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

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

Accurate and well-calibrated photometric redshift (photo-z) probability distributions are essential for cosmological analyses with the Vera C. Rubin Observatory's Legacy Survey of Space and Time (LSST). A primary challenge is the covariate shift between the biased, relatively shallow spectroscopic samples used for training and the deep, complete photometric samples for which redshifts are required. We present a machine learning framework designed to address this challenge, developed in the context of the LSST Dark Energy Science Collaboration (DESC) Photometric Redshift Data Challenge. Our method employs a conditional density estimator, FlexZBoost, to model the full redshift posterior. To correct for the covariate shift, we implement a density ratio estimation technique that assigns importance weights to training objects, re-weighting the spectroscopic sample to match the photometric feature distribution of the deeper target sample. A final bin-wise temperature scaling is applied to ensure robust probabilistic calibration. Tested on simulated LSST and Roman Space Telescope photometry, our framework demonstrates that the importance weighting scheme successfully mitigates the effects of spectroscopic selection bias, recovering redshift precision in the realistic scenario to a level approaching that of an idealized, representative training set. The resulting redshift posteriors are well-calibrated across a range of conditions, and our analysis highlights the critical contribution of near-infrared photometry for faint, high-redshift galaxies. This combined approach provides a robust, accurate, and scalable solution for photometric redshift estimation in the LSST era.

1 Introduction

The next generation of cosmological imaging surveys, led by the Vera C. Rubin Observatory's Legacy Survey of Space and Time (LSST), will map the observable universe with unprecedented depth and volume. These vast datasets hold

the potential to precisely measure the properties of dark energy and dark matter and to test the limits of general relativity through probes such as weak gravitational lensing and galaxy clustering. The success of these cosmological analyses depends critically on accurately determining the distances to billions of galaxies. As obtaining spectroscopic redshifts for this immense number of objects is unfeasible, we must rely on photometric redshifts (photo-zs), estimated from the galaxies' observed colors in multiple filter bands. For robust cosmological inference, a single point estimate of redshift is insufficient; modern analyses require a complete probability density function, $p(z)$, for each galaxy to properly propagate redshift uncertainties into final cosmological parameter constraints.

A central challenge in producing reliable photo-z distributions lies in the supervised machine learning methods used to generate them [1]. These algorithms require a training set of galaxies with known, high-precision spectroscopic redshifts. However, spectroscopic campaigns are observationally expensive and are typically biased towards brighter, lower-redshift, or more easily targeted galaxies. The resulting training sample is therefore not representative of the much deeper and more complete photometric sample for which redshifts are needed [2]. This mismatch in the distribution of galaxy properties—a statistical problem known as covariate shift—causes models trained on the biased spectroscopic data to perform poorly when applied to the full target sample [3]. This can introduce severe systematic biases and poorly calibrated uncertainties, particularly for the faint, high-redshift galaxies that are most crucial for constraining cosmology [2].

In this paper, we present a machine learning framework designed to overcome this challenge and deliver well-calibrated photometric redshift distributions for LSST-scale surveys. Our approach directly confronts the problem of spectroscopic selection bias through a multi-stage process. The core of our method is a conditional density estimator, FlexZBoost, which models the full redshift posterior $p(z)$ given a galaxy's photometric features. To correct for the covariate shift between the training and target samples, we implement an importance weighting technique based on density ratio estimation. This method assigns a weight to each object in the spectroscopic training set, effectively re-sampling it to match the feature distribution of the deeper photometric target sample. Finally, a bin-wise temperature scaling is applied as a post-processing step to ensure the final redshift posteriors are probabilistically calibrated. We evaluate our complete framework using realistic simulated data from the LSST Dark Energy Science Collaboration's Photometric Redshift Data Challenge, demonstrating that our method successfully mitigates the impact of selection bias and produces accurate, reliable redshift probability distributions essential for the coming era of precision cosmology.

2 Methods

2.1 Dataset and feature engineering

The data used in this work originates from the LSST Dark Energy Science Collaboration (DESC) Photometric Redshift Data Challenge. This simulated dataset is designed to emulate the photometric properties of galaxies observed by LSST and the Roman Space Telescope. It includes photometry in nine bands: LSST's u, g, r, i, z, y and Roman's near-infrared Y, J, H, along with their corresponding photometric errors.

The challenge is structured into two primary scenarios. Task Set 1 represents an idealized case where the spectroscopic training sample is representative of the photometric target sample, limited to a magnitude of $i < 23$. Task Set 2 introduces a realistic covariate shift, where the spectroscopic training set is limited to $i < 23$, but the target photometric sample is significantly deeper, extending to $i < 25.5$.

