Read It Back: Pretrained MLLMs Are Zero-Shot Reward Models for Text-to-Image Generation

Runhui Huang[1] **Qihui Zhang**[3] **Zhe Liu**[1] **Yu Gao**[2]
Jie Wu[2] **Hengshuang Zhao**[1]
[1]*The University of Hong Kong,* [2]*ByteDance Seed,* [3]*Peking University*

Abstract

In this paper, we propose **SpectraReward,** a training-free reward function that turns pretrained MLLMs into off-the-shelf reward models for image-generation reinforcement learning. Instead of asking the MLLM to judge a generated image or answer decomposed verification questions, SpectraReward measures how well the original prompt can be recovered from the generated image through a single image-conditioned, teacher-forced forward pass. We use the average image-conditioned prompt log-likelihood as the reward, directly reusing the MLLM's pretrained image-text alignment ability without preference labels, reward-model fine-tuning. We further introduce **Self-SpectraReward,** a special case for unified multimodal models where the policy's own understanding branch serves as the reward model for its generation branch, forming a closed-loop self-improving framework without external reward models or external knowledge. Extensive experiments validate SpectraReward through a broad image-generation RL study covering two diffusion models, three RL algorithms, nine reward MLLM backbones from four MLLM families spanning 4B to 235B parameters, and five out-of-distribution text-to-image benchmarks. Results show that both SpectraReward and Self-SpectraReward significantly and consistently improve generation performance and outperform prior MLLM-derived reward training methods. Further analysis reveals that larger reward MLLMs are not always better, while Self-SpectraReward can match or surpass much larger external reward models, suggesting that reward-policy alignment is a key factor for effective image-generation RL. Project Page: `https://huangrh99.github.io/SpectraReward/`

1 Introduction

Image generation has advanced rapidly in recent years, evolving from specialized text-to-image models [10, 24] to unified multimodal models (UMMs) [9, 41, 7, 46, 8, 16, 34] that integrate visual understanding and generation within a single architecture. Reinforcement learning has emerged in parallel as an effective post-training stage [28, 62, 61, 65], consistently lifting compositional fidelity and instruction-following. The success of an RL recipe, however, rests on two complementary pillars. The optimization algorithm governs the stability of long-horizon training, while the reward model determines the ceiling that the trained policy can ultimately approach. However, designing a practical reward model that remains both efficient and reliable is still challenging.

Recent studies have made substantial progress in reward modeling for image generation. One line of work builds reward models from large-scale human preference annotations, using these data to align visual-text representations or vision-language models with human judgments [22, 59, 54, 51, 47]. While effective, these methods depend on expensive annotation, difficult data collection, and costly iteration. Another line of work bypasses preference training by reusing pretrained MLLMs as zero-shot reward sources [23, 26, 17]. Direct scalar or logit-based feedback is training-free, but can be sensitive to judge calibration and scoring noise [23, 26]. Question-decomposition pipelines improve

Preprint.

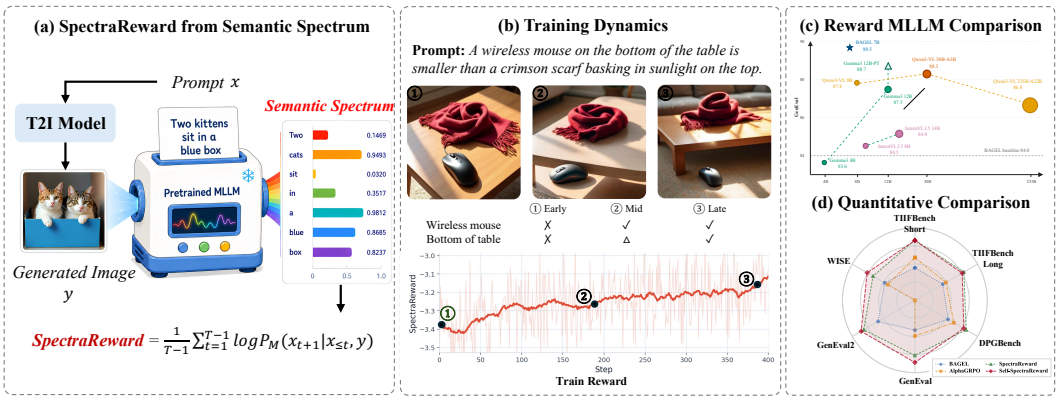

Figure 1: **Overview of SpectraReward.** (a) Pretrained MLLMs naturally induce a semantic spectrum that measures how well a generated image aligns with the prompt. SpectraReward aggregates this into a reward for T2I RL. (b) During RL training, SpectraReward steadily increases together with visible improvements in complex scene generation. (c) We study nine reward MLLM backbones from four MLLM families, with external reward MLLMs spanning three families and 4B to 235B parameters. Scaling the reward MLLM backbones brings non-monotonic gains. Qwen3-VL-30B-A3B achieves the best performance among external MLLMs, while Self-SpectraReward, using BAGEL's own understanding branch as the reward model, outperforms all external MLLMs. (d) Both SpectraReward and Self-SpectraReward bring significant and consistent improvements across all six downstream benchmarks compared to the baselines.

reliability by verifying atomic prompt requirements one by one, but introduce substantial engineering complexity [17]. This leaves open the design of an efficient, preference-label-free, and off-the-shelf reward model for open-sourced image generators.

We propose SpectraReward, a training-free reward function that turns any pretrained MLLM into an image-generation reward model by reusing its image-text alignment knowledge, without preference labels or additional training. Given a generated image, SpectraReward conditions a frozen MLLM on the image and performs a single teacher-forced forward pass on the prompt, producing a token-level likelihood profile that reflects how well each textual requirement can be read from the image. We refer to this profile as the *semantic spectrum* of the image-text pair, and aggregate it into a scalar reward using the mean prompt-token log-likelihood. In this way, SpectraReward uses the prompt-likelihood spectrum induced by the model's own image-conditioned language modeling objective as a deterministic, off-the-shelf reward signal, as shown in Figure 1 (a), instead of asking the MLLM to judge an image or answer decomposed questions.

We further specialize SpectraReward within unified multimodal models (UMMs), yielding a special case we call *Self-SpectraReward*. Since a UMM integrates image understanding and image generation into a single framework, its understanding branch could directly provide the reward signals for its generation branch. In this way, the model improves its own generation using its own understanding ability, without relying on any external model or external knowledge, forming a closed-loop self-improving framework in which the policy evolves by leveraging its own intrinsic capability. Beyond removing external dependencies, this design has a structural advantage. The understanding branch shares the same tokenizer, vision encoder, and pretraining distribution as the generation branch, so the way it reads an image is naturally aligned with the distribution from which the generation branch samples images. We therefore hypothesize that this intrinsic alignment makes the policy's own understanding branch a particularly suitable reward source for its own generation.

We validate SpectraReward through a broad image-generation RL study covering two diffusion models, three RL training algorithms, nine reward MLLM backbones from four MLLM families ranging from 4B to 235B parameters, and evaluate real generalization improvements on five out-of-distribution downstream text-to-image generation benchmarks. On both SD3.5-M [10] and BAGEL [9], SpectraReward brings significant improvements on all downstream benchmarks, showing the generalization ability of the proposed prompt-likelihood reward. With BAGEL as the policy model, both SpectraReward and Self-SpectraReward consistently improve the BAGEL baseline by +10.0 on TIIF-Bench [48] and +4.3/5.5 on GenEval [13], respectively. SpectraReward also

outperforms the strong RL baseline, AlphaGRPO [17], by 6.3 and 2.1 gains on TIIF-Bench and GenEval, respectively. Moreover, Self-SpectraReward exhibits better improvements compared to SpectraReward with similar-scale MLLM backbones, and comparable or even better performance compared to 4x or 30x larger MLLMs, e.g., +1.2 on GenEval, +2.1 on GenEval2 [21], and +2 on WISE [31] compared to SpectraReward with the best MLLM backbone. These results support our hypothesis that distributional alignment between the reward model and the policy can be as important as the scale of the reward model. The contributions of the paper are summarized as follows:

- We propose SpectraReward, a training-free and off-the-shelf reward function that turns any pretrained MLLM into a reward model for text-to-image generation. By calculating the image-conditioned prompt-token likelihood, SpectraReward reuses pretrained image-text alignment without preference labels, reward-model fine-tuning, scalar judging, or question decomposition.
- We introduce Self-SpectraReward, the special case in which the policy's own understanding branch in a unified multimodal model serves as the reward model. This forms a closed-loop self-improving framework without external reward models or external knowledge, where the model leverages its own understanding ability to improve generation performance.
- We conduct a broad image-generation RL study covering two generator backbones, three RL algorithms, nine reward MLLM backbones from four MLLM families spanning 4B–235B parameters, and five downstream benchmarks. The results show the effectiveness of SpectraReward and Self-SpectraReward, and reveal that the benefits of scaling the reward model are not monotonic and that the well-aligned self-reward model can match or even outperform much larger external models.

