Calibrating MLLM-as-a-judge via Multimodal Bayesian Prompt Ensembles
Supplementary Material

A. Additional Experimental Details

	Factor	Levels/Value	Notes
Exps.	prompts	5, 10, 20	Num. prompts in $\boldsymbol{a}$
	samples	5, 10, 20, 50	Num. samples in $\mathcal{D}_{val}$
	clusters	4, 8, 16, 32, 64	Num. clusters K
Seeds	train	3 Unique	Seed for training
	data	50 Unique	Seed for sampling $\mathcal{D}_{val}$
	cluster	5 Unique	Seed for sampling $\mathcal{D}_{sup}$
Clustering	method	Spherical	KMeans version
	samples	$256 \times K$	Num. clustering samples
	init	3	Num. random inits
	niter	1,000	Num. training iterations
Optimization	Optim	L-BFGS	Optimizer
	lr	0.01	Learning rate
	history	50	History size
	max iter	100	Max iterations
	search	strong wolfe	Search func
ϕ_I	Model	CLIP-ViT-B16	[52]
	Weights	laion2b_s34b_b88k	

Table 5. Summary of experimental configurations..

See table Tab. 5 for details on experimental factors, clustering configuration, and other hyperparameters.

B. Generating Instruction Prompts

We employ a variety of methods to contruct our prompt set $\boldsymbol{a}$. We both manually, and with the aid of GPT, construct lists of personas, prompt templates, and task instruction criteria. We also take these original templates and create "*augmented*" versions by flipping the order of inputs in the template and changing the response glyph (*e.g.* "*A)*" *vs.* "*1.*"). We provide a sample prompt for reference —

```
You are a technical expert at
 ↪ evaluating 'text-to-image'
 ↪ alignment and aesthetics. Your task
 ↪ is to assess the quality of two
 ↪ images generated from the same
 ↪ prompt. The criteria for evaluation
 ↪ are as follows:

Image Quality - The image should have a
 ↪ well-balanced composition with
 ↪ clear framing and focal point,
 ↪ effective brightness and contrast
 ↪ in lighting, appealing color
 ↪ harmony and saturation, sharp focus
 ↪ with visible fine details and
 ↪ minimal digital noise, and be high
 ↪ resolution without pixelation.

Image Artifacts - The image should not
 ↪ have any obvious artifacts
 ↪ including excessive blur,
 ↪ occlusion, warping, or other
 ↪ issues.

Be objective in your evaluation, do not
 ↪ consider attributes like age, race,
 ↪ gender, or other demographic
 ↪ information. Respond with a single
 ↪ letter only.

[IMAGE 1]
[IMAGE 2]
Caption: [INPUT PROMPT]

Which of the two images do you prefer?
A) I prefer the first image.
B) I prefer the second image.
```

C. Derivation of Multimodal Mixture-of-Bayesian Prompt Ensembles

$$
\begin{aligned}
\log p(y|x) &= \log \sum_z \sum_a p(y, a, z|x) \\
&= \log \sum_z \sum_a p(y|x, a) p(a|z) p(z|x) \\
&= \log \sum_z \sum_a p(y|x, a) p(a|z) p(z|x) * \frac{q(a|z)}{q(a|z)} \\
&= \log E_{a\sim q(a|z), z\sim p(z|x)} \left[p(y|x, a) \frac{p(a|z)}{q(a|z)} \right] \\
&\geq E_{a\sim q(a|z), z\sim p(z|x)} \left[\log p(y|x, a) + \log \frac{p(a|z)}{q(a|z)} \right] \\
&= E_{a\sim q(a|z), z\sim p(z|x)} \left[\log p(y|x, a) \right] - E_{a\sim q(a|z), z\sim p(z|x)} \left[\log \frac{q(a|z)}{p(a|z)} \right] \\
&= \left[\sum_z p(z|x) \sum_a q(a|z) \log p(y|x, a) \right] - \left[\sum_z p(z|x) KL(q(a|z)||p(a|z)) \right] \\
&= \sum_z p(z|x) \left[\left[\sum_a q(a|z) \log p(y|x, a) \right] - KL\left(q(a|z)||p(a|z)\right) \right]
\end{aligned}
$$