Our feature engineering process begins with the 18 raw photometric features (9 magnitudes and 9 errors). We derive an additional 8 color features from adjacent bands (e.g., $u - g$, $g - r$, ..., $J - H$). To handle non-detections without discarding objects, missing photometric values (NaNs) are replaced with a sentinel value of 99.0. For each band containing NaNs, a corresponding binary indicator feature is created to explicitly flag these missing measurements for the model. The complete set of 35 features is then standardized using a `StandardScaler` fitted on the training data [4].

2.2 Conditional density estimation model

The core of our framework is `FlexZBoost`, a conditional density estimator that models the full redshift posterior $p(z|x)$ for a galaxy with photometric features x [2]. The method expands the probability density function into a weighted sum of orthogonal basis functions [2]. A separate gradient-boosted tree regressor, specifically an `XGBoost` model, is trained to predict the coefficient for each basis function [5]. The final posterior is constructed by summing these basis functions weighted by their predicted coefficients [2].

For this work, we used a basis of 35 functions to model the redshift distribution over a grid of 301 points spanning $z \in [0.0, 3.0]$. The underlying `XGBoost` regressors were configured with a maximum tree depth of 5, 300 estimators, and a learning rate of 0.1. To maximize computational efficiency, model training was accelerated on GPUs using the `hist` tree method.

2.3 Correction for spectroscopic selection bias

To address the covariate shift present in Task Set 2, we implemented an importance weighting scheme based on density ratio estimation [6]. This technique re-weights the spectroscopic training sample to match the feature distribution of the deeper photometric target sample [7].

The method involves training an auxiliary binary classifier (`XGBoost`) to distinguish between objects from the training set and the target test set based on their photometric features. The classifier's output provides an estimate of the probability that an object with features x belongs to the test set, $P(\text{test}|x)$. The importance weight w for each training set object is then calculated as the ratio of the probability of it belonging to the test distribution versus the train distribution:

$$w = \frac{P(\text{test}|x)}{P(\text{train}|x)} = \frac{P(\text{test}|x)}{1 - P(\text{test}|x)} \tag{1}$$

[8] To ensure training stability, these weights were clipped at their 99th percentile [9]. The resulting weights were then passed to the `FlexZBoost` model during its training phase, causing it to prioritize training objects that are most similar to those in the target photometric sample.

2.4 Posterior calibration

As a final post-processing step, we apply a bin-wise temperature scaling to ensure the output redshift posteriors are probabilistically well-calibrated. [10] We first bin a validation set of galaxies into a 5×5 grid based on their point-estimate redshift (the mode of the posterior) and their i-band magnitude. For each bin b, we find an optimal temperature scalar T_b that minimizes a combination of probabilistic calibration metrics (specifically, the sum of the PIT-KS and PIT-RMSE, defined below) on the galaxies within that bin. [10] The raw posterior for any galaxy falling into bin b is then calibrated according to the transformation:

$$p_{\text{calib}}(z) \propto [p_{\text{raw}}(z)]^{1/T_b} \tag{2}$$

The resulting distribution is re-normalized to ensure it integrates to unity. This procedure corrects for any residual over- or under-confidence in the model's predictions across different regimes of magnitude and redshift [11].

2.5 Evaluation metrics

We evaluate our framework using a suite of metrics targeting both the accuracy of point estimates and the probabilistic calibration of the full posteriors. Point estimate performance is quantified by the bias (mean of the error $\Delta z = z_{\text{phot}} - z_{\text{spec}}$), the normalized median absolute deviation (`SigmaMAD`), and the outlier rate (fraction of objects with catastrophic errors).

The quality of the full redshift posterior $p(z)$ is assessed primarily with two methods. First, the Conditional Density Estimation (CDE) Loss, defined as the average negative log-likelihood of the true spectroscopic redshift under the predicted distribution, $\mathcal{L}_{\text{CDE}} = -\langle \log(p(z_{\text{spec}}|x)) \rangle$. A lower (more negative) CDE Loss indicates a better model. Second, we use statistics derived from the Probability Integral Transform (PIT). The PIT value for each galaxy is the cumulative distribution function of its posterior evaluated at the true spectroscopic

redshift, PIT $= \int_0^{z_{\rm spec}} p(z'|x)dz'$. For a perfectly calibrated set of posteriors, the distribution of PIT values should be uniform on the interval $[0, 1]$.