2 Related Work

Reward Models for Image Generation. Reinforcement learning from human feedback (RLHF) [32] is a standard paradigm for aligning LLMs [11, 38], and has been adapted to T2I models [4] via reward models trained on extensive human preferences or prompt–image alignment data [59, 60, 22, 59, 30]. While these reward models provide critical alignment signals, their effectiveness is fundamentally bottlenecked by an inherent bias against text-independent aesthetic details [1]. To overcome these capacity limits, a second line of research upgrades the reward architecture to fine-tuned Multimodal Large Language Models (MLLMs), which successfully raise the performance ceiling [60, 51, 14]. However, this requires large amounts of annotated data, substantial training overhead, and an inevitable shift in distribution away from the MLLM's pretraining knowledge. More recently, an emerging paradigm dispenses with both preference labels and reward-model fine-tuning. This approach leverages pretrained MLLMs as zero-shot judges, capable of assessing image quality and aesthetic reasoning directly [18, 37]. These methods typically operate either by emitting a direct scalar score or by decomposing the generation prompt into verifiable atomic questions answered by the MLLM [17]. While concurrent work such as PromptEcho [29] explores a cross-entropy function of this concept, its evaluation is restricted to domain-specific T2I backbones.

Self-Rewarding in Unified Multimodal Models. Self-rewarding mechanisms, wherein a model leverages its own evaluations to supervise its training, have demonstrated significant efficacy in LLM alignment [63, 53, 39]. This paradigm naturally extends to Unified Multimodal Models (UMMs) [42, 58, 52, 64], which inherently integrate both understanding and generation capabilities. Alongside broader efforts to align multimodal models using self-generated preference data and iterative self-evolution [40, 19], recent concurrent studies have explored the specific intersection of self-rewarding and UMMs: SRUM [20] introduces an aggregation of global and local signals from the UMM's comprehension branch, while GvU [33] utilizes the token-level logits of the understanding branch as dense reward signals. Although these studies align with the intuition that a policy's internal comprehension can effectively guide its own generation, they are confined to UMM-internal evaluations and fail to investigate how this reward compares against external reward models.

3 Method

In this section, we first introduce the SpectraReward reward function in Section 3.1. We then specialize it to unified multimodal models in Section 3.2, yielding Self-SpectraReward. Finally, Section 3.3 analyzes the discriminative properties of the reward signal. Figure 2 compares SpectraReward with previous MLLM-based reward functions.

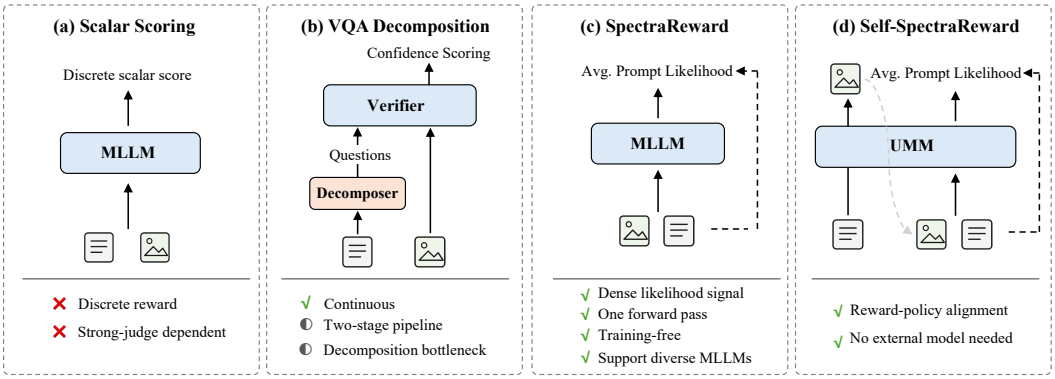

Figure 2: **Comparison of MLLM-based reward functions.** (a) Scalar scoring directly asks an MLLM to assign a discrete image-text alignment score, making the reward sensitive to judge calibration and scoring noise. (b) VQA decomposition converts the prompt into atomic questions and aggregates verifier confidence, but introduces a two-stage pipeline and depends on the quality of question decomposition. (c) SpectraReward computes the image-conditioned prompt likelihood through a single teacher-forced forward pass. The resulting token-level likelihoods form a semantic spectrum, whose average is used as the scalar reward. (d) Self-SpectraReward instantiates the same prompt-likelihood reward within a unified multimodal model by using the policy's own understanding branch, removing the need for an external reward MLLM and improving reward-policy alignment.

3.1 SpectraReward

Let G_θ be a text-to-image policy that samples an image $y \sim G_\theta(\cdot \mid x)$ from a prompt $x = (x_1, \ldots, x_T)$, and let $\mathcal{M}$ be the frozen pretrained MLLM. Since pretrained MLLMs already learn strong image-text alignment from large-scale pretraining, our goal is to extract a scalar reward $R_{\mathcal{M}}(x, y)$ from the pretrained MLLMs that measures how well the generated image y realizes the prompt x, while requiring no preference labels, reward-model fine-tuning, or other auxiliary pipeline.

The key idea of SpectraReward is to evaluate the generated image by asking a simple inverse question: how well the generated image y can be translated back into the prompt x from the perspective of $\mathcal{M}$. To compute this, we feed the image y as the visual condition to $\mathcal{M}$, and then perform a single teacher-forced forward pass over the prompt tokens x. The final reward function is as follows:

$$R_{\mathcal{M}}(x, y) = \frac{1}{T-1} \sum_{t=1}^{T-1} \log p_{\mathcal{M}}(x_{t+1} \mid x_{\leq t}, y). \tag{1}$$

This reward is the mean image-conditioned prompt log-likelihood under $\mathcal{M}$. The token-wise likelihoods form the *semantic spectrum* of the image-text pair, describing how the visual evidence in y supports the semantics of x across prompt tokens. SpectraReward aggregates this semantic spectrum into a scalar reward, where a higher reward $R(x, y)$ indicates that the prompt is more predictable from the image and that the generated image better supports the semantic content of the prompt.

The key advantage is that this reward function directly invokes the most fundamental capability of pretrained MLLMs, namely understanding the image and associating it with the corresponding language description. Thus, rather than training a separate reward model or prompting the MLLM to produce an abstract judgment score, SpectraReward reuses the fundamental visual-language grounding ability already learned during MLLM pretraining.

3.2 Self-SpectraReward

When G_θ is a unified multimodal model, it contains both a generation branch G_θ^{gen} and an understanding branch G_θ^{und}. Self-SpectraReward instantiates SpectraReward by using G_θ^{und} as the reward MLLM for images sampled by G_θ^{gen}. In other words, the model scores its own generated images by measuring how likely the original prompt is under its own image-conditioned understanding branch. This forms a closed-loop self-improving framework in which the model uses its own understanding ability to improve the generation quality. Moreover, this self-rewarding design eliminates the need

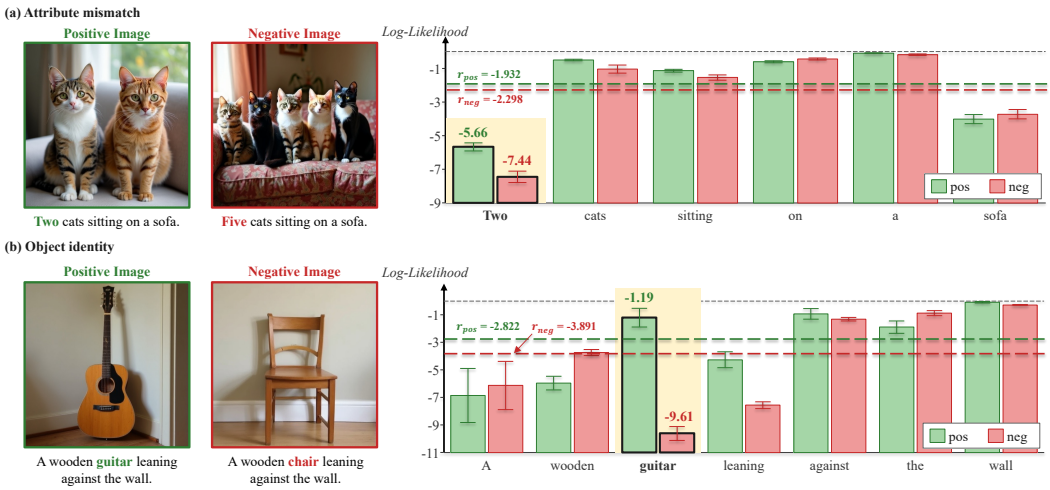

Figure 3: **Token-level semantic sensitivity of SpectraReward.** Positive and negative images are evaluated using the same positive prompt. For attribute mismatch, instantiated as a counting error, the negative image mainly lowers the likelihood of "Two"; for object identity mismatch, replacing the guitar with a chair sharply lowers the likelihood of "guitar". Bars show image-conditioned prompt-token log-likelihoods with error bars calculated over four pairs, and dashed lines show the resulting sequence-level reward value, i.e., SpectraReward.

for external reward models, avoids using additional computational resources to serve large-scale MLLMs, and simplifies the RL training pipeline.

Reward-policy alignment. Self-SpectraReward couples two dual capabilities within the same unified model. The generation branch maps text to images, while the reward function uses the understanding branch to measure how well the generated image can be translated back to the original text. Since the two branches share the same tokenizer, vision encoder, and pretraining distribution, the reward signal is compatible with the policy's own generation knowledge. This does not require the understanding branch to be stronger than external MLLMs in general visual understanding. Instead, it provides a reward signal that is better calibrated to the policy's own generated-image distribution. Optimizing the generation branch with this self-reward therefore encourages the text-to-image policy to become more consistent with its own image-to-text understanding ability. We further analyze the reward-policy alignment effect in Section 4.3.