Assuming uniform priors for $p(a|z)$ and parameterizing $q(a|z) = w_{za}$, we can write the training objective for **M**ultimodal **M**ixture-of-**B**ayesian Prompt Ensembles to maximize this lower-bound on the log-likelihood of all M datapoints in $\mathcal{D}_{val}$ as:

$$
\arg\max_{\mathbf{w}} \sum_{j=1}^{M} \sum_z p(z|x_j) \left[\sum_a w_{za} \log p(y_j^*|x_j, a) - \sum_a w_{za} \log w_{za} \right] \tag{11}
$$

D. Additional Experiment Configurations

prompts	samples	Expected Calibration Error (↓)					Max Calibration Error (↓)					AUC Precision-Recall (↑)				
		4	**8**	**16**	**32**	**64**	**4**	**8**	**16**	**32**	**64**	**4**	**8**	**16**	**32**	**64**
5	5	**.113**	**.113**	**.113**	**.113**	**.113**	**.245**	**.244**	**.245**	**.245**	**.246**	.834	.835	.835	.835	**.836**
	10	**.107**	.107	.108	.108	**.108**	**.239**	**.239**	**.239**	**.239**	**.239**	.837	.837	.838	.838	**.838**
	20	**.108**	.108	.108	.108	**.109**	**.241**	**.240**	**.241**	.241	**.242**	.837	.837	.838	.838	**.838**
	50	**.107**	**.107**	.107	.107	**.107**	**.235**	**.235**	.236	.236	**.237**	.839	.839	.839	.839	**.839**
10	5	**.092**	.092	.093	.092	**.093**	**.199**	**.200**	.201	.201	**.202**	.842	.842	.843	.844	**.844**
	10	**.094**	**.094**	.095	.095	**.095**	**.194**	**.195**	.196	.196	**.197**	.843	.843	.844	.844	**.845**
	20	**.090**	**.090**	.091	.091	**.091**	**.188**	**.188**	.189	.189	**.189**	.844	.844	.845	.845	**.845**
	50	**.088**	**.088**	.088	.088	**.089**	**.187**	**.187**	.188	.188	**.188**	.845	.845	.845	.845	**.845**
20	5	**.078**	.079	.080	.080	**.080**	**.170**	.172	.172	.172	**.173**	.845	.846	.847	.847	**.848**
	10	**.081**	**.081**	.082	.082	**.082**	**.168**	**.169**	**.169**	**.169**	**.169**	.846	.847	.847	.848	**.848**
	20	**.080**	**.080**	.080	.080	**.081**	**.159**	**.159**	**.159**	**.160**	**.161**	.847	.847	.848	.848	**.848**
	50	**.076**	.076	.076	.076	**.077**	**.153**	**.153**	**.153**	**.154**	**.154**	.848	.848	.849	.849	**.849**

Table 6. Expected calibration error. Lower is better. Multiple cluster sizes. FDR controlled with Benjamini-Yekutieli [5]. HPSv2.

