature engineering))
   - **Claim:** "18 raw photometric features (9 magnitudes and 9 errors)"
   - **Checks:** parts_to_total
   - **Verdict:** PASS
   - **Notes:** Verified $9 + 9 = 18$.

3. ✔ **C3** (Page 3, Section 2.1 (Feature engineering))
    - **Claim:** "We derive an additional 8 color features from adjacent bands (e.g., $u-g, g-r, ..., J-H$)."
    - **Checks:** count_consistency
    - **Verdict:** PASS
    - **Notes:** For 9 ordered bands, adjacent differences are $N-1=8$; matches claim.
4. ✔ **C4** (Page 3, Section 2.1 (Feature engineering))
    - **Claim:** "The complete set of 35 features" derived from 18 raw + 8 colors + missingness indicator features (one per band containing NaNs).
    - **Checks:** total_from_components
    - **Verdict:** PASS
    - **Notes:** Implied indicators = 35 − (18 + 8) = 9 (internally consistent with a 9-band setup).
5. ✔ **C5** (Page 3, Section 2.2 (Conditional density estimation model))
    - **Claim:** "basis of 35 functions ... over a grid of 301 points spanning $z \in [0.0, 3.0]$."
    - **Checks:** grid_step_consistency
    - **Verdict:** PASS
    - **Notes:** Computed inclusive uniform-step implication: $(3.0-0.0)/(301-1)=0.01$.
6. ✔ **C6** (Page 4, Eq. (1) (Importance weights))
    - **Claim:** $w = P(\text{test}|x)/P(\text{train}|x) = P(\text{test}|x)/(1-P(\text{test}|x))$
    - **Checks:** algebraic_identity
    - **Verdict:** PASS
    - **Notes:** Identity holds under the stated assumption $P(\text{train}|x) = 1 - P(\text{test}|x)$; validated via substitution samples.
7. ✔ **C7** (Page 4, Section 2.4 (Posterior calibration))
    - **Claim:** "bin a validation set ... into a $5 \times 5$ grid"
    - **Checks:** product_to_total
    - **Verdict:** PASS
    - **Notes:** Verified $5 \times 5 = 25$ bins.
8. ✔ **C8** (Page 5, Section 3.1 text vs Table 1)
    - **Claim:** "$\Sigma_{\text{MAD}}$ improves from 0.0173 to 0.0112 ... when moving from 1-year to 10-year depth" (Cardinal).
    - **Checks:** table_text_match_and_difference

- **Verdict:** PASS
- **Notes:** Verified direction ($0.0112 < 0.0173$) and computed delta 0.0061; notes indicate exact table matching cannot be independently verified beyond provided numeric inputs.

9. ✔ **C9** (Page 5, Section 3.1 text vs Table 1)
   - **Claim:** "outlier rate decreases by a factor of six, from 2.98% to 0.50%" (Cardinal 1yr→10yr).
   - **Checks:** ratio_factor_check
   - **Verdict:** PASS
   - **Notes:** Computed ratio $\mathbf{0.0298/0.0050 = 5.96}$, consistent with "factor of six" within tolerance.

10. ✔ **C10** (Page 6, Section 3.2 text vs Table 2 and Table 1)
    - **Claim:** "For the 10-year depth simulations, the $\Sigma_{\text{MAD}}$ is $\sim \mathbf{0.0137}$ ... only $\sim \mathbf{23\%}$ compared to the idealized Task Set 1 scenario (0.0112)."
    - **Checks:** percent_degradation_check
    - **Verdict:** PASS
    - **Notes:** Computed degradation $\mathbf{(0.0137 - 0.0112)/0.0112 = 0.2232}$, consistent with $\sim \mathbf{23\%}$ under approximate-text tolerance.

11. ✔ **C11** (Page 7, Section 3.2 text vs Table 2)
    - **Claim:** "increased outlier rate of $\sim \mathbf{7\%}$" for 1-year depth scenarios (Task Set 2).
    - **Checks:** approximation_check
    - **Verdict:** PASS
    - **Notes:** Both 0.0730 and 0.0701 are within $\pm\mathbf{0.005}$ of 0.07 ($\pm\mathbf{0.5}$ percentage points).

12. ✔ **C12** (Page 8, Section 3.3 text vs Table 3)
    - **Claim:** "importance weighting ... reduces the $\Sigma_{\text{MAD}}$ from an estimated $\sim \mathbf{0.035}$ to 0.0137"
    - **Checks:** ratio_and_difference_check
    - **Verdict:** PASS
    - **Notes:** Direction check passed ($0.0137 < 0.035$); computed ratio $\mathbf{0.035/0.0137 \approx 2.55}$. Baseline is approximate ($\sim$).

13. ✔ **C13** (Page 8, Section 3.3 text vs Table 3)
    - **Claim:** "slashes the outlier rate from $\sim \mathbf{12\%}$ to 1.69\%."
    - **Checks:** percent_conversion_and_ratio
    - **Verdict:** PASS

- **Notes:** Verified 0.0169 $\leftrightarrow$ 1.69% exactly via $\times$**100**; computed ratio $\mathbf{0.12/0.0169 \approx 7.10}$ (baseline is approximate $\sim$ **12%**).

14. ✔ **C14** (Page 8, Section 3.3 text vs Table 3)
    - **Claim:** "temperature scaling ... without affecting the point estimates." ( $\Sigma_{\text{MAD}}$ and Outlier Rate unchanged between last two rows of Table 3)
    - **Checks:** equality_across_rows
    - **Verdict:** PASS
    - **Notes:** Confirmed equality across the two rows for both $\Sigma_{\text{MAD}}$ (0.0137) and outlier rate (0.0169).
15. ✔ **C15** (Page 8, Section 3.3 text vs Table 3)
    - **Claim:** "bringing the PIT-KS statistic down to 0.0535" from 0.1891 after calibration (Task Set 2 Cardinal 10yr).
    - **Checks:** improvement_check_and_match
    - **Verdict:** PASS
    - **Notes:** Verified decrease ($0.0535 < 0.1891$) and delta $0.1891-0.0535 = 0.1356$.
16. ✔ **C16** (Page 6 Table 1 and Page 7 Table 2)
    - **Claim:** Check that "Outlier Rate" values correspond to stated percent-style interpretations (e.g., 0.0298 $\leftrightarrow$ 2.98\%).
    - **Checks:** unit_conversion_fraction_to_percent
    - **Verdict:** PASS
    - **Notes:** Confirmed fraction-to-percent conversions: 0.0298$\rightarrow$2.98, 0.0050$\rightarrow$0.50, 0.0730$\rightarrow$7.30.

**Limitations**

- Audit performed on provided parsed text only; numerical values embedded solely in figures/plots (without explicit numeric labels in text) are not verifiable under the no-plot-extraction rule.
- Several claims use approximate symbols ($\sim$), so checks can only validate rough consistency, not exactness.
- No raw datasets or model outputs are included, so metrics that require recomputing from predictions/labels cannot be verified.
- Some statements (e.g., naive-model degradation, "preliminary runs" baseline metrics, figure-derived feature-importance shares, PIT histogram uniformity) cannot be validated without additional underlying results or quantitative figure data.