

Embedding-Space Posterior Calibration for Zero-Shot Vision-Language Models

Zaiwang Gu^{1*}, Yingzhi Xia^{1*}, Yanyu Xu², Xulei Yang¹, Jun Cheng^{1†}

¹Institute of Advanced Intelligence and Computing (IAIC), A*STAR, Singapore

²Joint SDU–NTU Centre for Artificial Intelligence Research (C-FAIR),

Shandong University, China

{gu_zaiwang, xia_yingzhi, yang_xulei, cheng_jun}@a-star.edu.sg, xu_yanyu@sdu.edu.cn

*Equal contribution.

†Corresponding author.

Abstract

Zero-shot vision-language models (VLMs) such as CLIP demonstrate strong open-vocabulary recognition capabilities, yet their prediction confidence scores are often poorly calibrated. Existing calibration methods often estimate prediction stability by applying multiple augmentations in input space, introducing substantial computational overhead and limiting practical deployment. To this end, this paper proposes an embedding-space posterior calibration approach for zero-shot VLMs, which shifts the sampling process from the input space to the embedding space. Our approach defines a Confidence Stability Score that measures how stable the predicted confidence remains under posterior image-embedding perturbations. This score is then converted into a sample-wise temperature, so that unstable samples receive stronger confidence smoothing while stable samples remain comparable to their original predictions. Compared with the original CLIP baseline, our method reduces the mean Expected Calibration Error from 5.83 to 2.05, achieving a 64.8% relative reduction. Meanwhile, our method achieves this improvement with only a minor increase in GFLOPs, from 8.83 to 8.84.

1 Introduction

Zero-shot vision-language models (VLMs), such as CLIP (Radford et al., 2021) and SigLIP (Zhai et al., 2023), have demonstrated strong recognition capabilities across diverse domains. By aligning image and text representations in a shared feature or embedding space, these models can classify images by comparing an image embedding with a set of textual class embeddings, without requiring task-specific supervised training. However, zero-shot VLMs often suffer from poor confidence calibration, despite their strong recognition performance. The model may assign high confidence to an incorrect prediction, or produce overconfident

probabilities (Sharifdeen et al., 2025; Dong et al., 2025; Yoon et al., 2024). These miscalibration cases limit the reliability of VLMs in downstream safety-critical systems, where the confidence score is often used to decide whether a prediction should be trusted. Therefore, improving calibration for zero-shot VLMs is important for reliable deployment.

Existing calibration methods improve the calibration of VLMs mainly through post-hoc logit adjustment (Guo et al., 2017; Baumann et al., 2026) or test-time augmentation (Farina et al., 2024; Yoon et al., 2024; Huang et al., 2025; Yan et al., 2025). For example, test-time augmentation estimates reliability by applying multiple input transformations and aggregating the resulting predictions, but requires repeated forward passes through the image encoder. For large-scale VLMs, such repeated image encoding introduces substantial computational overhead, making input-space stability estimation inefficient for practical deployment. This motivates estimating prediction stability after the image encoder, rather than repeatedly applying the encoder to multiple augmented inputs. Intuitively, a reliable prediction should remain stable under small and plausible perturbations, regardless of whether these perturbations are applied in the image space or the embedding space. We refer to this property as *consistency*. Motivated by this observation, we measure consistency directly in the embedding space after a single image-encoder forward pass, thereby avoiding the expensive repeated image encoding required by input-space augmentation.

Since the image encoder is executed only once, embedding-space perturbation is significantly more efficient than input-space augmentation. However, naively adding random noise to the embedding lacks a clear probabilistic interpretation and may fail to capture meaningful local variations around the model prediction. Therefore, a principled approach is required to construct meaningful pertur-

bations around a VLM embedding.

In this paper, we propose Embedding-Space Posterior Calibration (EPC), a lightweight calibration framework for zero-shot VLMs. The key idea of EPC is to estimate prediction reliability by examining whether the predicted confidence remains stable under embedding-space posterior perturbations, rather than relying on multiple input-space augmentations. Specifically, EPC first reformulates deterministic image embeddings as posterior embedding distributions, from which posterior embedding samples are generated around the original embedding. It then measures the confidence consistency of the predicted class across these posterior samples and summarizes this consistency into a Confidence Stability Score (CSS). Finally, EPC converts CSS into a sample-wise temperature for adaptive calibration, such that predictions with unstable confidence are softened more strongly, while stable predictions receive weaker smoothing. In this way, EPC uses embedding-space consistency as a sample-wise calibration cue, improving probability reliability while avoiding repeated image-encoder forward passes and maintaining the zero-shot nature of VLMs.

The main contributions of this paper are summarized as follows:

1. We propose Embedding-Space Posterior Calibration (EPC), a calibration framework for zero-shot vision-language models based on embedding-space posterior sampling.
2. We introduce a Confidence Stability Score (CSS), which measures confidence consistency across posterior embedding samples and enables adaptive sample-wise temperature for calibration.
3. We empirically show that EPC reduces calibration error for zero-shot VLMs while maintaining zero-shot recognition accuracy with minimal additional computational cost.

2 Related Work

2.1 Calibration of Vision-Language Models

Confidence calibration aims to align predicted confidence with the true probability of correctness. Classical post-hoc calibration methods (Guo et al., 2017), such as temperature scaling, rescale model logits with a global temperature parameter. Although effective for supervised classifiers, directly applying these methods to zero-shot VLMs is challenging because VLMs are often deployed without

labeled validation data from the target task. Moreover, the confidence of CLIP-like models is determined jointly by visual representations and textual class embeddings, making calibration more complex than in standard closed-set classifiers. Recent studies (Baumann et al., 2026; Sharifdeen et al., 2025; Yoon et al., 2024) have investigated calibration for CLIP and other VLMs. Some methods adapt temperature scaling to zero-shot inference, while others examine calibration under prompt learning, fine-tuning, or domain adaptation. However, many existing approaches (Lafon et al., 2025) either rely on labeled validation data, use a global calibration parameter, or focus on adapted VLMs. In contrast, our method performs efficient post-hoc calibration for zero-shot VLMs by assigning an adaptive temperature to each test sample based on its posterior confidence stability.