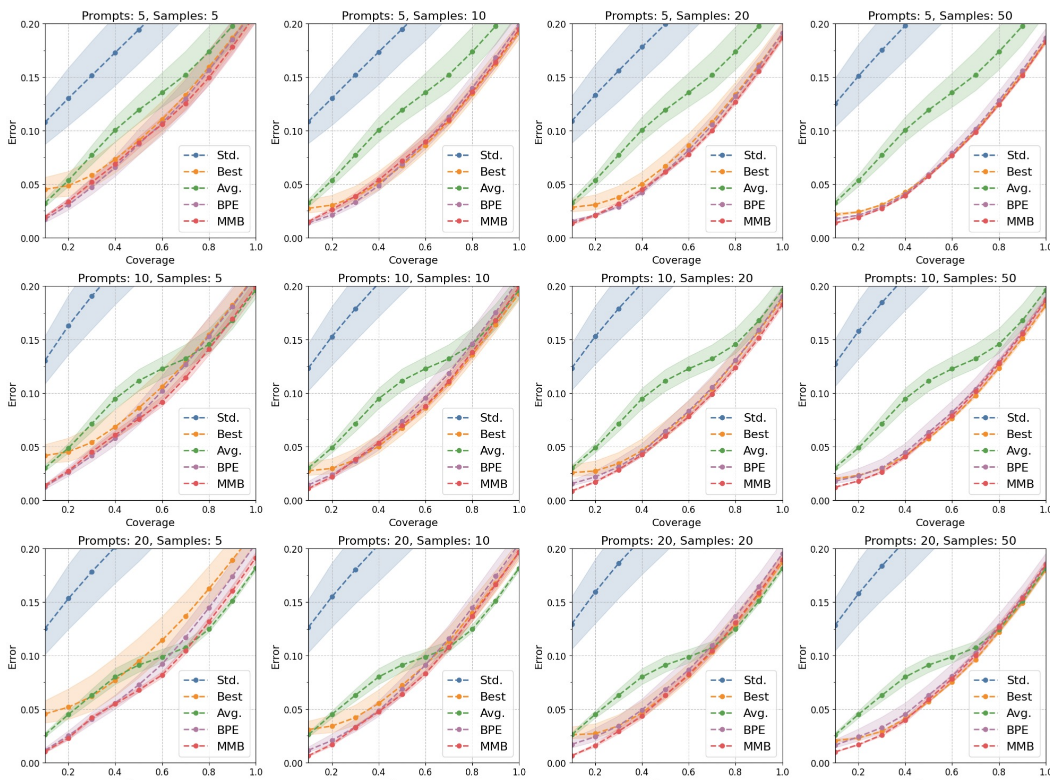
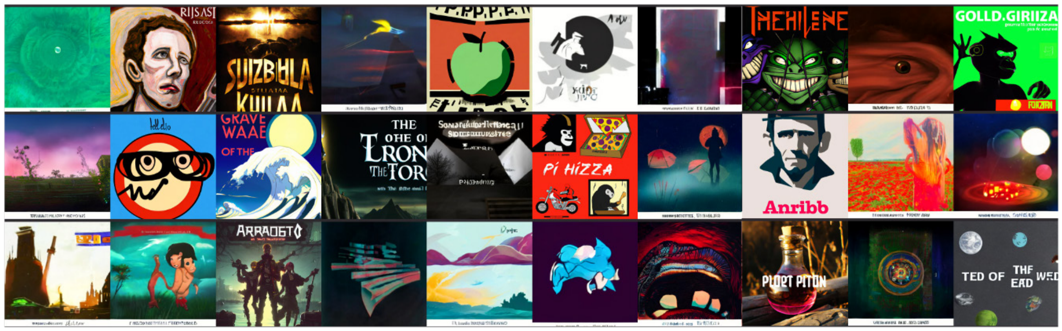
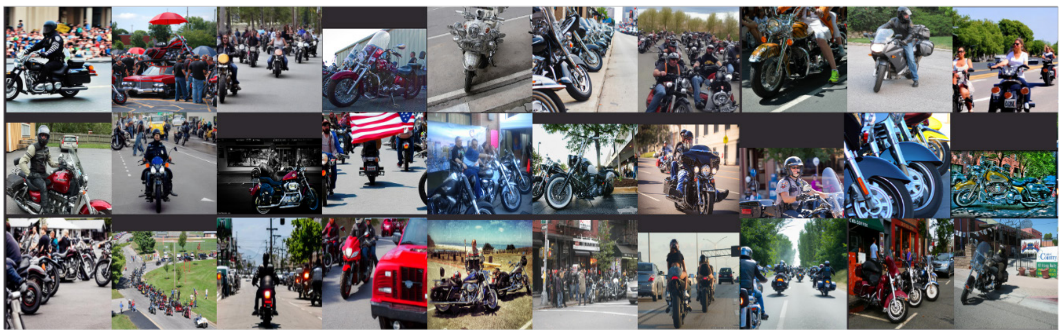


Figure 5. Error-coverage curves for different models. HPSv2.

