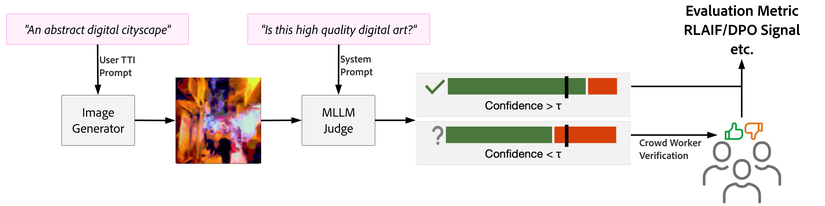
MMB: Calibrating MLLM-as-a-judge via Multimodal Bayesian Prompt Ensembles

Eric Slyman — Adobe
Mehrab Tanjim — Adobe
Kushal Kafle — Adobe
Stefan Lee — Oregon State Univ.

TLDR: We present MMB, a dynamic prompt ensemble for improving MLLM calibration, and use it to reduce SOTA calibration error on HPSv2 by up to 43%

Problem: When should we trust MLLM-as-a-Judge Outputs?

Calibration aligns confidence with correctness, preventing overconfident mistakes and surfacing uncertainty so practitioners can set thresholds, defer to humans, and make risk-aware decisions.



Prior Work: Bayesian Prompt Ensembles (BPE)

BPE [1] uses a small validation set to select a variational inference-based weighted average over candidate prompts to improve calibration on the ensemble model's predictions.

$$p(label|image) = \int llm(image, prompt)p(prompt)$$

Baysean Prompt Ensemble (BPE) Objective Function

$$\arg\max_{\mathbf{w}} \underbrace{\sum_a w_a \left[\sum_{j=1}^{M} \log p(y_j^* \mid x_j, a)\right]}_{\text{Maximize Likelihood}} - \underbrace{\sum_a w_a \log w_a}_{\text{Maximize Entropy}}$$

$$\underbrace{a = \{a_i\}_{i=1}^{n}}_{\text{Finite Prompt Set}} \quad \underbrace{w_a = \frac{q(a)}{NC} \; s.t. \; w_a \geq 0, \sum_a w_a = 1}_{\text{Weights on Probability Simplex}} \quad \underbrace{p(a) \approx C}_{\text{Uniform Prior}}$$

Limitation: What if prompts aren't equally relevant *a priori*?

The VI procedure of BPE assumes all prompts are equally relevant a priori. However, recent work has shown that small domain-adapting variations can result in a large impact on performance [2].

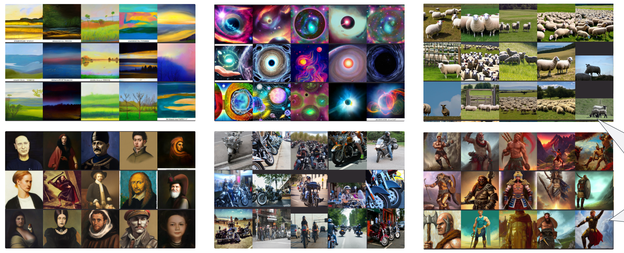


Image sets from K-Means clustering on HPSv2.

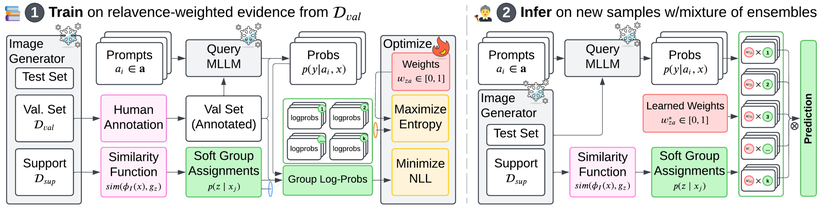
Do we expect the same prompts to have similar performance on all sets?

Plausible Persona: Wildlife Photographer

Plausible Persona: Character Concept Artist

Solution: Multimodal Mixture of Bayesian Prompt Ensembles

We propose MMB, a dynamic BPE that weights prompts based on the inputted image's relevance.