2.2 Uncertainty Quantification in Vision-Language Models

Uncertainty quantification (Chun et al., 2025; Baumann et al., 2026; Ju et al., 2025; Chun et al., 2022; Chun, 2024) aims to estimate when a model prediction should be trusted. For VLMs, this problem is particularly important because zero-shot predictions are often made under domain shift and without task-specific supervision. Existing studies have explored probabilistic adapters (Upadhyay et al., 2023), conformal prediction (Silva-Rodríguez et al., 2025), Bayesian approximation (Baumann et al., 2026), and multimodal uncertainty estimation to quantify predictive uncertainty and identify unreliable VLM predictions. More closely related to our work, Bayesian approaches (Baumann et al., 2026) approximate posterior distributions over selected components of CLIP-like models and propagate the resulting uncertainty to the joint embedding space. These methods provide a principled way to model uncertainty over image and text embeddings without retraining the full VLM. Our method builds on this posterior embedding distribution to measure the stability of Maximum A Posteriori (MAP) class confidence across posterior samples in embedding space.

2.3 Consistency in Calibration

Consistency (Li et al., 2023; Moon et al., 2020; Huang et al., 2025) is often used for model calibration, which provides a natural indicator of prediction reliability. A reliable prediction should remain stable during training or under different test-

time perturbations. For example, prediction consistency (Li et al., 2023) and correctness (Moon et al., 2020) during training process are proposed for classification calibration. For VLMs, this principle has also been widely used in test-time augmentation test-time adaptation. Test-time augmentation methods (Yoon et al., 2024; Yan et al., 2025) often encourage prediction consistency across augmented views of each test image. However, most existing consistency-based methods estimate stability in the input space or require test-time optimization. They typically rely on multiple augmented image views, repeated image-encoder inference, or additional optimization steps at test time, leading to non-negligible computational overhead. In contrast, our method estimates consistency directly in the embedding space after a single image-encoder forward pass, enabling efficient consistency estimation for calibration.

3 Method

We propose Embedding-space Posterior Calibration (EPC), a lightweight post-hoc calibration framework for zero-shot VLMs. As illustrated in Fig. 1, this section presents EPC in four stages. First, we briefly review how zero-shot VLMs, such as frozen CLIP, produce image and text embeddings. Second, we describe how deterministic VLM embeddings are reformulated as local posterior embedding distributions and how posterior samples are drawn in the embedding space. Third, we introduce the Confidence Stability Score (CSS), which measures the stability of predicted confidence across posterior embedding samples. Finally, we show how CSS is converted into a sample-wise temperature for adaptive confidence calibration.

3.1 CLIP Feature Embedding

Given the i^{th} image $\mathbf{x}_i^{\text{IMG}} \in \mathbb{R}^{p_{\text{IMG}}}$ and the j^{th} text description $\mathbf{x}_j^{\text{TXT}} \in \mathbb{R}^{p_{\text{TXT}}}$, we first extract their modality-specific representations using the image encoder $\phi : \mathbb{R}^{p_{\text{IMG}}} \rightarrow \mathbb{R}^{d_{\text{IMG}}}$ and the text encoder $\psi : \mathbb{R}^{p_{\text{TXT}}} \rightarrow \mathbb{R}^{d_{\text{TXT}}}$ of a pretrained vision-language model, such as CLIP (Radford et al., 2021). The extracted image and text features are then mapped into a shared joint embedding space through linear projection matrices $\mathbf{P} \in \mathbb{R}^{d \times d_{\text{IMG}}}$ and $\mathbf{Q} \in \mathbb{R}^{d \times d_{\text{TXT}}}$, respectively. Formally, the projected image and text embeddings are defined as

$$\mathbf{g}_i = \mathbf{P}\phi(\mathbf{x}_i^{\text{IMG}}), \quad \mathbf{h}_j = \mathbf{Q}\psi(\mathbf{x}_j^{\text{TXT}}), \quad (1)$$

where $\mathbf{g}_i, \mathbf{h}_j \in \mathbb{R}^d$ denote the image and text embeddings in the shared joint embedding space, as illustrated in Fig. 1.

3.2 Embedding-space Posterior Sampling

One core contribution of EPC is embedding-space posterior sampling. It performs the sampling process in the embedding space rather than applying multiple image augmentations in the input space. Instead of naively adding random noise to the embedding, we first reformulate the deterministic CLIP embedding as a posterior embedding distribution

Posterior Embedding Distribution. Following previous work, we apply a Laplace approximation (Scannell et al., 2024; MacKay, 1992) to the image projection layer of CLIP, namely the last linear image projection matrix $\mathbf{P}$, while keeping the image encoder ϕ , the text encoder ψ , and the text projection matrix $\mathbf{Q}$ fixed. This design models uncertainty on the image side only, while treating text embeddings as deterministic class anchors.

Let

$$\mathbf{z}_i^{\text{IMG}} = \phi(\mathbf{x}_i^{\text{IMG}}). \quad (2)$$

Then, the deterministic image embedding $\mathbf{g}_i = \mathbf{P}\mathbf{z}_i^{\text{IMG}}$ is reformulated as

$$p(\mathbf{g}_i | \mathbf{x}_i^{\text{IMG}}, \mathcal{D}_{\text{LA}}) = \mathcal{N}(\boldsymbol{\mu}_i^g, \boldsymbol{\Sigma}_i^g), \quad (3)$$

where

$$\boldsymbol{\mu}_i^g = \mathbf{P}_{\text{MAP}}\mathbf{z}_i^{\text{IMG}}. \quad (4)$$

Here, $\boldsymbol{\Sigma}_i^g$ denotes the input-dependent image embedding covariance induced by the Laplace posterior over the image projection layer $\mathbf{P}$. The Laplace approximation is applied only to $\mathbf{P}$, while the image encoder ϕ , the text encoder ψ , and the text projection matrix $\mathbf{Q}$ are kept fixed. Thus, the text embeddings are treated as deterministic class anchors. $\mathcal{D}_{\text{LA}}$ denotes the data used to construct the Laplace posterior and does not contain labels from the target evaluation benchmarks. Implementation details of the Laplace posterior and covariance construction are provided in Appendix A.4. Compared with the original deterministic CLIP embedding, this posterior formulation allows us to effectively quantify how sensitive a prediction is to structured perturbations in the image embedding space.