E. Additional Qualitative Examples



(a) A non-cohesive cluster that results in near-average weights across prompts due to low-validation sample match.

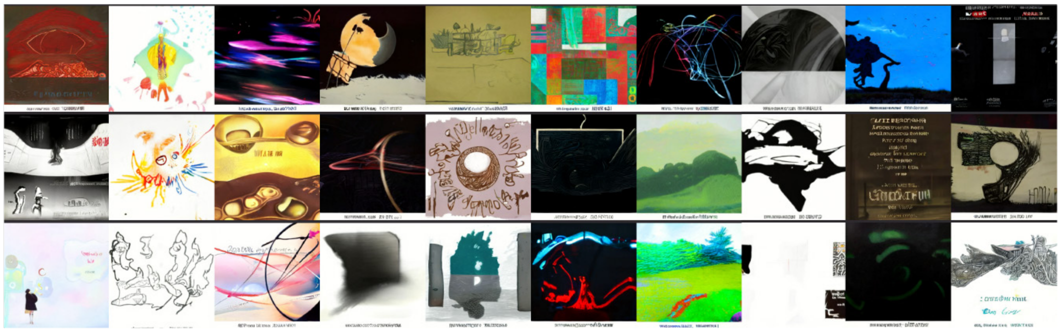


(b) A cohesive cluster which can be matched with validation samples, but does not have any highly weighted prompts.

(c) "*You are a **photographer** skilled in assessing lighting, focus, and overall image sharpness [...]*"

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

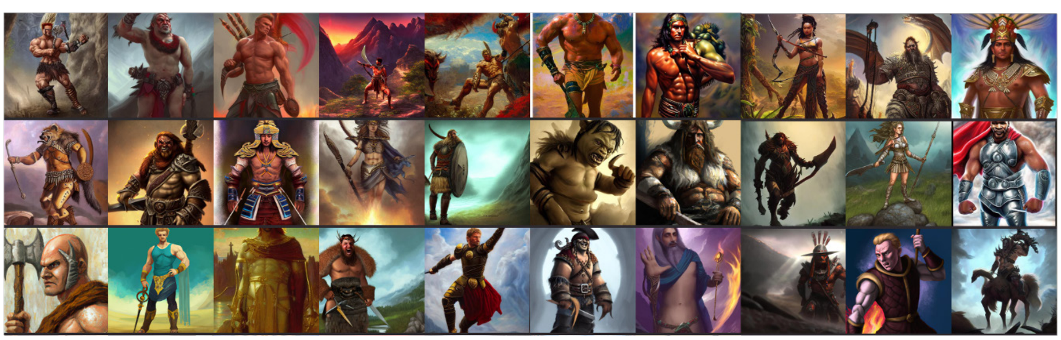
Figure 6. Image clusters and their corresponding highest weighted prompts (or lack thereof) when using K=64, N=200.



(a) A non-cohesive cluster that results in near-average weights across prompts due to low-validation sample match.

(b) "*You are a* ***graphic designer*** *with experience in visual clarity and technical image quality [...]*"

(c) "*You are an* ***art historian*** *with a keen eye for visual composition and color balance [...]*"



(d) "*You are an* ***AI ethics specialist*** *focusing on ensuring accurate and unbiased image representations [...]*"

Figure 7. Image clusters and their corresponding highest weighted prompts (or lack thereof) when using K=64, N=200.