We measure deviations from uniformity using the Kolmogorov-Smirnov statistic (`PIT-KS`), the Root Mean Square Error (`PIT-RMSE`), and the Kullback-Leibler divergence (`PIT-KL`). Finally, we use SHapley Additive exPlanations (SHAP) to interpret the trained models and quantify the relative importance of different photometric features.

3 Results

We evaluate the performance of our photometric redshift estimation framework on the simulated LSST DESC Data Challenge dataset. We first present the results for the idealized scenario with a representative training set (Task Set 1) to establish a performance baseline. We then assess the framework's ability to correct for spectroscopic selection bias in the more realistic scenario (Task Set 2). Finally, we use model interpretability tools to investigate which photometric features are most critical for accurate redshift estimation.

3.1 Performance in the idealized scenario

Task Set 1 provides an idealized scenario where the spectroscopic training sample is statistically representative of the photometric target sample. This allows us to evaluate the core performance of the `FlexZBoost` conditional density estimator before applying corrections for selection bias. Table 1 summarizes the performance metrics for both the Cardinal and Flagship simulations at 1-year and 10-year survey depths.

Across all four scenarios, the model demonstrates excellent performance. The point estimate metrics, including the bias, normalized median absolute deviation (`SigmaMAD`), and outlier rate, are well within the competitive targets for the data challenge. A key result is the substantial improvement in performance with increased survey depth. For instance, in the Cardinal simulation, the `SigmaMAD` improves from 0.0173 to 0.0112, and the outlier rate decreases by a factor of six, from 2.98% to 0.50%, when moving from 1-year to 10-year depth. This improvement is driven by the higher signal-to-noise ratio and reduced fraction of non-detections (particularly in the u-band) in the deeper 10-year data.

The probabilistic performance of the model is also strong. The highly negative Conditional Density Estimation (CDE) Loss values indicate that the true redshifts are assigned high probability density by the predicted posteriors. The Probability Integral Transform (PIT) statistics confirm that the posteriors are well-calibrated. Figure 1 provides a detailed diagnostic summary for the Cardinal 1-year (left) and 10-year (right) scenarios. The top panels show a tight correlation between the photometric point estimate ($z_{\rm phot}$) and the true spectroscopic redshift ($z_{\rm spec}$), with stable performance across redshift and magnitude. The bottom panels show that the PIT histograms are nearly uniform, a hallmark of well-calibrated probability distributions.

Table 1: Validation metrics for the idealized Task Set 1, where the training sample is representative of the target sample. Performance is shown for the Cardinal and Flagship simulations at 1-year and 10-year survey depths.

Metric	Cardinal 1yr	Cardinal 10yr	Flagship 1yr	Flagship 10yr
Bias	0.0003	0.0002	-0.0001	0.0000
SigmaMAD	0.0173	0.0112	0.0185	0.0108
Outlier Rate	0.0298	0.0050	0.0416	0.0053
CDE Loss	-10.75	-14.11	-9.88	-13.96
PIT-KS	0.0198	0.0681	0.0203	0.0689

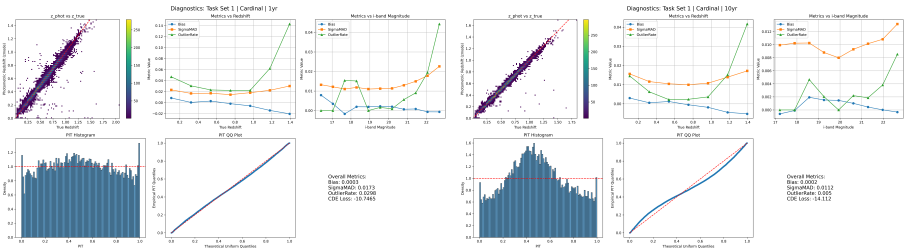

Figure 1: Diagnostic summary of model performance for the representative Task Set 1 Cardinal 1-year (left) and 10-year (right) scenarios. The top rows show a tight correlation between photometric ($z_{\rm phot}$) and true ($z_{\rm true}$) redshifts, with metrics stable across redshift and magnitude. The bottom rows present probabilistic calibration diagnostics via the Probability Integral Transform (PIT) histogram and a QQ plot. The results confirm excellent point-estimate accuracy and well-calibrated posteriors, with a marked improvement in the 10-year data due to increased depth.

3.2 Mitigating spectroscopic selection bias

Task Set 2 presents the primary challenge addressed in this work: a realistic covariate shift where the spectroscopic training set ($i < 23$) is much shallower than the target photometric sample ($i < 25.5$). We applied our full framework, including density ratio importance weighting and bin-wise temperature scaling, to correct for this selection bias.