Posterior Sampling. Given a posterior embedding distribution, we study how to extract a calibration-specific stability signal from posterior

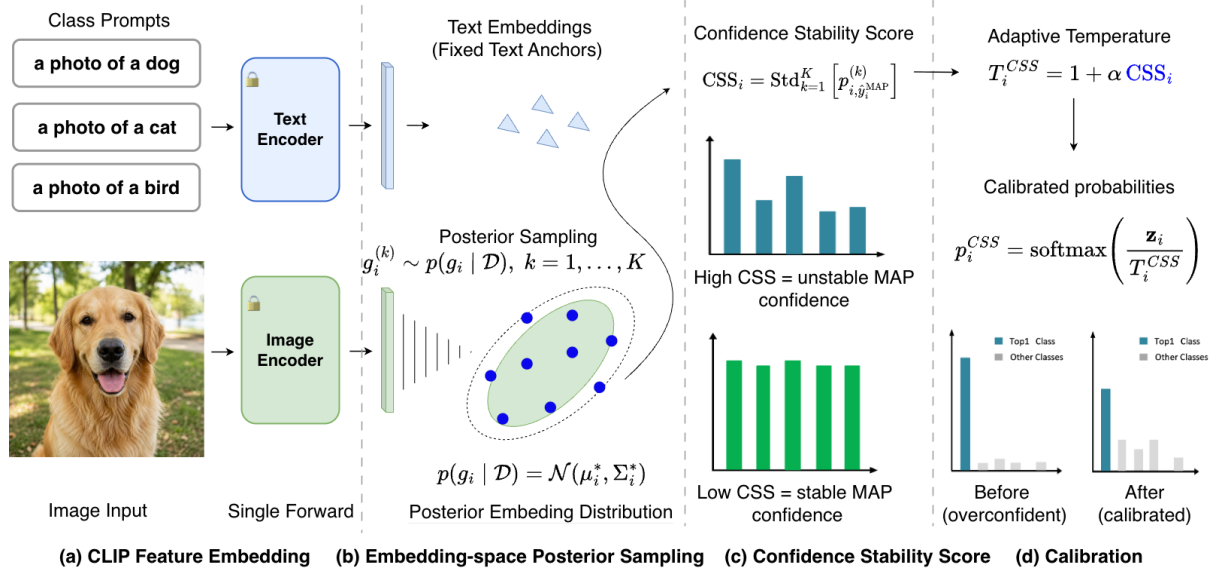


Figure 1: **Overview of the proposed embedding-space posterior calibration for vision-language models.** We sample posterior image embeddings from a frozen vision-language model and measure the stability of the MAP-class confidence across perturbations. The resulting Confidence Stability Score (CSS) is converted into a sample-wise temperature to calibrate the prediction probabilities.

samples. For each image, we draw K embedding samples from its posterior embedding distribution:

$$\mathbf{g}_i^{(k)} \sim p(\mathbf{g}_i | \mathbf{x}_i^{\text{IMG}}, \mathcal{D}_{\text{LA}}), \quad k = 1, \dots, K. \quad (5)$$

These samples can be viewed as structured posterior perturbations around the original embedding, rather than arbitrary random perturbations in the feature space. Compared with arbitrary Gaussian noise in the feature space, our proposed posterior sampling is model-aware and produces multiple embedding variants around the original representation.

Embedding-space posterior sampling has several advantages. First, it avoids repeated image-encoder forward passes required by input-space test-time augmentation, thereby reducing the computational cost. Second, posterior embedding samples provide a natural way to probe whether the model’s prediction is locally stable: a well-calibrated high-confidence prediction should remain confident and consistent under plausible posterior perturbations, whereas an unreliable prediction is more likely to exhibit unstable confidence across samples.

3.3 Confidence Stability Score

Based on embedding-space posterior sampling, we propose the Confidence Stability Score (CSS) to measure the stability of predicted confidence under posterior image-embedding perturbations. The key intuition is that a reliable prediction should remain

stable under plausible samples from the posterior embedding distribution. In contrast, if the predicted confidence changes significantly across posterior samples, the prediction is likely to be uncertain and poorly calibrated.

For the c^{th} text class, we use the deterministic text embedding $\mathbf{h}_c = \mathbf{Q}^\psi(\mathbf{x}_c^{\text{TXT}})$ as its class embedding. The logit for the k^{th} posterior image-embedding sample is computed as

$$\ell_{i,c}^{(k)} = \tau \left(\hat{\mathbf{g}}_i^{(k)} \right)^\top \hat{\mathbf{h}}_c, \quad (6)$$

where τ is the CLIP logit scale and $\hat{\cdot}$ denotes ℓ_2 normalization.

We refer to the class predicted from the posterior mean image embedding as the MAP-class prediction:

$$\hat{y}_i^{\text{MAP}} = \arg \max_c \tau \left(\hat{\boldsymbol{\mu}}_i^g \right)^\top \hat{\mathbf{h}}_c. \quad (7)$$

For each posterior image-embedding sample, we compute the corresponding class probability as

$$p_{i,c}^{(k)} = \frac{\exp(\ell_{i,c}^{(k)})}{\sum_{c'=1}^C \exp(\ell_{i,c'}^{(k)})}. \quad (8)$$

We define CSS as the standard deviation of the MAP-class probability across posterior image-embedding samples:

$$\text{CSS}_i = \text{Std}_{k=1}^K \left[p_{i,\hat{y}_i^{\text{MAP}}}^{(k)} \right]. \quad (9)$$

A larger CSS_i indicates that the confidence of the MAP class is more sensitive to posterior embedding perturbations, suggesting higher uncertainty in the embedding-space prediction. Conversely, a smaller CSS_i means that the model produces stable confidence under posterior sampling, indicating a more reliable prediction.

3.4 Calibration

After obtaining the posterior-sample logits and the corresponding CSS, we perform sample-wise temperature calibration. Instead of using a single global temperature for all test samples, we assign an adaptive temperature to each sample according to its CSS value. This design allows unstable samples to receive stronger confidence smoothing while keeping stable samples less affected.

We compute the average posterior-sample logit for the c^{th} class as

$$\bar{\ell}_{i,c} = \frac{1}{K} \sum_{k=1}^K \ell_{i,c}^{(k)}, \quad c \in \{1, \dots, C\}, \quad (10)$$

where C denotes the number of classes. The resulting average logit vector for the i^{th} image is

$$\bar{\ell}_i = [\bar{\ell}_{i,1}, \bar{\ell}_{i,2}, \dots, \bar{\ell}_{i,C}]. \quad (11)$$

Given the CSS signal, we define the adaptive temperature as

$$T_i^{\text{CSS}} = 1 + \alpha \text{CSS}_i, \quad (12)$$

where α is a scaling coefficient that controls the strength of CSS-based sample-wise calibration. The calibrated probability distribution is then computed as

$$\mathbf{p}_i^{\text{CSS}} = \text{softmax} \left(\frac{\bar{\ell}_i}{T_i^{\text{CSS}}} \right). \quad (13)$$

This formulation directly converts posterior confidence instability into an adaptive calibration temperature. When CSS_i is large, the sample is considered less reliable, and a larger temperature is applied to soften the predicted probability distribution. When CSS_i is small, the temperature remains close to one, preserving the original prediction confidence.