F. MJBench Synthetic Preference Examples

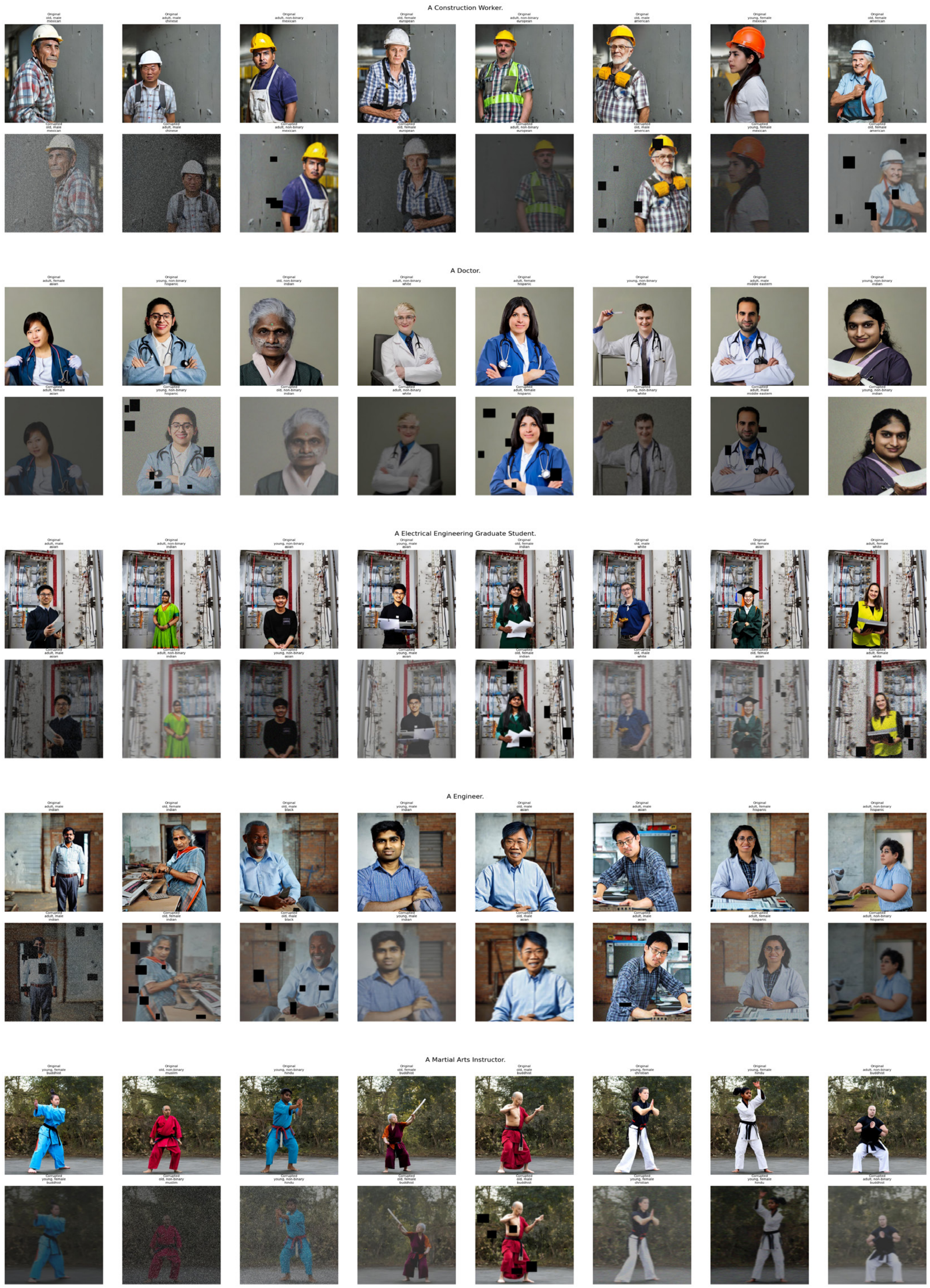


Figure 8. Synthetic images preference pairs generated from MJBench-Bias.