Multimodal LLMs as Customized Reward Models for Text-to-Image Generation

Supplementary Material

A. Implementation Details

We fine-tune Phi-3.5-vision on the training set introduced above via the standard pairwise ranking loss or 2-dimensional GPM [4] loss when only preference is necessary, such as the evaluation in MJ-Bench. We finetune the safety model on UnsafeBench via cross-entropy loss. We train LLaVA-Reward with batch size 8 and gradient accumulation size 4 on 4 NVIDIA A6000 GPUs for one epoch, with a learning rate of 2e-4. We fine-tuned the LoRA adapter with rank 128, the visual projector, and the SkipCA value head described in Section 3, with other parameters frozen.

B. Limitations

Multimodal LLMs have shown impressive performance in image understanding and multimodal reasoning. However, the multimodal reward model still suffers from the absence of high-quality training data, which can lead to issues such as reward hacking [3]. Currently, LLaVA-Reward is adapted using preference data derived from generations of a limited set of models. In future work, we aim to enhance the robustness and capacity of LLaVA-Reward by incorporating more comprehensive and diverse training data.

C. Additional Quantitative and Visual Results

Besides the GenEval scores in Tab. 4, we show the overall performance of 3 reward models in GenEval and DrawBench using SD v2.1 and SDXL here in Tab. 9. Additional visual examples of diffusion inference-time scaling via FK steering (SDXL) using LLaVA-Reward and baselines are shown here in Fig. 4, Fig. 5 and Fig. 6.

References

[1] Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, et al. Inference-time scaling for diffusion models beyond scaling denoising steps. *arXiv preprint arXiv:2501.09732*, 2025. 2, 3, 4

[2] Raghav Singhal, Zachary Horvitz, Ryan Teehan, Mengye Ren, Zhou Yu, Kathleen McKeown, and Rajesh Ranganath. A general framework for inference-time scaling and steering of diffusion models. *arXiv preprint arXiv:2501.06848*, 2025. 1

[3] Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking, 2025. 1

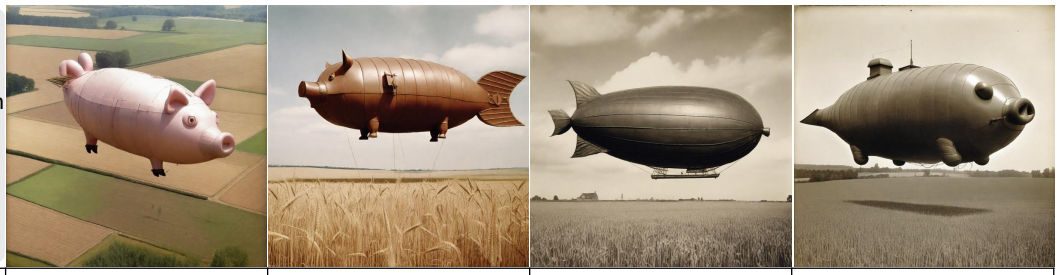
[4] Yifan Zhang, Ge Zhang, Yue Wu, Kangping Xu, and Quanquan Gu. General preference modeling with preference representations for aligning language models. *arXiv preprint arXiv:2410.02197*, 2024. 1

Table 9. Diffusion inference-time scaling results of FK steering [2] using the prompts of GenEval and DrawBench. We report the mean performance on each metrics. We use **bold** and underline to indicate the best and second-best results.

Dataset	Reward Model	Model	HPSv2	ClipScore	VQAscore	LLaVA-Reward	ImageReward
GenEval	None	SD v2.1	0.285	0.300	0.620	−0.045	0.303
	CLIPScore		0.292	**0.314**	0.663	0.018	0.660
	ImageReward		0.298	0.312	0.737	0.078	**1.179**
	LLaVA-Reward		**0.301**	0.311	**0.743**	**0.226**	0.976
	None	SDXL	0.307	0.303	0.681	0.180	0.677
	CLIPScore		0.316	0.317	0.694	0.252	1.046
	ImageReward		0.316	0.316	0.718	0.270	**1.247**
	LLaVA-Reward		**0.323**	**0.319**	**0.721**	**0.397**	1.073
DrawBench	None	SD v2.1	0.261	0.296	0.649	0.023	0.849
	CLIPScore		0.267	0.314	0.706	0.070	1.042
	ImageReward		0.274	0.312	0.702	0.124	**1.332**
	LLaVA-Reward		**0.276**	**0.321**	**0.729**	**0.192**	1.147
	None	SDXL	0.276	0.327	0.826	0.244	1.309
	CLIPScore		0.282	0.348	**0.890**	0.308	1.540
	ImageReward		0.286	0.343	0.881	0.299	**1.627**
	LLaVA-Reward		**0.288**	**0.350**	0.887	**0.352**	1.542

Prompt

An old photograph of a 1920s airship shaped like a pig, floating over a wheat field.



	w/o inference-time scaling	FK steering (LLaVA-Reward)	FK steering (ImageReward)	FK steering (CLIPscore)
HPSv2	0.298	**0. 340**	0. 286	0.288
CLIPscore	0.288	0. 339	0.289	**0.342**
VQAScore	0.427	0.739	0.519	**0.909**
ImageReward	1.649	**1.961**	1.524	1.802
LLM Grader	7.750	**8.400**	7.000	8.300
LLaVA-Reward (Ours)	0.250	**0.602**	0.262	0.463