4 Experiments

4.1 Experiment Setup

Datasets. We conduct zero-shot VLM recognition and calibration experiments on six benchmarks,

including fine-grained recognition (FLOWERS-102 (Nilsback and Zisserman, 2008), FOOD-101 (Bossard et al., 2014)), coarse classification (CIFAR-10/100 (Krizhevsky and Hinton, 2009)), distribution shift (IMAGENET-R (Hendrycks et al., 2021)) and video frames (UCF101 (Soomro et al., 2012)). For active few-shot learning, we conduct experiments on the OfficeHome (Venkateswara et al., 2017) dataset, including OH-Art, OH-ClipArt, and OH-Product, as well as the ImageNet variant ImageNet-R.

Evaluation Metrics. Following previous work (Baumann et al., 2026), we report Expected Calibration Error (ECE), Negative Log Predictive Density (NLPD) and Accuracy (ACC) on each dataset to evaluate the performance. Since the main goal of our method is model calibration rather than accuracy improvement, ECE and NLPD are main metrics while ACC is a tracking metric where we aim to maintain the performance without any significant deterioration. NLPD is a balanced metric that reflects both predictive accuracy and probability calibration. The main study uses CLIP-ViT-B/32 (Radford et al., 2021) for zero-shot VLM classification experiments, while the cross-backbone ablation study extends the evaluation to SigLIP-Base (Zhai et al., 2023) using the standard zero-shot prompt template.

4.2 Zero-shot Calibration Performance

We first evaluate whether the proposed method improves the reliability of zero-shot VLM predictions without decreasing recognition accuracy. The experiment compares our method with CLIP (Radford et al., 2021), temperature-scaled CLIP (Guo et al., 2017), TTA (Farina et al., 2024), and BayesVLM (Baumann et al., 2026) on multiple zero-shot recognition benchmarks. We report ECE and NLPD as main metrics while additionally monitoring ACC in Table 1.

As shown in Table 1, our method achieves the best calibration performance (2.05 in mean ECE and 0.98 in NLPD) among all compared methods, while it also achieves the best mean ACC of 75.42 for zero-shot VLM recognition. Compared with original CLIP method (Radford et al., 2021), our proposed method increases the mean ACC from 75.15 to 75.42, and decreases the mean NLPD from 1.07 to 0.98 and mean ECE from the 5.83 to 2.05. Our method reduces the primary metric ECE by 64.8%, which indicates that our method benefits the calibration performance of the most commonly

Table 1: Zero-shot VLM calibration and recognition performance on benchmarks. The best result is marked in blue, and the second-best result is marked in orange. The * denotes tracking metrics where our goal is to maintain the performance without any significant deterioration.

Metrics	Methods	FLOWERS-102	FOOD-101	CIFAR-10	CIFAR-100	IMAGENET-R	UCF101	MEAN
ECE ↓	CLIP (Radford et al., 2021)	6.59	3.91	1.45	6.31	5.20	11.52	5.83
	CLIP (temp. scaling) (Guo et al., 2017)	5.51	4.74	1.88	3.07	4.80	3.61	3.93
	TTA (Farina et al., 2024)	9.63	4.18	2.02	5.27	2.88	11.75	5.96
	BayesVLM (Baumann et al., 2026)	4.22	1.69	0.72	1.92	1.78	3.57	2.32
	Ours	3.12	1.77	0.91	1.69	1.66	3.14	2.05
NLPD ↓	CLIP (Radford et al., 2021)	1.90	0.70	0.21	0.97	1.07	1.59	1.07
	CLIP (temp. scaling) (Guo et al., 2017)	1.67	0.69	0.21	0.94	1.04	1.46	1.00
	TTA (Farina et al., 2024)	1.86	0.67	0.35	1.26	0.90	1.50	1.09
	BayesVLM (Baumann et al., 2026)	1.73	0.68	0.20	0.95	1.03	1.44	1.01
	Ours	1.69	0.67	0.20	0.94	0.99	1.44	0.98
ACC ↑*	CLIP (Radford et al., 2021)	68.99	80.21	93.61	73.76	74.52	59.82	75.15
	CLIP (temp. scaling) (Guo et al., 2017)	68.99	80.21	93.61	73.76	74.52	59.82	75.15
	TTA (Farina et al., 2024)	68.87	81.68	88.54	65.64	78.29	63.07	74.51
	BayesVLM (Baumann et al., 2026)	68.87	80.43	93.62	73.63	74.45	61.43	75.40
	Ours	69.33	80.58	93.54	73.79	74.65	60.61	75.42

Table 2: Ablation on different sampling strategies. Naive Gaussian applies isotropic random noise, while posterior embedding sampling uses local curvature to construct uncertainty-aware perturbations. We report ECE where a lower ECE indicates a better calibration.

Strategy	Flowers-102	Food-101	CIFAR-10	CIFAR-100	ImageNet-R	UCF-101	Mean
Naive Gaussian Sampling	5.04	4.27	1.19	2.34	4.39	2.58	3.30
Posterior Sampling	3.12	1.77	0.91	1.69	1.66	3.14	2.05

used CLIP model. The TTA baseline (Farina et al., 2024) outperforms our method in zero-shot VLM recognition on the FOOD-101, IMAGENET-R and UCF101 datasets, while our method achieves better recognition performance on the remaining datasets. These results indicate that TTA can provide recognition gains in certain zero-shot settings, but its effectiveness varies across datasets. For the calibration, our method improves the NLPD by 10.1% from 1.09 to 0.98, and the ECE by 65.6% from 5.96 to 2.05. Compared to the recent BayesVLM (Baumann et al., 2026), our method reduces the ECE by 11.6% from 2.32 to 2.05, while still maintaining the competitive results in the metrics of ACC and NLPD. Overall, these results support the effectiveness of posterior confidence stability calibration for reliable zero-shot VLM recognition.

4.3 Ablation Studies

We conduct ablation studies to better explain the effectiveness of the proposed embedding-space posterior calibration for zero-shot VLMs.