3.3 Analysis of SpectraReward

In this section, we analyze three questions: (1) whether the language prior affects the accuracy of the reward signal; (2) whether token-level likelihoods are sensitive to visual misalignment; (3) whether the resulting scalar reward can reliably distinguish samples within the same prompt group.

Language-prior cancellation. A natural concern is that Equation 1 contains the MLLM's language prior over the prompt, since some words are easier to predict regardless of the image. One possible correction is to subtract a text-only likelihood term, yielding a PMI-style reward $\log p(x \mid y) - \log p(x)$. However, since we focus on using group-relative reinforcement learning methods, this correction is unnecessary. For a fixed prompt, the text-only likelihood is shared by all generated images in the same group, and therefore cancels out when computing the group-relative advantage. Thus, PMI normalization may affect absolute reward calibration across prompts, but it does not provide additional optimization signal within each prompt group.

Token-level semantic sensitivity. We further examine whether the semantic spectrum responds to local visual errors. We construct positive and negative image pairs for two common prompt-following failures, attribute mismatch and object identity mismatch. The attribute mismatch is instantiated as a counting error, where the prompt asks for two cats but the negative image contains five. For each pair, we evaluate both images using the original positive prompt. Figure 3 shows that the likelihood drop is concentrated on the mismatched semantic word. The counting mismatch mainly lowers the likelihood of "Two", while replacing the guitar with a chair sharply lowers the likelihood of "guitar".

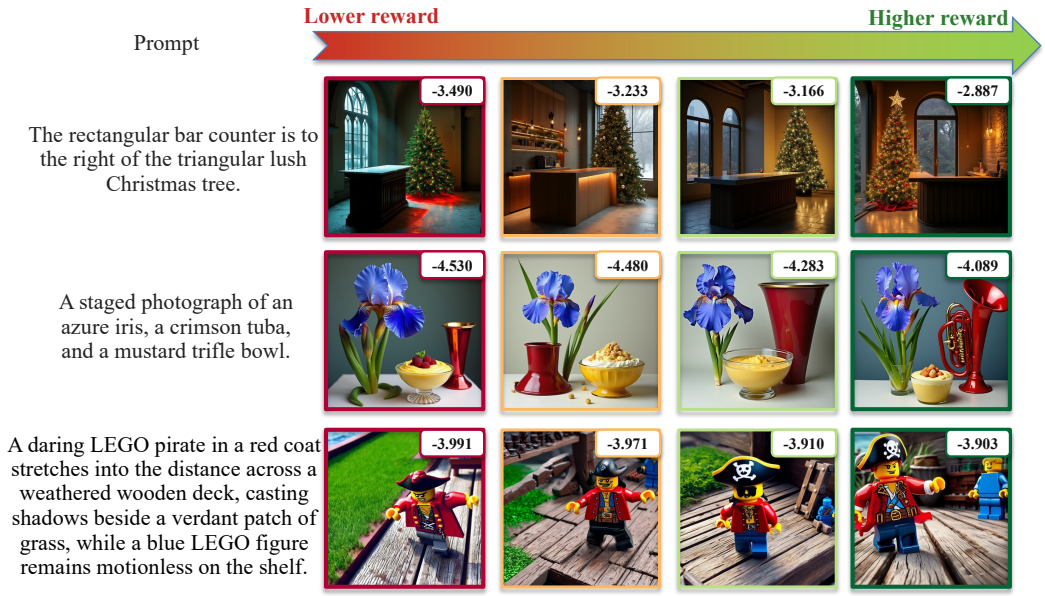

Figure 4: **The visual interpretation of SpectraReward.** The reward ranking is consistent with the visual quality ranking.

The sequence-level reward, shown by the dashed lines, also remains higher for the positive images. This indicates that token-level prompt likelihoods are sensitive to targeted semantic errors, and their average, i.e., the calculation used by SpectraReward, still provides a reliable scalar reward.

Reward ranking reliability. What matters for group-relative RL is whether the reward correctly orders different rollouts from the same prompt. As shown in Figure 4, SpectraReward assigns higher rewards to images that better satisfy the prompt and lower rewards to images with missing objects, wrong attributes, or incorrect spatial relations. This indicates that the image-conditioned prompt likelihood provides a reliable group-wise ranking signal for text-to-image RL.

4 Experiments

4.1 Experimental Setup

We use AWM as the default reinforcement learning algorithm, following the optimizer ablation in Section 4.4, where AWM achieves the best overall performance among the tested RL algorithms. Our policy model is BAGEL [9], a native AR-Diffusion unified multimodal model pretrained for both image understanding and image generation. Unless otherwise specified, SpectraReward uses Qwen3-VL-30B-A3B [2] as the reward MLLM backbone, while Self-SpectraReward uses BAGEL's own understanding branch for reward computation.

Training settings. We use $T = 16$ sampling steps during training and stochastically sample 6 steps from the first 10 denoising steps for training. Each training step uses 32 prompts, and the group size is $G = 16$. The training image resolution is 512×512. We use a classifier-free guidance scale of 4 and a timestep shift of 3. The learning rate is 1e-4. We trained on 32 A100 GPUs with batch size of 2 and gradient accumulation of 8. The total training steps are 380. Other hyperparameters follow the official AWM implementation. We train on AlphaGRPO20k [17].

4.2 Main Results

We use our SpectraReward as the proxy reward during training and evaluate the real improvement of generation performance on five out-of-distribution benchmarks. All downstream benchmarks use different evaluation protocols and out-of-distribution test sets: GenEval [13], TIIF-Bench [48], DPG-Bench [15], Geneval2 [21], and WISE [31]. We compare against the state-of-the-art generation-only models, including SD3 Medium [10] and FLUX.1 dev [24], and unified multimodal models such as

Table 1: **Results on Text-to-Image benchmarks**. S and L denote Short and Long prompts, respectively. For WISE, the second score denotes evaluation with CoT. AlphaGRPO is trained on the reasoning text-to-image generation task. **Bold** indicates the best performance. The results of BAGEL are reproduced.

Model	TIIF Bench ↑								WISE ↑	DPGBench ↑	GenEval ↑	GenEval2 ↑
	Basic		Advanced		Designer		Overall					
	S	L	S	L	S	L	S	L	Overall	Score	Score	Soft TIFA_{GM}
Generation Only Models												
SD3 Medium [10]	78.3	77.8	61.5	59.6	63.2	67.3	64.8	64.8	0.4	84.1	74.0	21.3
FLUX.1 dev [24]	83.1	78.7	65.8	68.5	70.7	71.5	71.1	71.8	0.5	83.8	82.0	21.1
Unified Multimodal Models												
Show-o [57]	73.1	75.8	55.0	50.9	53.7	50.4	59.7	58.9	0.35	-	69.0	-
JanusPro [7]	79.3	78.3	59.7	58.8	65.8	60.3	66.5	65.0	0.35	84.2	80.0	-
Inference on 512 Resolution												
BAGEL	81.7	86.1	73.7	77.6	84.7	82.1	75.2	78.6	-	85.0	84.0	23.6
AlphaGRPO [17]	85.5	84.2	77.4	78.9	84.3	86.6	78.9	79.5	-	86.0	85.0	-
SpectraReward	87.7	87.4	**84.9**	85.4	88.1	**90.3**	**85.2**	84.8	-	**88.1**	88.3	31.8
Self-SpectraReward	**89.7**	**87.6**	83.7	**85.7**	**90.3**	86.2	85.1	**84.3**	-	87.7	**89.5**	**33.9**
Inference on 1024 Resolution												
BAGEL	83.4	83.7	75.2	76.7	79.8	73.5	76.4	76.2	0.53/0.70	85.1	86.6	23.5
AlphaGRPO [17]	84.8	85.9	79.9	78.4	80.2	80.2	78.7	79.5	0.50/0.69	85.9	87.4	-
SpectraReward	**88.4**	87.3	**83.2**	81.8	82.5	81.0	**83.3**	81.7	**0.53**/0.74	86.5	88.3	32.6
Self-SpectraReward	87.4	**87.6**	81.5	**83.2**	**87.7**	**84.7**	82.7	**82.5**	0.52/**0.76**	**87.0**	**89.8**	**34.3**

Show-o [57], JanusPro [7], and our baseline BAGEL [9]. We compare with AlphaGRPO [17], which also applies the RL training on Bagel with the proposed MLLM-derived DVReward.

Table 1 compares SpectraReward and Self-SpectraReward with representative text-to-image generators and prior MLLM-derived reward training methods. Overall, both SpectraReward and Self-SpectraReward achieve comparable performance and consistently improve over the BAGEL baseline and AlphaGRPO across all evaluated benchmarks. Specifically, at 512 resolution, SpectraReward improves BAGEL on TIIF-Bench overall short/long prompts by +10.0/+6.2, while Self-SpectraReward achieves a comparable 85.1/84.3 and further brings a 5.5 gain on GenEval. Compared with AlphaGRPO, SpectraReward improves TIIF-Bench by +6.3/+5.3 on short/long prompts and +3.3 on GenEval, while Self-SpectraReward improves them by +6.2/+4.8 and +4.5, respectively.