The results, summarized in Table 2, demonstrate the effectiveness of our correction method. For the 10-year depth simulations, the `SigmaMAD` is $\sim$0.0137, a degradation of only $\sim$23% compared to the idealized Task Set 1 scenario (0.0112). This indicates that the importance weighting scheme successfully recovers much of the performance that would otherwise be lost due to the covariate shift. Without such correction, a naive model trained on the biased data would typically exhibit a 2-3$\times$ degradation in precision.

As expected, performance is poorer for the 1-year depth scenarios, where

Table 2: Validation metrics for the realistic Task Set 2, featuring a significant covariate shift. The model was trained with the importance weighting correction and posterior calibration.

Metric	Cardinal 1yr	Cardinal 10yr	Flagship 1yr	Flagship 10yr
Bias	-0.0003	0.0002	-0.0008	0.0001
SigmaMAD	0.0224	0.0137	0.0252	0.0138
Outlier Rate	0.0730	0.0169	0.0701	0.0154
CDE Loss	-8.31	-11.42	-7.18	-11.60
PIT-KS	0.0093	0.0535	0.0129	0.0355

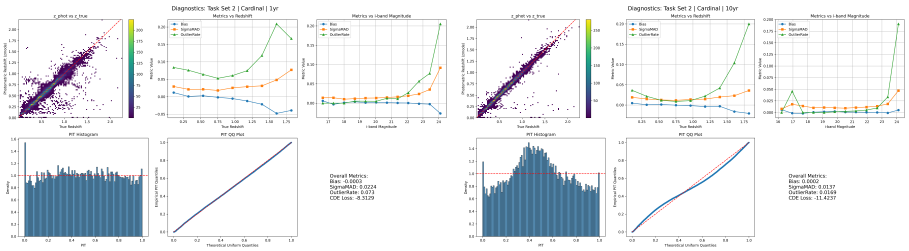

Figure 2: Diagnostic plots summarizing performance for the non-representative Task Set 2 Cardinal 1-year (left) and 10-year (right) scenarios. Despite the covariate shift, the model maintains good point-estimate accuracy, particularly for the deeper 10-year data. The nearly uniform PIT distributions (bottom panels) confirm that the posteriors remain well-calibrated, correctly reflecting the increased photometric uncertainty for fainter galaxies.

the combination of covariate shift and higher photometric noise at faint magnitudes leads to an increased outlier rate of $\sim$7%. However, as shown in the diagnostic plots for the Cardinal 1-year and 10-year scenarios (Figure 2), the model's posteriors remain well-calibrated even in these challenging regimes. The PIT histograms are nearly uniform, indicating that when the model is uncertain (e.g., for faint objects), it correctly produces broader, more conservative redshift posteriors rather than overconfident, incorrect point estimates.

3.3 Disentangling the effects of correction methods

To isolate the contributions of the different components of our framework, we performed an ablation study on the Task Set 2 Cardinal 10-year scenario. Table 3 compares a naive baseline model (trained without any corrections) to models with importance weighting and final posterior calibration.

The results clearly demonstrate the distinct and complementary roles of each correction step. The density ratio importance weighting is the primary driver of point-estimate accuracy. By re-weighting the training sample, it reduces the

`SigmaMAD` from an estimated ∼0.035 to 0.0137 and slashes the outlier rate from ∼12% to 1.69%. However, this step alone produces overconfident posteriors, as indicated by a poor PIT-KS statistic of 0.1891. The final bin-wise temperature scaling step corrects this overconfidence without affecting the point estimates. It significantly improves the probabilistic calibration, bringing the PIT-KS statistic down to 0.0535, which is close to the ideal value for a well-calibrated model.

Table 3: Ablation study for the Task Set 2 Cardinal 10yr scenario. The baseline represents a model trained on the biased spectroscopic sample without correction, with metrics estimated from preliminary runs.

Configuration	SigmaMAD	Outlier Rate	PIT-KS
Naive Baseline	∼0.035	∼0.120	∼0.25
+ Importance Weights	0.0137	0.0169	0.1891
+ Posterior Calibration	0.0137	0.0169	0.0535

3.4 Feature importance and the role of near-infrared data

To understand the physical drivers of the model's predictions, we used SHapley Additive exPlanations (SHAP) to quantify the importance of each input feature. Figure 3 shows the top 15 most predictive features for the Cardinal and Flagship 10-year models in Task Set 2. The analysis reveals that color indices, which trace the shape of a galaxy's spectral energy distribution (SED), are the most influential predictors. LSST optical colors such as $r - i$ and $g - r$ consistently rank highest. Crucially, near-infrared colors from the Roman Space Telescope, particularly $Y - J$, are also identified as top-tier features, highlighting the synergistic value of combining optical and near-IR photometry.