Naive Gaussian Noise Sampling vs. Posterior Sampling. We first compare posterior sampling with naive Gaussian noise sampling. As shown in Table 2, posterior sampling outperforms naive Gaussian feature noise on five out of six datasets,

reducing the mean ECE from 3.30 to 2.05. Specifically, ECE is reduced from 5.04 to 3.12 on Flowers-102, from 4.27 to 1.77 on Food-101, and from 4.39 to 1.66 on ImageNet-R. These results show that posterior embedding sampling provides substantial improvements on fine-grained recognition and distribution-shift benchmarks. We attribute this improvement to the fact that naive Gaussian noise treats all feature directions as equally uncertain, which may introduce perturbations that are not aligned with the model’s local predictive uncertainty.

Table 3: Ablation on the number of posterior samples. A small value of K is preferred for efficient deployment.

Samples K	ECE ↓	NLPD ↓
$K = 32$	2.06	0.9983
$K = 64$	2.05	0.9983
$K = 128$	2.05	0.9980
$K = 256$	2.07	0.9981

Effect of the Number of Posterior Samples. Table 3 reports the calibration performance under different numbers of posterior samples. As shown in Table 3, our method achieves strong calibration performance across all tested values of K . When

using $K = 128$ posterior samples, our method obtains a mean ECE of 2.05 and the mean NLPD of 0.9980 across six datasets. This result indicates that $K = 128$ provides the strongest overall performance among the evaluated settings. The experiment results also show that the performance is not very sensitive to the exact choice of K . Across $K = 32, 64, 128$, and 256 , the ECE remains within a small range from 2.05 to 2.07, while NLPD score also vary only marginally. Our proposed method maintains stable calibration performance over a wide range of K , which supports the robustness of the proposed posterior feature consistency signal.

Effect of Different Sampling Consistency Signals. We also compare CSS with three alternative signals: LFR, Dual, and Variance. Detailed definitions are provided in Appendix.

Table 4: Ablation on different sampling consistency signals.

Consistency Signals	ECE ↓	NLPD ↓
LFR	2.47	1.001
CSS	2.05	0.998
Dual	13.95	1.705
Variance	40.76	1.589

As shown in Table 4, the proposed CSS signal achieves the best performance across all calibration metrics. Compared with LFR, CSS reduces the mean ECE from 2.47 to 2.05, and improves mean NLPD from 1.001 to 0.998 across six datasets. This indicates that CSS provides a more effective estimate of posterior predictive stability for sample-wise calibration. Unlike LFR, which only captures limited aspects of sampling behavior, CSS directly reflects the confidence stability of posterior predictions and is therefore better aligned with calibration quality. Among all evaluated alternatives, CSS is the only signal that consistently achieves the best ECE and NLPD score. Overall, this ablation indicates that the design of the sampling consistency signal is critical for effective posterior feature calibration.

Disentangling Overall Smoothing from Sample-wise Alignment. We further disentangle two effects of CSS-based calibration: (i) whether posterior sampling provides a useful overall smoothing scale, and (ii) whether the posterior-derived uncertainty is correctly aligned with each individual test

sample, to better understand where the calibration improvement comes from.

We consider two controls. For *Permuted CSS*, we randomly permute the CSS values across test samples and use

$$T_i^{\text{perm}} = 1 + \alpha \text{CSS}_{\pi(i)}, \quad (14)$$

where π denotes a random permutation. This preserves the empirical distribution of sample-wise temperatures, including its mean, variance, and range, while destroying the correspondence between each sample and its posterior-derived CSS. We report the mean and standard deviation over 20 independent permutations.

For *Constant Temperature*, we retain only the average posterior-derived smoothing strength:

$$T^{\text{const}} = 1 + \alpha \frac{1}{N} \sum_{i=1}^N \text{CSS}_i. \quad (15)$$

This removes both sample-wise variation and sample-wise alignment. We use $\alpha = 1$ for all variants and otherwise keep the inference and ECE computation identical.

As shown in Table 5, both Permuted CSS and Constant T substantially improve the mean ECE compared with standard temperature scaling, reaching 2.14 and 2.18, respectively. This indicates that posterior-derived variation primarily provides a useful overall smoothing scale. Preserving the correct sample-wise CSS correspondence further reduces the mean ECE to 2.05, corresponding to a smaller additional benefit over the permuted and constant controls. Importantly, this advantage is not uniform across individual datasets: the permuted control can outperform the aligned CSS temperature on some benchmarks. Therefore, our results do not attribute the full calibration gain to sample-wise posterior alignment. Rather, the main effect arises from efficient posterior-derived smoothing, while sample-wise alignment provides a smaller additional improvement on average.

Effect of Weights of CSS-based Sample-wise Calibration. Table 6 studies the effect of the CSS calibration weight α . The coefficient α controls the strength of the temperature induced by the CSS score. Increasing α from 0.1 to 1.0 reduces mean ECE from 3.06 to 2.05 across six datasets. However, further increasing α leads to worse ECE, suggesting that excessive smoothing can make the predicted confidence under-aligned with empirical

Table 5: ECE comparison for disentangling the effect of overall posterior-derived smoothing from sample-wise CSS alignment.

Dataset	Temp. Scaling	CSS (Ours)	Permuted CSS	Constant T
CIFAR-10	1.88	0.91	0.85 ± 0.03	0.87
CIFAR-100	3.07	1.69	2.12 ± 0.07	2.00
Flowers-102	5.51	3.12	3.55 ± 0.13	3.91
Food-101	4.74	1.77	1.10 ± 0.07	1.11
ImageNet-R	4.80	1.66	1.22 ± 0.16	1.32
UCF101	3.61	3.14	4.03 ± 0.17	3.87
Mean	3.93	2.05	2.14	2.18

Table 6: Ablation on different weights of CSS-based calibration.

Weight α of CSS	ECE $\downarrow$	NLPD $\downarrow$
$\alpha = 0.1$	3.06	1.023
$\alpha = 0.5$	2.29	1.009
$\alpha = 1$	2.05	0.998
$\alpha = 1.5$	2.68	0.993
$\alpha = 2$	3.66	0.992

Table 7: Cross-backbone evaluation using SigLIP on three zero-shot benchmarks.