Consistent with AlphaGRPO [17], although training is performed at 512 resolution, the improvements also transfer to 1024-resolution inference. Specifically, Self-SpectraReward reaches 89.8 on GenEval and 34.3 on GenEval2, outperforming both BAGEL and AlphaGRPO. More importantly, Self-SpectraReward achieves 0.76 on the knowledge-grounded benchmark, WISE. These results suggest that SpectraReward provides a more effective supervision signal than decompositional verification for text-to-image RL.

4.3 Effect of reward MLLM backbone

We next study the effect of different reward MLLM backbones in SpectraReward, covering different families and scales. Table 2 reports the experimental results.

Consistent improvement of SpectraReward across diffusion models and reward MLLM families. On both SD3.5-M [10] and BAGEL, applying SpectraReward improves the baseline, highlighting the architecture-agnostic nature of the caption-likelihood reward. On BAGEL, different reward MLLMs from different families, including Gemma3 [43], InternVL3.5 [45], and Qwen3-VL [2], all bring improvements over the baseline on TIIF-Bench. This indicates that SpectraReward can extract useful reward signals from diverse MLLM backbones.

Self-SpectraReward shows that reward-policy alignment can rival external scale. Using BAGEL's own understanding branch as the reward MLLM, Self-SpectraReward achieves the best result among similar-scale reward models, matches the strongest 30B-class external reward, and outperforms 30x larger MLLMs, i.e., Qwen3-VL-235B-A22B [2]. This suggests that reward quality

Table 2: **Effect of different reward MLLM backbones.** Unless otherwise specified, reward MLLMs all use Instruct models. **Bold** indicates the best performance. The second best is underlined.

Base Model	Reward MLLM	GenEval↑	TIIF-Short↑	TIIF-Long↑
SD3.5-M	-	79.8	74.0	73.2
SD3.5-M	Qwen3-VL-8B	85.7	84.4	84.6
BAGEL	-	84.0	75.2	78.6
BAGEL	Gemma3-4B	83.6	83.1	83.0
BAGEL	Gemma3-12B-Pretrain	88.7	84.3	82.6
BAGEL	Gemma3-12B	87.5	82.0	82.6
BAGEL	InternVL3.5-8B	84.5	80.4	80.0
BAGEL	InternVL3.5-14B	84.9	78.5	79.9
BAGEL	Qwen3-VL-8B	87.8	85.0	83.8
BAGEL	Qwen3-VL-30B-A3B	88.3	**85.2**	**84.8**
BAGEL	Qwen3-VL-235B-A22B	86.8	84.3	83.4
BAGEL	BAGEL	**89.5**	85.1	84.3

Table 3: **Comparison of different RL algorithms and reward models.** When trained with Self-SpectraReward, AWM achieves the best downstream performance. Self-SpectraReward further outperforms prior reward models under comparable RL training. **Bold** indicates the best performance.

RL Algorithm	Reward Model	GenEval↑	TIIF-Short↑	TIIF-Long↑
-	-	84.0	75.2	78.6
AlphaGRPO	HPSv3	83.4	78.5	77.1
AlphaGRPO	UnifiedReward	83.7	79.2	77.3
AlphaGRPO	VIEScore	81.7	79.1	77.9
AlphaGRPO	DVReward	85.0	78.9	79.5
AWM	DVReward	88.0	83.3	83.7
FlowGRPO	Self-SpectraReward	85.3	82.1	79.8
DiffusionNFT	Self-SpectraReward	89.1	84.8	83.5
AWM	Self-SpectraReward	**89.5**	**85.1**	**84.3**

is not determined by scale alone. We attribute the success of Self-SpectraReward to its distributional alignment with the policy: Self-SpectraReward shares the tokenizer, vision encoder, and pretraining distribution with the generator, so the way it reads generated images is naturally matched to the distribution from which the policy samples them.

Reward MLLM scale helps, but is not monotonic. Within Qwen3-VL, scaling from 8B to 30B improves all reported metrics, but further scaling to 235B leads to a clear drop. Within the Gemma3 family, scaling from 4B to 12B improves GenEval, but does not consistently improve TIIF-Bench. We hypothesize that larger MLLMs may devote more capacity to advanced vision-language reasoning and instruction-following tasks, while the more fundamental image-conditioned captioning ability needed by SpectraReward may already saturate at a moderate scale. These results suggest that a 30B MLLM is sufficient to unlock most of the benefit of SpectraReward.

Pretraining-stage MLLMs can be better reward backbones for SpectraReward. Gemma3-12B-Pretrain consistently outperforms Gemma3-12B-Instruct across three evaluated benchmarks. We hypothesize that pretraining learns large-scale image-captioning and image-text alignment objectives that closely match our reward function, while instruction tuning emphasizes advanced visual tasks and multi-task instruction following that are less directly needed by SpectraReward.

4.4 Ablation Study

The effect of RL algorithm. We compare three image generation reinforcement learning methods in the Self-SpectraReward setting: FlowGRPO [28], AWM [61], and DiffusionNFT [65]. Table 3 shows that AWM achieves the best overall downstream performance, outperforming the SDE-based FlowGRPO and DiffusionNFT. We therefore apply AWM in the main experiments.

Table 4: **Ablation of Reward function.** Scalar Scoring asks the MLLM to rate the image from 1 to 5. VQA-Score asks "Does this figure show `{caption}`?" and uses $P(\text{yes})$ as the reward.

Reward function	GenEval↑	TIIF-S↑	TIIF-L↑
-	84.0	75.2	78.6
Scalar Scoring (1-5)	77.7	67.5	76.0
VQA-Score ($P(\text{yes})$)	83.1	77.4	78.9
Prompt Likelihood	**89.5**	**85.1**	**84.3**

Table 5: **Ablation of Reward granularity.** Token-level reward does not consistently outperform sequence-level reward, so we use sequence-level reward by default.

Method	GenEval↑	TIIF-S↑	TIIF-L↑
Sequence-level Adv.	**89.5**	**85.1**	84.3
Token-level Adv.	88.4	84.9	**85.1**
+ Group std	88.5	84.0	84.8
+ Per-token std	88.8	84.1	82.1

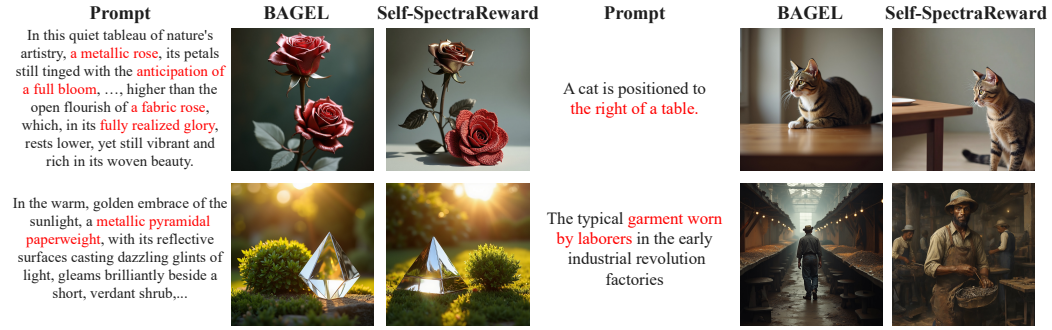

Figure 5: **Qualitative comparison.**

Discussion of Reward function. We compare SpectraReward with two alternative MLLM-based reward functions for measuring image-text alignment. As illustrated in Figure 5, the first is scalar scoring, which directly asks the MLLM to rate image-text alignment with a 1–5 score. The second is VQA-Score, which asks a yes/no question such as "Does this figure show `{text}`?" and uses the probability of the "yes" token as the reward. Table 4 shows that the reward function has a large impact on RL performance. Scalar scoring substantially degrades all benchmarks, including a 6.3 drop on GenEval compared with the baseline, while VQA-Score brings minor gains on TIIF-Bench but fails to improve GenEval. In contrast, the image-conditioned prompt likelihood used by SpectraReward achieves the best performance across all metrics, showing that likelihood-based reward is a more effective way to leverage MLLMs for image-generation RL.

Sequence-level vs. token-level advantage. We compare the default sequence-level advantage with token-level advantage. The former first averages token log-likelihoods into one sequence-level reward and then computes group-relative advantages, while the latter treats each token position as an individual reward signal, computes token-wise group advantages, and averages them as the sequence advantage. We test three normalization scopes for token-level std: global std, group std of all text tokens in the group, and per-token std. Although token-level advantage can emphasize visually grounded tokens with larger rollout differences, Table 5 shows no obvious improvement over sequence-level advantage. We use sequence-level reward by default for more stable training.

5 Conclusion

We introduced SpectraReward, a training-free reward function that turns pretrained MLLMs into off-the-shelf reward models for image-generation reinforcement learning by measuring image-conditioned prompt likelihood. Instead of relying on massive preference labels, reward-model fine-tuning, SpectraReward aggregates the token-level semantic spectrum induced by an MLLM into a simple and effective reward. We further introduced Self-SpectraReward for unified multimodal models, where the policy's own understanding branch serves as the reward model for its generation branch, forming a closed-loop self-improving framework without external reward models. Across generator backbones, RL algorithms, reward MLLM backbones, and out-of-distribution benchmarks, SpectraReward significantly and consistently improves text-to-image generation and outperforms prior MLLM-derived reward training methods. The results also show that larger reward MLLMs are not always better, while a well-aligned self-reward can match or surpass much larger external reward models. These findings suggest that reward function and reward-policy alignment are key factors for reinforcement learning in unified multimodal generation.