To address the conditional nature of performance between prompts and salient image sets (e.g., semantics, style), MMB introduces a random variable for image content: group.

$$p(label|image) = \int\int llm(image, prompt)p(prompt|group)p(group|image)$$

Assuming prompt ⫫ image|group and label ⫫ group|prompt

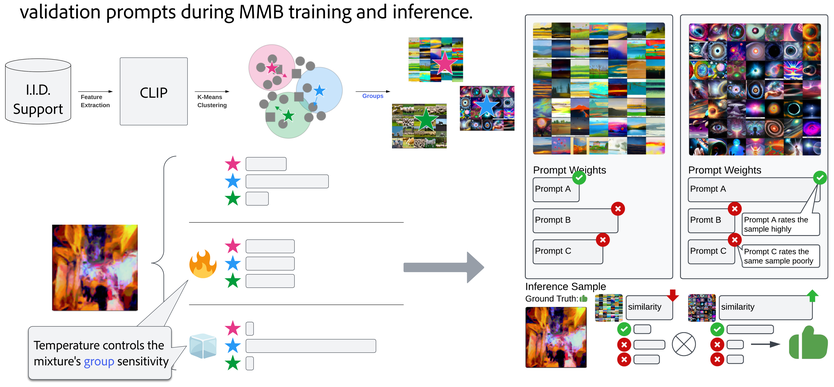
Multimodal Mixture of BPE (MMB) Objective Function

See paper for derivation and reparameterization

$$p(z|x) \propto e^{sim(image,\ group)\ /\ temp}$$

$$\arg\max_{\mathbf{w}} \sum_{j=1}^{M} \sum_z \underbrace{p(z|x_j)}_{\text{Soft Group Assignment}} [\underbrace{\sum_a w_{za} \log p(y_j^*|x_j, a)}_{\text{Per-Group Log Likelihood}} - \underbrace{\sum_a w_{za} \log w_{za}}_{\text{Per-Group Entropy Regularizer}}]$$

Clustering CLIP-embedded image generations affords a *"free"* support set for anchoring validation prompts during MMB training and inference.



Result: MMB Significantly Improves Calibration on HPSv2.

We perform n=150 random trials across varied setting on MMB (ours), BPE (sota) , and baseline ensemble methods for predicting human preference scores on HPSv2 [3].

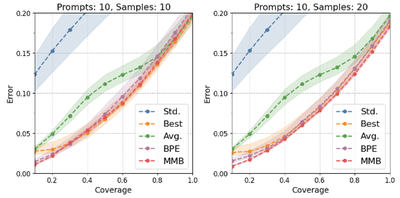
Methods	Std. Best, Average, BPE, MMB
Prompts	5, 10, 20
Samples	5, 10, 20, 50
Clusters	4, 8, 16, 32, 64
Train Seed	3 unique random values
Data Seed	50 unique random values
Cluster Seed	5 unique random values

*"You are a **landscape artist** skilled in assessing lighting, color and composition [...]"*

*"You are a **photographer skilled** in assessing lighting, focus, and style [...]"*

In our experiments, MMB demonstrates strong low-variance coverage-error trade off (*left*). MMB also reduces calibration error from BPE at 95% confidence according to a permutation test (*right*).

Bucket-Averaged Coverage vs. Area Curves

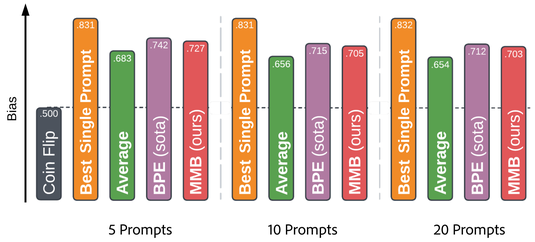


MMB (ours) Calibration W.r.t BPE (sota)

	5 Prompts	10 Prompts	20 Prompts
Calibration Error (avg.) ↓	-10%	-20%	-30%
Calibration Error (max) ↓	-22%	-34%	-43%
AUC PR ↑	No Diff.	+1%	+1%

Result: MMB Ameliorates Fairness Disparities W.r.t. BPE.

We construct a synthetic preference set on MJBench [4] where (un)preferred samples may be drawn for various occupations with controlled demographic identifiers to study preferential bias.



[1] Bayesian Prompt Ensembles: Model Uncertainty Estimation for Black-Box Large Language Models (Tonolini et. al 2024)
[2] Can LLM be a Personalized Judge? (Dong et al., 2024)
[3] Human Preference Score v2: A Solid Benchmark for Evaluating Human Preferences of Text-to-Image Synthesis (Wu et. al 2023)
[4] MJBench: Is your multimodal reward model really a good judge for text-to-image generation? (Chen et. al 2024)