Dataset	Method	ACC $\uparrow$	NLPD $\downarrow$	ECE $\downarrow$
Flowers-102	SigLIP	82.31	0.88	4.31
	SigLIP (temp. scaling)	82.31	0.84	6.28
	BayesVLM	82.44	0.86	4.87
	Ours	82.49	0.79	3.63
Food-101	SigLIP	88.81	0.38	1.66
	SigLIP (temp. scaling)	88.81	0.40	5.54
	BayesVLM	88.84	0.39	3.49
	Ours	88.95	0.36	2.17
UCF101	SigLIP	59.61	1.90	12.72
	SigLIP (temp. scaling)	59.61	1.77	6.69
	BayesVLM	59.69	1.82	9.60
	Ours	59.82	1.68	6.15

accuracy. Although $\alpha = 2.0$ achieves the lowest NLPD, it substantially increases ECE. In our study, we adopt $\alpha = 1.0$ as the default setting.

Effect of Different Backbones. Table 7 extends the cross-backbone evaluation to Flowers-102, Food-101, and UCF101. On Flowers-102 and UCF101, EPC substantially improves ECE and NLPD while maintaining or slightly improving accuracy. On Food-101, EPC achieves the best ACC and NLPD, but the uncalibrated SigLIP baseline retains the lowest ECE (1.66 versus 2.17 for EPC). Therefore, the results provide evidence that EPC transfers beyond CLIP, while also showing that its ECE improvement is not universal across datasets.

4.4 Application in Downstream Tasks

To further evaluate whether the proposed method can identify informative samples, we conduct active learning experiments in a cross-domain adaptation setting. Given an unlabeled target domain and a limited annotation budget, the goal is to se-

lect a small support set from the candidate pool for fine-tuning the zero-shot vision-language model. We compare our CSS-based selection strategies with logit-based uncertainty baselines, including BALD (Houlsby et al., 2011) and EPIG (Bickford Smith et al., 2023), and with targeted random selection. Specifically, we combine the proposed CSS with standard uncertainty, BALD, and EPIG, denoted as CSS+std, CSS+BALD, and CSS+EPIG, respectively.

As shown in Fig. 2, CSS-based selection improves fine-tuning performance across different subset sizes and target domains. In terms of weighted accuracy, CSS+EPIG and CSS+BALD generally outperform targeted random selection and achieve competitive or better results than their logit-based counterparts. The gains are particularly clear on OH-ClipArt, OH-Product, and IN-R, where CSS-guided sample selection yields higher weighted accuracy as the support set size increases. On OH-Art, CSS-based methods also remain competitive, with CSS+std achieving the best weighted accuracy when the subset size is 75. Across all evaluated domains, CSS-based methods lead to lower NLPD than targeted random selection. In particular, CSS+EPIG achieves the lowest or near-lowest NLPD on OH-Art, OH-ClipArt, and IN-R, suggesting that samples selected using the proposed consistency signal help the model become less overconfident after fine-tuning. On OH-Product, both CSS+BALD and CSS+EPIG obtain clearly lower NLPD than random selection and remain competitive with logit-based EPIG.

4.5 Computation Costs

In addition, we report the inference costs in GFLOPs in Table 8. Our method with posterior embedding samples ($K = 128$) introduces marginal overhead: the cost increases from 8.83 to 8.84 GFLOPs for CLIP-Base and from 47.00 to 47.02 GFLOPs for SigLIP-Base, where the input-space TTA requires 687.78 and 3652.05 GFLOPs.

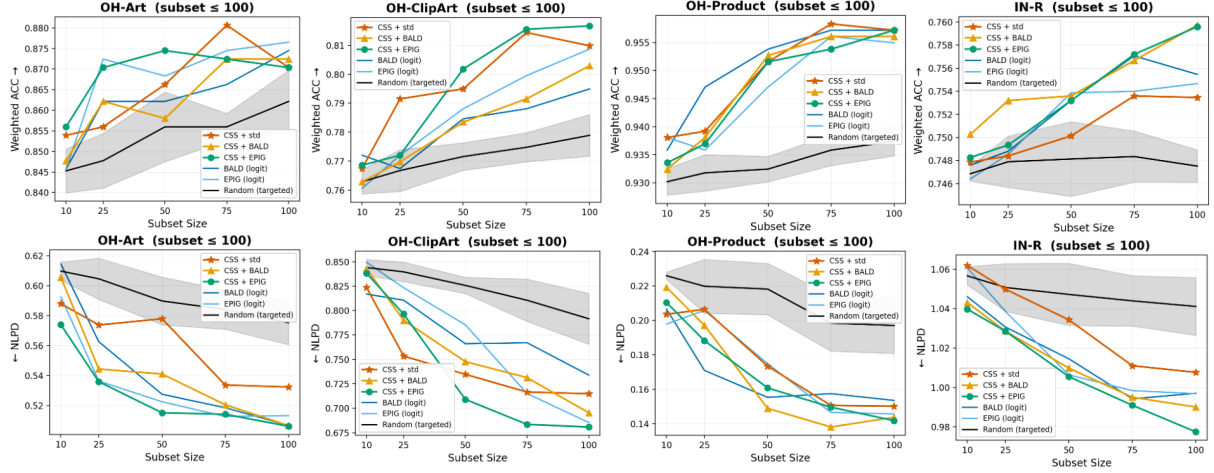


Figure 2: We evaluate active data selection under a limited support-set budget on OH-Art, OH-ClipArt, OH-Product, and IN-R. The first row reports weighted accuracy, where higher values are better, and the second row reports NLPD, where lower values are better. Compared with targeted random selection and logit-based uncertainty baselines, CSS-guided strategies generally achieve stronger fine-tuning performance.

Table 8: Comparison of computational cost across different model scales.

Model	Vanilla	TTA (Farina et al., 2024)	Ours
CLIP-Base	8.83	687.78	8.84
SigLIP-Base	47.00	3652.05	47.02

5 Conclusion

In this paper, we presented EPC, an efficient calibration method for zero-shot VLMs. EPC shifts the sampling process from the input space to the embedding space, avoiding the repeated image-encoder forward passes required by input-space test-time augmentation. By estimating confidence stability from posterior embedding samples, EPC adaptively softens uncertain predictions while preserving stable ones. Experimental results show that EPC improves calibration performance with only marginal computational overhead. These results suggest that embedding-space posterior sampling provides a practical and effective way toward reliable zero-shot vision-language calibration.