References

[1] Ying Ba, Tianyu Zhang, Yalong Bai, Wenyi Mo, Tao Liang, Bing Su, and Ji-Rong Wen. Enhancing reward models for high-quality image generation: Beyond text-image alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19022–19031, 2025.

[2] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[3] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *OpenAI blog*, 2023.

[4] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023.

[5] Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025.

[6] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *European Conference on Computer Vision*, pages 74–91. Springer, 2024.

[7] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.

[8] Yufeng Cui, Honghao Chen, Haoge Deng, Xu Huang, Xinghang Li, Jirong Liu, Yang Liu, Zhuoyan Luo, Jinsheng Wang, Wenxuan Wang, et al. Emu3. 5: Native multimodal models are world learners. *arXiv preprint arXiv:2510.26583*, 2025.

[9] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

[10] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

[11] Chujie Gao, Siyuan Wu, Yue Huang, Dongping Chen, Qihui Zhang, Zhengyan Fu, Yao Wan, Lichao Sun, and Xiangliang Zhang. Honestllm: Toward an honest and helpful large language model. *arXiv preprint arXiv:2406.00380*, 2024.

[12] Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*, 2024.

[13] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: an object-focused framework for evaluating text-to-image alignment. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 52132–52152, 2023.

[14] Yuan Gong, Xionghui Wang, Jie Wu, Shiyin Wang, Yitong Wang, and Xinglong Wu. Onereward: Unified mask-guided image generation via multi-task human preference learning. *arXiv preprint arXiv:2508.21066*, 2025.

[15] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

[16] Runhui Huang, Chunwei Wang, Junwei Yang, Guansong Lu, Yunlong Yuan, Jianhua Han, Lu Hou, Wei Zhang, Lanqing Hong, Hengshuang Zhao, et al. Illume+: Illuminating unified mllm with dual visual tokenization and diffusion refinement. *arXiv preprint arXiv:2504.01934*, 2025.

[17] Runhui Huang, Jie Wu, Rui Yang, Zhe Liu, and Hengshuang Zhao. Alphagrpo: Unlocking self-reflective multimodal generation in unified multimodal models via decompositional verifiable reward. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

[18] Ruixiang Jiang and Chang Wen Chen. Multimodal llms can reason about aesthetics in zero-shot. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 6634–6643, 2025.

[19] Songtao Jiang, Yan Zhang, Ruizhe Chen, Tianxiang Hu, Yeying Jin, Qinglin He, Yang Feng, Jian Wu, and Zuozhu Liu. Modality-fair preference optimization for trustworthy mllm alignment. *arXiv preprint arXiv:2410.15334*, 2024.

[20] Weiyang Jin, Yuwei Niu, Jiaqi Liao, Chengqi Duan, Aoxue Li, Shenghua Gao, and Xihui Liu. Srum: Fine-grained self-rewarding for unified multimodal models. *arXiv preprint arXiv:2510.12784*, 2025.

[21] Amita Kamath, Kai-Wei Chang, Ranjay Krishna, Luke Zettlemoyer, Yushi Hu, and Marjan Ghazvininejad. Geneval 2: Addressing benchmark drift in text-to-image evaluation. *arXiv preprint arXiv:2512.16853*, 2025.

[22] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.

[23] Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. Viescore: Towards explainable metrics for conditional image synthesis evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12268–12290, 2024.

[24] Black Forest Labs. Flux, 2024. URL `https://github.com/black-forest-labs/flux`.

[25] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*, 2024.

[26] Zongjian Li, Zheyuan Liu, Qihui Zhang, Bin Lin, Feize Wu, Shenghai Yuan, Zhiyuan Yan, Yang Ye, Wangbo Yu, Yuwei Niu, et al. Uniworld-v2: Reinforce image editing with diffusion negative-aware finetuning and mllm implicit feedback. *arXiv preprint arXiv:2510.16888*, 2025.

[27] Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, et al. Uniworld: High-resolution semantic encoders for unified visual understanding and generation. *arXiv preprint arXiv:2506.03147*, 2025.

[28] Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025.

[29] Jinlong Liu, Wanggui He, Peng Zhang, Mushui Liu, Hao Jiang, and Pipei Huang. Promptecho: Annotation-free reward from vision-language models for text-to-image reinforcement learning. *arXiv preprint arXiv:2604.12652*, 2026.

[30] Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15086–15095, 2025.

[31] Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Kunpeng Ning, Chaoran Feng, Bin Zhu, and Li Yuan. Wise: A world knowledge-informed semantic evaluation for text-to-image generation. *arXiv preprint arXiv:2503.07265*, 2025.

[32] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[33] Jiadong Pan, Liang Li, Yuxin Peng, Yu-Ming Tang, Shuohuan Wang, Yu Sun, Hua Wu, Qingming Huang, and Haifeng Wang. Learning to generate via understanding: Understanding-driven intrinsic rewarding for unified multimodal models. *arXiv preprint arXiv:2603.06043*, 2026.

[34] Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, et al. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025.

[35] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024.

[36] Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K Du, Zehuan Yuan, and Xinglong Wu. Tokenflow: Unified image tokenizer for multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2545–2555, 2025.

[37] Sheila Schoepp, Masoud Jafaripour, Yingyue Cao, Tianpei Yang, Fatemeh Abdollahi, Shadan Golestan, Zahin Sufiyan, Osmar R Zaiane, and Matthew E Taylor. The evolving landscape of llm-and vlm-integrated reinforcement learning. *arXiv preprint arXiv:2502.15214*, 2025.

[38] Zhihong Shao et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[39] Zhiqing Sun, Yikang Shen, Hongxin Zhang, Qinhong Zhou, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. Salmon: Self-alignment with instructable reward models. *arXiv preprint arXiv:2310.05910*, 2023.

[40] Wentao Tan, Qiong Cao, Yibing Zhan, Chao Xue, and Changxing Ding. Beyond human data: Aligning multimodal large language models by iterative self-evolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 7202–7210, 2025.

[41] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

[42] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

[43] Gemma Team. Gemma 3 technical report, 2025.

[44] Chunwei Wang, Guansong Lu, Junwei Yang, Runhui Huang, Jianhua Han, Lu Hou, Wei Zhang, and Hang Xu. Illume: Illuminating your llms to see, draw, and self-enhance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21612–21622, 2025.

[45] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

[46] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.

[47] Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025.

[48] Xinyu Wei, Jinrui Zhang, Zeqing Wang, Hongyang Wei, Zhen Guo, and Lei Zhang. Tiif-bench: How does your t2i model follow your instructions? *arXiv preprint arXiv:2506.02161*, 2025.

[49] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12966–12977, 2025.

[50] Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025.

[51] Jie Wu, Yu Gao, Zilyu Ye, Ming Li, Liang Li, Hanzhong Guo, Jie Liu, Zeyue Xue, Xiaoxia Hou, Wei Liu, et al. Rewarddance: Reward scaling in visual generation. *arXiv preprint arXiv:2509.08826*, 2025.

[52] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. In *Forty-first International Conference on Machine Learning*, 2024.

[53] Tianhao Wu, Weizhe Yuan, Olga Golovneva, Jing Xu, Yuandong Tian, Jiantao Jiao, Jason E Weston, and Sainbayar Sukhbaatar. Meta-rewarding language models: Self-improving alignment with llm-as-a-meta-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11548–11565, 2025.

[54] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

[55] Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, Song Han, and Yao Lu. Vila-u: A unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024. URL `https://arxiv.org/abs/2409.04429`.

[56] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13294–13304, 2025.

[57] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024.

[58] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024.

[59] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

[60] Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, et al. Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. *arXiv preprint arXiv:2412.21059*, 2024.

[61] Shuchen Xue, Chongjian Ge, Shilong Zhang, Yichen Li, and Zhi-Ming Ma. Advantage weighted matching: Aligning rl with pretraining in diffusion models. *arXiv preprint arXiv:2509.25050*, 2025.

[62] Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025.

[63] Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*, 2024.

[64] Shanshan Zhao, Xinjie Zhang, Jintao Guo, Jiakui Hu, Lunhao Duan, Minghao Fu, Yong Xien Chng, Guo-Hua Wang, Qing-Guo Chen, Zhao Xu, et al. Unified multimodal understanding and generation models: Advances, challenges, and opportunities. *arXiv preprint arXiv:2505.02567*, 2025.

[65] Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. Diffusionnft: Online diffusion reinforcement with forward process. *arXiv preprint arXiv:2509.16117*, 2025.

[66] Chunting Zhou, LILI YU, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *The Thirteenth International Conference on Learning Representations*, 2024.

Appendix

The outline of the Appendix is as follows.

- A Limitations and Future Work. We discuss the limitations of prompt-likelihood rewards, including dependence on MLLM visual reasoning, implicit physical and commonsense implications, and complementary reward signals.
- B Additional Implementation Details, including the reward prompt format, EOS masking, and the use of VAE features in Self-SpectraReward on BAGEL.
- C Additional Experimental Results, including the effect of EOS token removal and the effect of VAE features in reward calculation on BAGEL.
- D Detailed Benchmark Results, including category-level results on GenEval, TIIF-Bench, DPG-Bench, and WISE.
- E More Visualizations, including qualitative comparisons on spatial relations, counting, attribute binding, relative size, and long-prompt composition.
- F Broader Impacts.