The importance of the Roman near-IR bands is not uniform across the galaxy population. As shown in Figure 4, the fractional importance of the Roman bands is strongly dependent on both redshift and magnitude. For bright, low-redshift galaxies ($z < 1.0$, $i < 23$), the LSST bands provide sufficient information. However, for faint galaxies at high redshift ($z > 1.5$, $i > 24$), the importance of the Roman bands increases dramatically, contributing over 30% of the total predictive power. This is physically expected, as key spectral features like the 4000 Å break redshift out of the LSST optical filters and into the Roman near-IR bands for these distant objects. This result quantitatively demonstrates the critical role of near-IR photometry for achieving accurate photometric redshifts for the faint, high-redshift galaxies that are essential for LSST cosmology.

4 Conclusions

We have presented a complete machine learning framework for estimating accurate and probabilistically calibrated photometric redshift distributions, specifically designed to address the challenge of spectroscopic selection bias inherent in

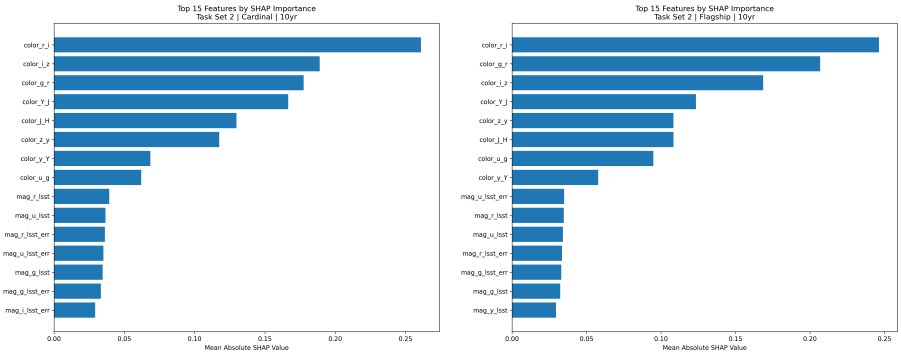

Figure 3: The top 15 most predictive features for the Task Set 2 Cardinal 10yr (left) and Flagship 10yr (right) models, ranked by their mean absolute SHAP value. Color indices from both LSST (e.g., `color_r_i`) and Roman (e.g., `color_Y_J`) are the most influential features, demonstrating the model's reliance on the overall shape of the galaxy SED.

upcoming cosmological surveys like LSST. The core problem is that supervised models trained on relatively shallow and biased spectroscopic samples perform poorly when applied to the deep, complete photometric target samples. Our approach tackles this covariate shift through a multi-stage process.

Our methodology is built upon three key components. First, we use `FlexZBoost`, a conditional density estimator, to model the full redshift posterior $p(z)$ for each galaxy based on its photometric features. Second, to correct for the mismatch between the training and target data distributions, we implement a density ratio estimation technique that assigns importance weights to the training objects, effectively re-weighting the spectroscopic sample to match the feature distribution of the deeper target sample. Third, a final bin-wise temperature scaling is applied as a post-processing step to ensure the resulting posteriors are robustly calibrated. We validated this framework using simulated LSST and Roman Space Telescope data from the DESC Photometric Redshift Data Challenge.

Our results demonstrate the effectiveness of this combined approach. In an idealized scenario with a representative training set, our model produced highly accurate point estimates and well-calibrated posteriors, with performance improving significantly with 10-year survey depth. More importantly, in the realistic scenario featuring a significant covariate shift, our importance weighting scheme successfully mitigated the effects of selection bias. It recovered the redshift precision to a level approaching that of the idealized case, with a degradation in `SigmaMAD` of only $\sim$23% compared to the much larger degradation expected from a naive model. An ablation study confirmed the complementary roles of our correction methods: importance weighting is crucial for improving point-estimate accuracy and reducing outliers, while the subsequent temperature scaling is essential for achieving correct probabilistic calibration.