Limitations

Approximation and OOD Scope. EPC relies on a local Laplace approximation over the image projection layer. This approximation should not be interpreted as an exact model of all sources of predictive uncertainty. It may become misleading when a test input lies far outside the representation regime covered by the source data, when the true parameter posterior is strongly non-Gaussian

or multimodal, or when uncertainty is dominated by earlier encoder layers. In these cases, the local posterior may remain narrow even for an incorrect prediction, causing EPC to underestimate uncertainty. Although ImageNet-R provides evaluation under a non-trivial shared-label distribution shift, our experiments do not cover semantic or open-set OOD inputs unrelated to the candidate class prompts.

Text-side Uncertainty. The current EPC formulation models uncertainty only on the image side and treats text embeddings as deterministic class anchors. This is a deliberate scope choice that keeps CSS interpretable as a sample-specific image-side reliability signal, but it does not capture prompt sensitivity, semantic ambiguity, or uncertainty in the textual class anchors. EPC can be combined with prompt ensembling by first aggregating multiple prompt embeddings into each class anchor, but explicit joint image-text uncertainty modeling is left for future work.

Acknowledgments

This research work was supported by the Agency for Science, Technology and Research (A*STAR) under its MTC Programmatic Funds (Grant No. M23L7b0021). This work was also supported in part by the Natural Science Foundation of Shandong Province (Grant No. ZR2024QF023), and the Taishan Scholars Program (Grant No. tsqn 202312067).

References

- Anton Baumann, Rui Li, Marcus Klasson, Santeri Mentu, Shyamgopal Karthik, Zeynep Akata, Arno Solin, and Martin Trapp. 2026. Post-hoc probabilistic vision-language models. In *International Conference on Learning Representations (ICLR)*.
- Freddie Bickford Smith, Andreas Kirsch, Sebastian Farquhar, Yarin Gal, Adam Foster, and Tom Rainforth. 2023. Prediction-oriented bayesian active learning. In *International conference on artificial intelligence and statistics*, pages 7331–7348. PMLR.
- Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101—mining discriminative components with random forests. In *European conference on computer vision*, pages 446–461. Springer.
- Sanghyuk Chun. 2024. Improved probabilistic image-text representations. In *International Conference on Learning Representations (ICLR)*.
- Sanghyuk Chun, Wonjae Kim, Song Park, Minsuk Chang, and Seong Joon Oh. 2022. Eccv caption: Correcting false negatives by collecting machine-and-human-verified image-caption associations for ms-coco. In *European Conference on Computer Vision (ECCV)*.
- Sanghyuk Chun, Wonjae Kim, Song Park, and Sangdoo Yun. 2025. Probabilistic language-image pre-training. In *International Conference on Learning Representations (ICLR)*.
- Hao Dong, Moru Liu, Jian Liang, Eleni Chatzi, and Olga Fink. 2025. To trust or not to trust your vision-language model’s prediction. *arXiv preprint arXiv:2505.23745*.
- Matteo Farina, Gianni Franchi, Giovanni Iacca, Massimiliano Mancini, and Elisa Ricci. 2024. Frustratingly easy test-time adaptation of vision-language models. *Advances in Neural Information Processing Systems*, 37:129062–129093.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.
- Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, and 1 others. 2021. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8340–8349.
- Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. 2011. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*.
- Chengsong Huang, Langlin Huang, Jixuan Leng, Jiacheng Liu, and Jiaxin Huang. 2025. Efficient test-time scaling via self-calibration. *arXiv preprint arXiv:2503.00031*.
- Li Ju, Max Andersson, Stina Fredriksson, Edward Glöckner, Andreas Hellander, Ekta Vats, and Prashant Singh. 2025. Exploiting the asymmetric uncertainty structure of pre-trained vlms on the unit hypersphere. *Advances in neural information processing systems*, 38:121317–121335.
- Alex Krizhevsky and Geoffrey Hinton. 2009. Learning multiple layers of features from tiny images.
- Marc Lafon, Yannis Karmim, Julio Silva-Rodríguez, Paul Couairon, Clément Rambour, Raphaël Fournier-Sniehotta, Ismail Ben Ayed, Jose Dolz, and Nicolas Thome. 2025. Vilu: Learning vision-language uncertainties for failure prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17807–17817.
- Chen Li, Xiaoling Hu, and Chao Chen. 2023. Confidence estimation using unlabeled data. In *The Eleventh International Conference on Learning Representations (ICLR)*.
- David JC MacKay. 1992. Information-based objective functions for active data selection. *Neural computation*, 4(4):590–604.
- Jooyoung Moon, Jihyo Kim, Younghak Shin, and Sangheum Hwang. 2020. Confidence-aware learning for deep neural networks. In *international conference on machine learning*, pages 7034–7044. PMLR.
- Maria-Elena Nilsback and Andrew Zisserman. 2008. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR.
- Aidan Scannell, Riccardo Mereu, Paul Chang, Ella Tamir, Joni Pajarinen, and Arno Solin. 2024. Function-space parameterization of neural networks for sequential learning. In *International Conference on Learning Representations*.
- Ashshak Sharifdeen, Muhammad Akhtar Munir, Sanoojan Baliah, Salman Khan, and Muhammad Haris Khan. 2025. O-tp: Orthogonality constraints for calibrating test-time prompt tuning in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19942–19951.
- Julio Silva-Rodríguez, Ismail Ben Ayed, and Jose Dolz. 2025. Conformal prediction for zero-shot models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19931–19941.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Uddeshya Upadhyay, Shyamgopal Karthik, Massimiliano Mancini, and Zeynep Akata. 2023. Problm: Probabilistic adapter for frozen vision-language models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1899–1910.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. 2017. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027.

Rui Yan, Jin Wang, Hongyu Qu, Xiaoyu Du, Dong Zhang, Jinhui Tang, and Tieniu Tan. 2025. [Test-v: Test-time support-set tuning for zero-shot video classification](#). In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 2143–2151. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Hee Suk Yoon, Eunseop Yoon, Joshua Tian Jin Tee, Mark Hasegawa-Johnson, Yingzhen Li, and Chang Yoo. 2024. C-TPT: Calibrated test-time prompt tuning for vision-language models via text feature dispersion. In *International Conference on Learning Representations*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986.