A Limitations and future work

SpectraReward relies on the image-conditioned prompt likelihood of pretrained MLLMs, so its reward quality is bounded by the visual understanding and reasoning ability of the chosen reward backbone. Since the likelihood is computed only over the input prompt, the reward mainly captures explicit semantic alignment between generated images and prompts, and may under-emphasize implicit visual implications that are not directly expressed in the text. For example, a prompt such as "hot coffee" may imply steam through physical and commonsense reasoning, but such implications may be weakly reflected in the likelihood of the original prompt tokens. This limitation may be alleviated by using stronger MLLM backbones with rich world knowledge, or by extending SpectraReward to reasoning text-to-image generation, where intermediate reasoning can make such implicit requirements explicit. In addition, SpectraReward does not directly optimize criteria such as aesthetics and safety, which could be addressed by combining it with complementary specialized reward models.

B Additional Implementation Details

When using SpectraReward, we feed the generated image as the visual condition and evaluate the original prompt with teacher forcing. We do not add any general prefix to ask the reward MLLM to describe the image. Although such a prefix can change absolute likelihood values, it does not substantially affect the relative ranking within each prompt group, which is the quantity used by group-relative RL. Therefore, we refrain from using a prefix for all reward MLLMs by default. The only exception is the InternVL3.5 family, where we empirically find that adding the prefix "Describe the image" gives more stable reward values.

We exclude the [EOS] token from the SpectraReward computation. The [EOS] likelihood is often dominated by sequence termination rather than image-text alignment, and it can disproportionately affect short prompts because it contributes a large fraction of the averaged token likelihood. The ablation in Appendix C shows that masking [EOS] improves GenEval while keeping TIIF-Bench performance comparable.

For Self-SpectraReward on BAGEL, we use the model's own understanding branch to score generated images. BAGEL contains both semantic visual features and VAE features. By default, we include the VAE feature in Self-SpectraReward because it gives better average performance in our ablation. We discuss the effect of VAE inputs in Appendix C.

C Additional Experimental Results

Effect of EOS token removal. When inspecting the per-token likelihood distribution, we find that the EOS token often receives much lower likelihood than regular prompt tokens. Because SpectraReward averages token log-likelihoods, this unusually low-likelihood token can significantly

Table 6: **The effect of including the EOS token in the reward calculation.**

Include EOS token	GenEval↑	TIIF-Short↑	TIIF-Long↑
Yes	88.5	**85.5**	**84.3**
No	**89.5**	85.1	**84.3**

Table 7: **Effect of VAE features in Self-SpectraReward reward computation.** BAGEL provides two visual inputs to its understanding branch, ViT features from the semantic encoder and VAE features from the generation encoder. Using both inputs improves GenEval and TIIF-Short.

Use VAE input	GenEval↑	TIIF-Short↑	TIIF-Long↑
No	87.8	83.6	**84.7**
Yes	**89.5**	**85.1**	84.3

shift the final reward, especially for short prompts. However, EOS only indicates sequence termination and does not carry semantic information about image-text alignment. We therefore mask the EOS token when computing SpectraReward.

Effect of VAE feature in the reward calculation on BAGEL. BAGEL uses separate encoders for generation and understanding, i.e., ViT for understanding and VAE for generation, and can optionally include the VAE feature when conducting understanding tasks. In Self-SpectraReward, we use BAGEL's understanding backbone to provide reward signals for generation. Here, we ablate the effect of using the VAE feature in Self-SpectraReward. Table 7 shows that including the VAE feature obtains better performance on average, and the improvement is more significant on the TIIF-Long benchmark.

D Detailed Benchmark Results

This section provides the full category-level benchmark results that complement the aggregate numbers in Table 1. We include detailed breakdowns on GenEval, TIIF-Bench, DPG-Bench, and WISE to examine whether the gains from SpectraReward are concentrated in a single score or hold across different types of text-to-image requirements. Overall, the detailed results show that both SpectraReward and Self-SpectraReward improve the BAGEL baseline across compositional generation, fine-grained instruction following and knowledge-grounded generation. The improvements also remain visible when the same 512-resolution trained model is evaluated at 1024 resolution, suggesting that the learned reward signal transfers beyond the training sampling resolution.

GenEval. Table 8 reports the complete GenEval breakdown. Compared with BAGEL and AlphaGRPO, SpectraReward and Self-SpectraReward improve the overall score at both 512 and 1024 inference resolutions. The largest improvements appear in the more compositional categories, including counting, spatial position and color-attribute binding, where successful generation requires not only object presence but also correct attribute binding to objects. Self-SpectraReward achieves the best overall score at both resolutions, indicating that using the policy model's own understanding branch as the reward model can provide a particularly well-aligned signal for structured prompt following.

TIIF-Bench. Table 9 gives the full TIIF-Bench results on short and long prompts. TIIF-Bench separates basic following, advanced following, and designer-oriented prompts, making it a useful stress test for instruction fidelity beyond object-level matching. Both SpectraReward and Self-SpectraReward substantially improve the BAGEL baseline and AlphaGRPO on the overall short and long scores. The improvements are observed in basic attribute and relation following, advanced combinations of attribute/relation/reasoning, and the designer-oriented real-world split. The text subcategory remains challenging in absolute terms, but both methods increase the score by a large margin over BAGEL, suggesting that image-conditioned prompt likelihood provides useful supervision even for fine-grained instruction details.

Table 8: **Evaluation of text-to-image generation ability on GenEval benchmark.** 'Gen. Only' stands for an image generation model, and 'Unified' denotes a model that has both understanding and generation capabilities. † refers to the methods using the LLM rewriter. Our model's results and BAGEL all use the LLM rewriter.

Model	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attri.	Overall↑
Gen. Only Models							
PixArt-α [6]	98.0	50.0	44.0	80.0	8.0	7.0	48.0
Emu3-Gen [46]	98.0	71.0	34.0	81.0	17.0	21.0	54.0
SDXL [35]	98.0	74.0	39.0	85.0	15.0	23.0	55.0
DALL-E 3 [3]	96.0	87.0	47.0	83.0	43.0	45.0	67.0
SD3-Medium [10]	99.0	94.0	72.0	89.0	33.0	60.0	74.0
FLUX.1-dev† [24]	98.0	93.0	75.0	93.0	68.0	65.0	82.0
Unified Models							
SEED-X [12]	97.0	58.0	26.0	80.0	19.0	14.0	49.0
TokenFlow-XL [36]	95.0	60.0	41.0	81.0	16.0	24.0	55.0
ILLUME [44]	99.0	86.0	45.0	71.0	39.0	28.0	61.0
Transfusion [66]	-	-	-	-	-	-	63.0
Emu3-Gen† [46]	99.0	81.0	42.0	80.0	49.0	45.0	66.0
Show-o [57]	98.0	80.0	66.0	84.0	31.0	50.0	68.0
Janus-Pro-7B [7]	99.0	89.0	59.0	90.0	79.0	66.0	80.0
MetaQuery-XL† [34]	-	-	-	-	-	-	80.0
Inference on 512 resolution							
BAGEL	99.1	95.0	74.1	90.4	71.0	74.8	84.0
AlphaGRPO† (RT2I)	98.1	95.7	82.5	91.0	74.3	68.8	85.1
AlphaGRPO†	98.8	97.0	75.6	91.0	69.8	73.3	84.2
SpectraReward†	98.4	96.0	83.1	93.4	79.8	79.0	88.3
Self-SpectraReward†	98.4	97.7	84.4	93.6	81.5	81.5	89.5
Inference on 1024 resolution							
BAGEL	98.4	94.7	81.3	94.7	74.0	76.3	86.6
AlphaGRPO (RT2I)	98.8	95.2	82.8	93.6	76.3	77.8	87.4
AlphaGRPO	99.1	96.0	80.3	94.7	71.0	75.5	86.1
SpectraReward†	98.1	96.2	86.2	94.2	78.5	76.8	88.3
Self-SpectraReward†	98.4	95.2	88.8	95.7	82.3	78.3	89.8

DPG-Bench. Table 10 further evaluates prompt following with the decomposed categories of DPG-Bench. Since BAGEL is already a strong baseline on this benchmark, the absolute margins are smaller than those on TIIF-Bench. Nevertheless, SpectraReward obtains the best 512-resolution overall score and improves the global, relation, and other categories, while Self-SpectraReward gives the strongest attribute score. At 1024 resolution, Self-SpectraReward achieves the best overall score among the compared BAGEL-based methods. These results indicate that the reward learned from prompt likelihood helps not only aggregate instruction following, but also the relation- and attribute-sensitive subskills measured by DPG-Bench.

WISE. Table 11 reports the WISE benchmark, which emphasizes world-knowledge and semantic reasoning in text-to-image generation. When combined with self-CoT, however, both methods clearly improve over BAGEL and AlphaGRPO with the same inference strategy. Self-SpectraReward with self-CoT reaches the best overall WISE score and improves multiple knowledge domains, including space, biology, physics, and chemistry. This suggests that SpectraReward improves prompt-following ability and helps the model respond better to the CoT text.

E More Visualizations

Figure 6 provides additional examples of reward ranking in the same rollout group. Samples with higher SpectraReward reward better satisfy object, attribute, and spatial constraints, supporting the use of image-conditioned prompt likelihood as a group-wise ranking signal.