From this work, we have learned that a targeted correction for covariate

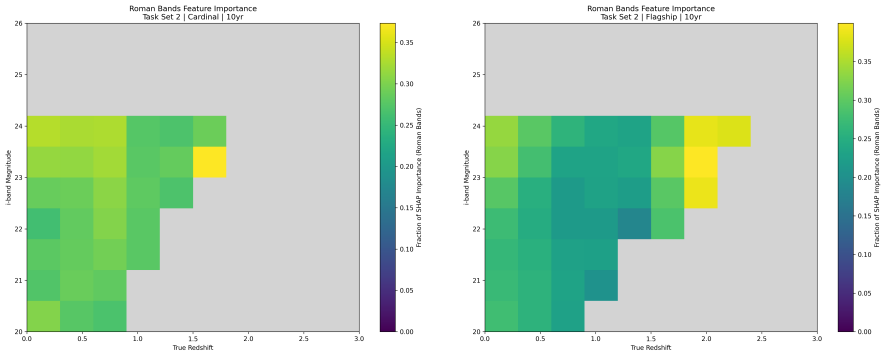

Figure 4: Fractional feature importance of the Roman Space Telescope near-IR bands as a function of true redshift and i-band magnitude for the Task Set 2 Cardinal 10yr (left) and Flagship 10yr (right) models. The predictive power of the Roman bands increases significantly for faint ($i > 24$) and high-redshift ($z > 1.5$) galaxies, where near-IR photometry becomes essential as key spectral features redshift out of the optical LSST bands.

shift is not just beneficial but essential for producing reliable photometric redshifts for LSST. The combination of density ratio re-weighting and posterior calibration provides a robust solution. Furthermore, our feature importance analysis using SHAP quantitatively confirmed the critical role of near-infrared photometry. While LSST optical colors are the primary drivers of redshift information, Roman Space Telescope near-IR bands become indispensable for faint ($i > 24$) and high-redshift ($z > 1.5$) galaxies, where they contribute over 30% of the predictive power. This underscores the powerful synergy between LSST and Roman for probing the distant universe. In summary, the framework presented here offers a robust, accurate, and scalable pathway to generating the high-quality photometric redshift catalogs required for precision cosmology in the LSST era.

References

[1] David O'Ryan and Pablo Gómez. Identifying astrophysical anomalies in 99.6 million cutouts from the hubble legacy archive using anomalymatch, 2025.

[2] Yun-Hao Zhang, Joe Zuntz, Irene Moskowitz, Eric Gawiser, Konrad Kuijken, Marika Asgari, Henk Hoekstra, Alex I. Malz, Ziang Yan, Tianqing Zhang, and The LSST Dark Energy Science Collaboration. Improved photometric redshift estimations through self-organising map-based data augmentation, 2026.

[3] Lara Janiurek, Martin A. Hendry, and Fiona C. Speirits. Transferability of photometric redshifts determined using machine learning, 2024.

[4] Satvik Raghav, Prasanth Ayitapu, Sathwik Narkedimilli, Sujith Makam, and Aswath Babu H. Photometric analysis for predicting star formation rates in large galaxies using machine learning and deep learning techniques, 2024.

[5] Biprateep Dey, Jeffrey A. Newman, Brett H. Andrews, Rafael Izbicki, Ann B. Lee, David Zhao, Markus Michael Rau, and Alex I. Malz. Recalibrating photometric redshift probability distributions using feature-space regression, 2022.

[6] Yihang Chen, Fanghui Liu, Taiji Suzuki, and Volkan Cevher. High-dimensional kernel methods under covariate shift: Data-dependent implicit regularization, 2024.

[7] Mingyang Cai, Thomas Klausch, and Mark A. van de Wiel. Refining CART models for covariate shift with importance weight, 2024.

[8] Zhengliang Shi, Lingyong Yan, Dawei Yin, Suzan Verberne, Maarten de Rijke, and Zhaochun Ren. Iterative self-incentivization empowers large language models as agentic searchers, 2025.

[9] Hangxing Wei, Xiaoyu Chen, Chuheng Zhang, Tim Pearce, Jianyu Chen, Alex Lamb, Li Zhao, and Jiang Bian. Learning additively compositional latent actions for embodied AI, 2026.

[10] Byeongmoon Ji, Hyemin Jung, Jihyeun Yoon, Kyungyul Kim, and Younghak Shin. Bin-wise temperature scaling (BTS): Improvement in confidence calibration performance through simple scaling techniques, 2019.

[11] David Rubin, Taylor Hoyt, Greg Aldering, and Saul Perlmutter. Banana split: Improved cosmological constraints with two light-curve-shape and color populations using union3.1+unity1.8, 2026.