Prompt

a photo of a bench and a snowboard.

	w/o inference-time scaling	FK steering (LLaVA-Reward)	FK steering (ImageReward)	FK steering (CLIPscore)
HPSv2	0.248	**0. 290**	0. 268	0.252
CLIPscore	0.291	**0. 332**	0.324	0. 314
VQAScore	0.967	0.969	**0.981**	0.649
ImageReward	0.121	**1.775**	1.540	-1.009
LLM Grader	7.950	**8.050**	7.750	7.500
LLaVA-Reward (Ours)	0.337	**0.566**	0.291	0. 279

Prompt

a photo of a black kite and a green bear.

	w/o inference-time scaling	FK steering (LLaVA-Reward)	FK steering (ImageReward)	FK steering (CLIPscore)
HPSv2	0.312	**0.345**	0.286	0.285
CLIPscore	0.275	**0.337**	0.291	0.311
VQAScore	0.139	**0.770**	0.211	0.171
ImageReward	0.147	**1.652**	1.537	0.183
LLM Grader	5.400	**7.160**	6.480	5.760
LLaVA-Reward (Ours)	-0.306	**0.460**	0.083	-0.193

Figure 4. Examples of diffusion inference-time scaling via FK steering (SDXL) using 5 different reward models with the prompt from GenEval. The LLM grader is conducted using GPT-4o with prompts from Ma et al. [1].

Prompt

a photo of a blue handbag and a white cell phone.

	w/o inference-time scaling	FK steering (LLaVA-Reward)	FK steering (ImageReward)	FK steering (CLIPscore)
HPSv2	0.292	0.311	0.286	**0.317**
CLIPscore	0.291	0.302	0.292	**0.342**
VQAScore	0.942	**0.971**	0.909	0.899
ImageReward	1.696	**1.798**	1.729	1.740
LLM Grader	7.880	**8.040**	7.480	7.560
LLaVA-Reward (Ours)	0.412	**0.512**	0.503	0.299

Prompt

a photo of a train above a potted plant.

	w/o inference-time scaling	FK steering (LLaVA-Reward)	FK steering (ImageReward)	FK steering (CLIPscore)
HPSv2	0.315	0.297	**0.321**	0.296
CLIPscore	0.296	**0.331**	0.299	0.282
VQAScore	0.904	0.888	0.909	**0.971**
ImageReward	1.012	**1.540**	0.870	0.977
LLM Grader	**7.480**	7.400	6.240	6.080
LLaVA-Reward (Ours)	0.363	**0.451**	0.389	0.297

Prompt

a photo of a brown oven and a purple train.

	w/o inference-time scaling	FK steering (LLaVA-Reward)	FK steering (ImageReward)	FK steering (CLIPscore)
HPSv2	0.266	**0.295**	0.275	0.276
CLIPscore	0.273	0.312	0.311	**0.328**
VQAScore	0.176	**0.769**	0.245	0.253
ImageReward	-1.305	**0.445**	-1.334	-1.466
LLM Grader	4.800	**6.040**	5.480	5.160
LLaVA-Reward (Ours)	-0.125	**0.178**	-0.400	-0.217

Figure 5. Examples of diffusion inference-time scaling via FK steering (SDXL) using 5 different reward models with the prompt from GenEval. The LLM grader is conducted using GPT-4o with prompts from Ma et al. [1].

Prompt

A storefront with 'Google Brain Toronto' written on it.

	w/o inference-time scaling	FK steering (LLaVA-Reward)	FK steering (ImageReward)	FK steering (CLIPscore)
HPSv2	0.295	**0.309**	0.286	0.301
CLIPscore	0.384	**0.417**	0.330	0.349
VQAScore	0.869	**0.983**	0.365	0.695
ImageReward	1.757	**1.765**	1.697	1.694
LLM Grader	7.000	**7.320**	7.000	7.080
LLaVA-Reward (Ours)	0.330	**0.490**	0.279	0.230

Prompt

Photo of an athlete cat explaining it's latest scandal at a press conference to journalists.

	w/o inference-time scaling	FK steering (LLaVA-Reward)	FK steering (ImageReward)	FK steering (CLIPscore)
HPSv2	0.299	**0.354**	0.339	0.341
CLIPscore	0.286	0.347	**0.359**	0.344
VQAScore	0.565	**0.626**	0.531	0.614
ImageReward	1.566	**1.732**	1.455	1.661
LLM Grader	**8.120**	8.000	7.680	7.920
LLaVA-Reward (Ours)	0.188	**0.289**	0.186	0.183

Prompt

McDonalds Church.

	w/o inference-time scaling	FK steering (LLaVA-Reward)	FK steering (ImageReward)	FK steering (CLIPscore)
HPSv2	0.331	**0.332**	0.318	0.317
CLIPscore	0.275	0.320	0.306	**0.332**
VQAScore	0.395	**0.459**	0.218	0.435
ImageReward	1.738	**1.909**	1.576	1.560
LLM Grader	7.720	**7.960**	7.480	7.880
LLaVA-Reward (Ours)	-0.007	**0.254**	-0.098	0.159

Figure 6. Examples of diffusion inference-time scaling via FK steering (SDXL) using 5 different reward models with the prompt from DrawBench. The LLM grader is conducted using GPT-4o with prompts from Ma et al. [1].