Figure 7 provides additional qualitative comparisons between the BAGEL baseline and Self-SpectraReward. The red text marks the key differences between the baseline BAGEL and our model, covering spatial relations, counting, attribute binding, relative size, and object co-occurrence

Table 9: **Performance of proprietary models and state-of-the-art open-source models on TIIF-Bench testmini subset.** Evaluated systems are grouped into (i) diffusion-based open-source models, (ii)autoregressive open-source models, and (iii) proprietary models. The results of SpectraReward and BAGEL are evaluated by GPT-4.1. "Inf. SRR" indicates executing the inference-time self-reflective refinement.

Model	Overall		Basic Following								Advanced Following												Designer	
			Avg		Attribute		Relation		Reasoning		Avg		Attribute +Relation		Attribute +Reasoning		Relation +Reasoning		Style		Text		Real World	
	short	long	short	long	short	long	short	long	short	long	short	long	short	long	short	long	short	long	short	long	short	long	short	long
Diffusion based Open-Source Models																								
FLUX.1 dev	71.09	71.78	83.12	78.65	87.05	83.17	87.25	80.39	75.01	72.39	65.79	68.54	67.07	73.69	73.84	73.34	69.09	71.59	66.67	66.67	43.83	52.83	70.72	71.47
SD XL	54.96	42.13	65.72	53.28	59.33	50.83	77.57	62.57	60.32	46.57	49.73	36.22	47.82	35.57	56.22	45.34	52.59	36.09	73.33	60.00	16.83	0.83	50.92	41.59
SD 3	67.46	66.09	78.32	77.75	83.33	79.83	82.07	78.82	71.07	74.07	61.46	59.56	61.07	64.07	68.84	70.34	50.96	57.84	66.67	76.67	59.83	20.83	63.23	67.34
SD 3.5 L	71.15	66.96	78.34	79.56	79.50	76.50	80.96	83.21	72.46	78.71	67.67	61.18	66.46	61.89	73.53	74.15	60.03	61.53	73.33	63.33	70.52	42.52	64.43	66.39
AR based Open-Source Models																								
Llamagen	41.67	38.22	53.00	50.00	48.33	42.33	59.57	60.32	51.07	47.32	35.89	32.61	38.82	31.57	40.84	47.22	49.59	46.22	46.67	33.33	0.00	0.00	39.73	35.62
Show-o	59.72	58.86	73.08	75.83	74.83	79.83	78.82	78.32	65.57	69.32	53.67	50.38	60.95	56.82	68.59	68.96	66.46	56.22	63.33	66.67	3.83	2.83	55.02	50.92
Infinity	62.07	62.32	73.08	75.41	74.33	76.83	72.82	77.57	72.07	71.82	56.64	54.98	60.44	55.57	74.22	64.71	60.22	59.71	80.00	73.33	10.83	23.83	54.28	56.89
JanusPro	66.50	65.02	79.33	78.25	79.33	82.33	78.32	73.32	80.32	79.07	59.71	58.82	66.07	56.20	70.46	70.84	67.22	59.97	60.00	70.00	28.83	33.83	65.84	60.25
Inference on 512 resolution																								
BAGEL	75.21	78.56	81.73	86.11	85.50	88.00	84.99	85.39	74.69	84.94	73.66	77.61	77.68	81.55	67.77	76.48	78.58	77.86	90.00	90.00	33.03	40.72	84.70	82.09
AlphaGRPO (RT2I)	78.92	79.48	85.46	84.15	88.50	85.50	88.12	86.56	79.77	80.38	77.41	78.85	81.05	82.77	74.38	80.52	79.30	75.83	90.00	83.33	44.80	53.85	84.33	86.57
AlphaGRPO	79.05	79.50	85.56	83.32	89.50	85.50	85.34	83.60	81.86	80.86	77.12	79.87	78.59	83.95	71.81	78.94	83.02	79.36	86.67	93.33	51.13	45.25	83.58	84.70
SpectraReward	85.23	84.81	87.67	87.44	90.00	90.50	88.12	89.45	84.90	82.38	84.85	85.39	84.34	87.52	84.99	84.95	86.41	86.02	93.33	93.33	66.97	58.82	88.06	90.30
Self-SpectraReward	85.13	84.33	89.69	87.58	88.00	92.00	93.56	87.37	87.50	83.38	83.73	85.70	85.17	86.63	83.05	90.72	84.65	82.38	93.33	93.33	60.63	57.01	90.30	86.19
Inference on 1024 resolution																								
BAGEL	76.42	76.15	83.44	83.72	87.50	89.00	86.03	84.35	76.77	77.81	75.16	76.68	79.34	83.12	70.38	74.58	78.36	75.58	93.33	86.67	36.20	40.72	79.85	73.51
AlphaGRPO (RT2I)	78.70	79.48	84.83	85.92	89.00	89.50	88.81	88.36	76.69	79.89	78.42	79.21	79.38	84.09	77.48	77.37	81.05	78.84	90.00	90.00	45.70	47.06	80.22	80.22
AlphaGRPO	77.74	78.09	85.39	82.90	89.00	89.50	87.31	82.96	79.85	76.25	75.62	77.49	79.46	83.32	73.28	76.13	76.48	75.56	90.00	90.00	42.53	44.80	81.72	84.33
SpectraReward	83.31	81.71	88.44	87.26	91.50	92.00	88.81	88.81	85.02	80.98	83.18	81.83	87.89	85.41	81.67	82.23	81.97	80.23	86.67	90.00	63.80	54.75	82.46	80.97
Self-SpectraReward	82.66	82.54	87.38	87.60	90.50	93.00	89.80	85.68	81.86	84.10	81.49	83.15	84.33	83.39	80.88	84.25	81.16	84.71	86.67	83.33	61.09	59.73	87.69	84.70

in long prompts. Across these cases, Self-SpectraReward more often preserves the requested relation between objects while also keeping the main visual content recognizable. For example, it places the toy car on modeling clay and separates the pen from the fabric blanket. These examples support the quantitative gains in instruction-following benchmarks by showing that the learned reward improves fine-grained prompt grounding rather than only global image quality.

F Broader Impacts

SpectraReward lowers the cost of constructing reward signals for image-generation RL by reusing frozen pretrained MLLMs, without preference annotation or reward-model fine-tuning. This may make reinforcement learning for text-to-image generation more accessible to researchers with limited annotation or training resources, and can support studies on prompt following, compositional generation, and unified multimodal self-improvement.

At the same time, easier RL optimization for image generation may amplify the existing risks of text-to-image models. Stronger prompt following can be misused for generating misleading visual content, imitating protected styles, or producing unsafe images. Since SpectraReward derives rewards from pretrained MLLMs, it may inherit potential biases and visual misjudgments from the reward MLLM backbone during policy optimization. For Self-SpectraReward, the closed-loop reward further encourages the model to align with its own understanding branch, which may not always reflect human preferences or safety requirements. We therefore recommend using SpectraReward together with standard safety filters, bias auditing, and human evaluation when deploying or releasing models trained with this reward.

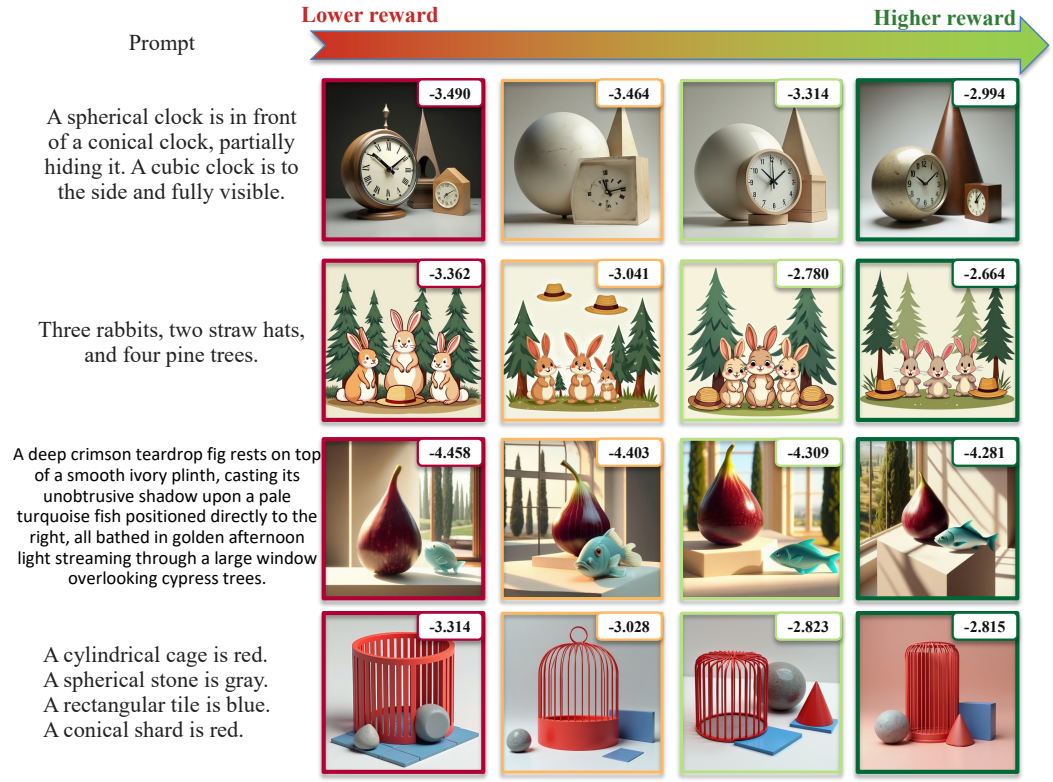

Figure 6: **Additional reward-ranking examples.** For each prompt, samples are ordered from lower to higher SpectraReward reward.