A Appendix

Given an input image x , we draw K posterior feature samples $\{\mathbf{z}^{(k)}\}_{k=1}^K$ from the perturbed image projection layer. Each sampled feature is first normalized and then compared with the normalized text embeddings. Let $\mathbf{t}_c$ denote the text embedding of class c . The sampled logits are computed as

$$\ell_c^{(k)}(x) = s \cdot \left\langle \frac{\mathbf{z}^{(k)}}{\|\mathbf{z}^{(k)}\|_2}, \frac{\mathbf{t}_c}{\|\mathbf{t}_c\|_2} \right\rangle, \quad (16)$$

where s is the CLIP logit scale. The corresponding sampled predictive distribution is

$$p^{(k)}(x) = \text{softmax} \left(\ell^{(k)}(x) \right). \quad (17)$$

For the MAP image embedding $\mathbf{z}^{(0)}$, we similarly compute

$$\ell_c^{(0)}(x) = s \cdot \left\langle \frac{\mathbf{z}^{(0)}}{\|\mathbf{z}^{(0)}\|_2}, \frac{\mathbf{t}_c}{\|\mathbf{t}_c\|_2} \right\rangle, \quad (18)$$

and define the MAP prediction as

$$\hat{y}^{(0)} = \arg \max_c \ell_c^{(0)}(x). \quad (19)$$

A.1 Label Flip Rate.

LFR measures how frequently the posterior samples change the top-1 prediction compared with the MAP prediction:

$$s_{\text{LFR}}(x) = \frac{1}{K} \sum_{k=1}^K \mathbf{1} [\hat{y}^{(k)} \neq \hat{y}^{(0)}], \quad (20)$$

where

$$\hat{y}^{(k)} = \arg \max_c \ell_c^{(k)}(x). \quad (21)$$

A larger LFR indicates stronger label-level instability under posterior perturbations.

A.2 Monte Carlo Logit Variance.

Variance measures the Monte Carlo variance of sampled logits on the MAP top-1 class. Let

$$\bar{\ell}_c(x) = \frac{1}{K} \sum_{k=1}^K \ell_c^{(k)}(x) \quad (22)$$

be the mean sampled logit for class c . The class-wise unbiased Monte Carlo variance is

$$v_c(x) = \frac{1}{K-1} \sum_{k=1}^K \left(\ell_c^{(k)}(x) - \bar{\ell}_c(x) \right)^2. \quad (23)$$

The top-1 logit variance used as the calibration signal is

$$s_{\text{Var}}(x) = v_{\hat{y}^{(0)}}(x). \quad (24)$$

Unlike CSS, which measures probability-level confidence stability, this signal operates directly in the logit space.

A.3 Dual Signal.

Dual is a zero-label fusion score that combines multiple sampling-based instability signals. Given LFR, CSS, and the variance signal $s_{\text{Var}}(x)$, we first standardize each signal over the evaluation set. For a generic signal $s(x)$, its standardized form is

$$\tilde{s}(x) = \frac{s(x) - \mu_s}{\sigma_s + \epsilon}, \quad (25)$$

where μ_s and σ_s denote the mean and standard deviation of $s(x)$ over all samples, and ϵ is a small constant for numerical stability.

Table 9: ECE under different CSS-to-temperature mappings.

Mapping	CIFAR-10	CIFAR-100	Flowers-102	Food-101	ImageNet-R	UCF101	Mean
Quadratic ($a = 1, b = 1$)	0.99	1.83	3.18	2.22	2.07	2.94	2.21
Quadratic ($a = 0.5, b = 0.5$)	0.88	2.33	3.18	0.75	1.24	4.94	2.22
Exponential ($a = 0.75$)	0.86	1.86	2.99	1.19	0.97	4.02	1.98
Exponential ($a = 1$)	0.94	1.79	2.99	2.01	1.93	2.72	2.06
Sigmoid ($a = 1, b = 1$)	0.91	2.54	3.13	0.59	1.14	5.35	2.28
Sigmoid ($a = 0.5, b = 1$)	0.98	3.32	3.28	0.55	1.94	6.46	2.75
Linear ($a = 1$)	0.91	1.69	3.12	1.77	1.66	3.14	2.05

The Dual score is then defined as a weighted sum:

$$s_{\text{Dual}}(x) = \lambda_{\text{LFR}} \tilde{s}_{\text{LFR}}(x) + \lambda_{\text{CSS}} \tilde{s}_{\text{CSS}}(x) + \lambda_{\text{Var}} \tilde{s}_{\text{Var}}(x). \quad (26)$$

In our implementation, unless otherwise specified, we use equal weights:

$$\lambda_{\text{LFR}} = \lambda_{\text{CSS}} = \lambda_{\text{Var}} = 1. \quad (27)$$

This signal is designed to aggregate label-level instability, confidence-level instability, and posterior logit dispersion without using any ground-truth labels.

A.4 Laplace Posterior Implementation

We apply a last-layer Laplace approximation to the image projection matrix P , while keeping the image encoder ϕ , text encoder ψ , and text projection matrix Q fixed. We use the precomputed curvature factors released by BayesVLM, following its Hessian-estimation protocol.

The curvature estimation uses up to 327,680 image-text pairs sampled from 60 LAION-400M WebDataset shards. These samples come from the heterogeneous web-scale image-caption distribution of LAION-400M and are independent of the downstream evaluation benchmarks. Although BayesVLM estimates curvature factors for both image and text projection layers, EPC uses only the image-side curvature factors and the associated optimized prior precision.

Given the posterior covariance of the image projection parameters, the input-dependent embedding covariance Σ_i^g is obtained by propagating the image feature z_i^{IMG} through the linear projection posterior. The resulting posterior is constructed once per backbone and reused for all downstream datasets.

Importantly, no target-domain images, downstream labels, or target-domain validation data are used to construct the posterior or to select the EPC hyperparameters.

A.5 Alternative CSS-to-Temperature Mappings

We further evaluate whether EPC depends critically on the linear CSS-to-temperature mapping. Let $s = \text{CSS}_i$. We compare the following monotonic mappings:

$$T_{\text{linear}}(s) = 1 + as, \quad (28)$$

$$T_{\text{quad}}(s) = 1 + as + bs^2, \quad (29)$$

$$T_{\text{sigmoid}}(s) = 1 + a(2\sigma(bs) - 1), \quad (30)$$

$$T_{\text{exp}}(s) = \exp(as). \quad (31)$$

All mappings satisfy $T(0) = 1$ under the evaluated parameterizations.

The exponential mapping with $a = 0.75$ obtains the lowest mean ECE of 1.98, while the default linear mapping remains competitive at 2.05. These results indicate that EPC is not tied to a particular functional form. We retain the linear mapping because of its simplicity, single-parameter formulation, and ease of deployment, rather than because it is theoretically optimal.