Table 10: **Evaluation of text-to-image generation ability on the DPG-Bench [15] benchmark.** * indicates our reproduced results.

Method	Global↑	Entity↑	Attribute↑	Relation↑	Other↑	Overall↑
Gen. Only Models						
Hunyuan-DiT [25]	84.59	80.59	88.01	74.36	86.41	78.87
PixArt-Σ [6]	86.89	82.89	88.94	86.59	87.68	80.54
DALLE3 [3]	90.97	89.61	88.39	90.58	89.83	83.50
SD3-medium [10]	87.90	91.01	88.83	80.70	88.68	84.08
FLUX.1-dev [24]	82.10	89.50	88.70	91.10	89.40	84.00
OmniGen [56]	87.90	88.97	88.47	87.95	83.56	81.16
Unified Multimodal Models						
Show-o [57]	79.33	75.44	78.02	84.45	60.80	67.27
EMU3 [46]	85.21	86.68	86.84	90.22	83.15	80.60
TokenFlow-XL [36]	78.72	79.22	81.29	85.22	71.20	73.38
Janus [49]	82.33	87.38	87.70	85.46	86.41	79.68
Janus Pro [7]	86.90	88.90	89.40	89.32	89.48	<u>84.19</u>
BLIP3-o 4B [5]	-	-	-	-	-	79.36
BLIP3-o 8B [5]	-	-	-	-	-	81.60
BAGEL [9]	88.94	90.37	91.29	90.82	88.67	85.07
UniWorld-V1 [27]	83.64	88.39	88.44	89.27	87.22	81.38
OmniGen2 [50]	88.81	88.83	90.18	89.37	90.27	83.57
Inference on 512 resolution						
BAGEL*	88.94	90.37	91.29	90.82	88.67	85.07
AlphaGRPO (RT2I)	89.99	**92.20**	88.49	90.89	89.12	85.98
AlphaGRPO	84.99	90.93	91.22	92.51	90.11	86.25
SpectraReward	**93.80**	92.17	90.78	**92.95**	**93.08**	**88.08**
Self-SpectraReward	92.40	91.94	**92.12**	92.77	88.57	87.73
Inference on 1024 resolution						
BAGEL*	87.42	**92.46**	90.75	91.92	84.96	85.17
AlphaGRPO (RT2I)	87.42	**92.46**	90.75	91.92	84.96	85.87
AlphaGRPO	**89.21**	89.43	90.20	**92.26**	90.39	85.08
SpectraReward	85.54	91.95	92.67	91.27	89.14	86.46
Self-SpectraReward	85.33	91.23	**93.14**	89.97	**92.06**	**86.97**

Table 11: **Comparison of world knowledge reasoning on WISE.** WISE examines the complex semantic understanding and world knowledge for T2I generation. 'Gen. Only' stands for an image generation model, and 'Unified' denotes a model that has both understanding and generation capabilities.

Type	Model	Cultural	Time	Space	Biology	Physics	Chemistry	Overall↑
Gen. Only	SDXL [35]	0.43	0.48	0.47	0.44	0.45	0.27	0.43
	SD3.5-large [10]	0.44	0.50	0.58	0.44	0.52	0.31	0.46
	PixArt-Alpha [6]	0.45	0.50	0.48	0.49	0.56	0.34	0.47
	FLUX.1-dev [24]	0.48	0.58	0.62	0.42	0.51	0.35	0.50
Unified	Janus [49]	0.16	0.26	0.35	0.28	0.30	0.14	0.23
	VILA-U [55]	0.26	0.33	0.37	0.35	0.39	0.23	0.31
	Show-o-512 [57]	0.28	0.40	0.48	0.30	0.46	0.30	0.35
	Janus-Pro-7B [7]	0.30	0.37	0.49	0.36	0.42	0.26	0.35
	Emu3 [46]	0.34	0.45	0.48	0.41	0.45	0.27	0.39
	MetaQuery-XL [34]	0.56	0.55	0.62	0.49	0.63	0.41	0.55
	BAGEL [9]	0.44	0.55	0.68	0.44	0.60	0.39	0.52
	BAGEL w/ Self-CoT [9]	0.76	0.69	0.75	0.65	0.75	0.58	0.70
	AlphaGRPO	0.44	0.55	0.64	0.46	0.62	0.46	0.53
	AlphaGRPO w/ Self-CoT	0.75	0.70	0.74	0.66	0.77	0.64	0.71
	SpectraReward	0.42	0.51	0.67	0.48	0.65	0.47	0.53
	SpectraReward w/ Self-CoT	0.76	0.70	0.80	0.72	0.80	0.65	0.74
	Self-SpectraReward	0.40	0.53	0.67	0.43	0.64	0.45	0.52
	Self-SpectraReward w/ Self-CoT	0.75	0.71	0.82	0.73	0.84	0.69	0.76

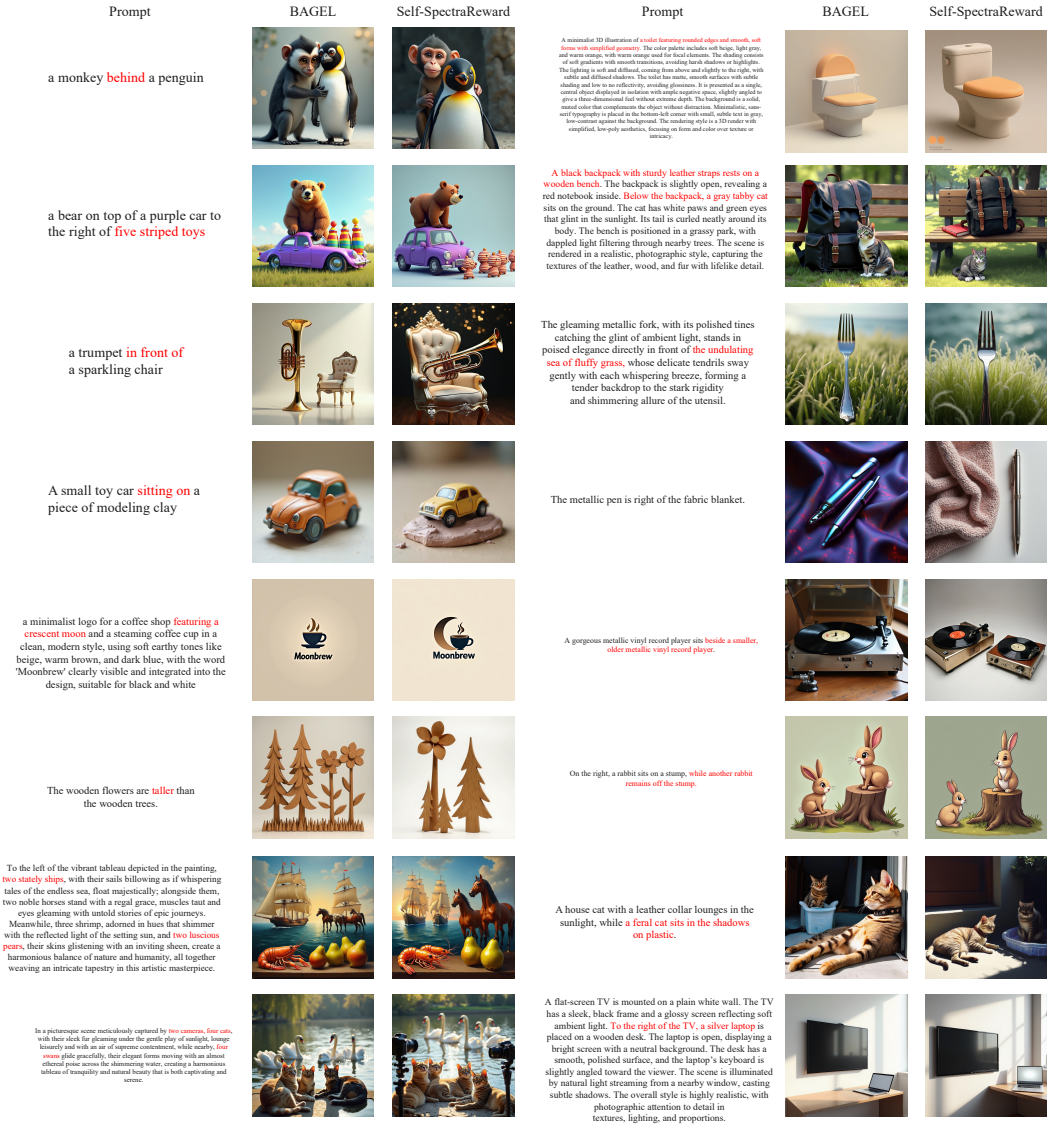

Figure 7: **Additional qualitative comparison.** Red text highlights the prompt constraints that are especially sensitive to spatial relation, counting, attribute binding, or long-prompt composition. Self-SpectraReward more consistently satisfies these highlighted constraints than the BAGEL baseline.