RubricRL: Simple Generalizable Rewards for Text-to-Image Generation

Xuelu Feng[1] Yunsheng Li[2] Ziyu Wan[2] Zixuan Gao[3] Junsong Yuan[1]
Dongdong Chen[2*] Chunming Qiao[1]

[1]University at Buffalo, USA [2]Microsoft AI, USA [3]Nikola Tesla STEM High School, USA *Corresponding Author
{xuelufen,jsyuan,qiao}@buffalo.edu, {yunshengli, ziyuwan,dongdong.chen}@microsoft.com, zixuan.go.1@gmail.com

Fig. 1: Visual examples of our RubricRL on two language backbones. Equipped with interpretable and user-controlled criteria, RubricRL improves SFT models' performance to generate high-quality images.

Abstract. Reinforcement learning (RL) has recently emerged as a promising approach for aligning text-to-image generative models with human preferences. A key challenge, however, lies in designing effective and interpretable rewards. Existing methods often rely on either composite metrics (*e.g.*, CLIP, OCR, and realism scores) with fixed weights or a single scalar reward distilled from human preference models, which can limit interpretability and flexibility. We propose RubricRL, a simple and general framework for rubric-based reward design that offers greater interpretability, composability, and user control. Instead of using a black-box scalar signal, RubricRL dynamically constructs a structured rubric for each prompt—a decomposable checklist of fine-grained visual criteria such as object correctness, attribute accuracy, OCR fidelity, and realism—tailored to the input text. Each criterion is independently evaluated by a multimodal judge (*e.g.*, o4-mini), and a prompt-adaptive weighting mechanism emphasizes the most relevant dimensions. This design not only produces interpretable and modular supervision signals for policy optimization (*e.g.*, GRPO or DiffusionNFT), but also enables

users to directly adjust which aspects to reward or penalize. Experiments with autoregressive and diffusion text-to-image models demonstrate that RubricRL improves prompt faithfulness, visual detail, and generalizability, while offering a flexible and extensible foundation for interpretable RL alignment across text-to-image architectures.

Keywords: Rubric reward · Reinforcement learning · Text-to-image generation

1 Introduction

Reinforcement learning (RL) has recently emerged as a promising approach for aligning generative models [3,7–9,21,32,33,36–38,40,42] with human preferences. In large language models, frameworks such as RLHF [30] and RLVF [35,56] have demonstrated that policy optimization guided by preference-based feedback can significantly enhance faithfulness, style, and usability. Extending this paradigm to text-to-image generation, including both autoregressive (AR) and diffusion architectures, offers a principled way to optimize models directly for human-aligned visual quality rather than likelihood-based objectives. However, the effectiveness of RL in visual domains critically depends on reward design: constructing evaluation signals that are accurate, interpretable, and generalizable across prompts, domains, and architectures remains a core challenge.

Existing text-to-image RL frameworks can be broadly categorized into multi-reward mixtures and unified scalar reward models. As shown in Fig. 2 (a), multi-reward systems (*e.g.*, X-Omni [15], AR-GRPO [57], Flow-GRPO [28], DanceGRPO [53]) combine heterogeneous objectives, such as CLIP-based image–text similarity [31], OCR accuracy [10], realism [52], and attribute consistency, to jointly encourage alignment and visual quality. While such approaches improve coverage, they depend on manually tuned weighting schemes that can be brittle across prompts and domains, and offer limited interpretability. Unified reward models (*e.g.*, OneReward [16], Pref-GRPO [45], LLaVA-Reward [60]) in Fig. 2 (b), instead learn a single scalar reward from pairwise human preference data. This simplifies optimization but can obscure the reasoning behind rewards, limit extensibility, and make it difficult for developers to control which visual aspects are prioritized.

In this paper, we propose RubricRL, a simple and general framework for rubric-based rewards design in text-to-image models as illustrated in Fig. 2 (c). Rather than relying on opaque scalar signals, RubricRL dynamically selects a structured rubric for each prompt, *i.e.*, a decomposable checklist of fine-grained visual criteria such as object correctness, attribute accuracy, OCR fidelity, compositional coherence, and realism. Each criterion is independently evaluated by a multimodal judge (*e.g.*, GPT-o4-mini), while a prompt-adaptive weighting mechanism highlights the most relevant dimensions. This produces interpretable, modular supervision signals that integrate naturally into policy optimization frameworks such as GRPO [20], PPO [34] or DiffusionNFT [58].

By expressing rewards in human-readable and decomposable form, RubricRL transforms reward evaluation from a black-box heuristic into an auditable process, where developers can directly inspect, extend, or adjust which aspects of generation are rewarded or penalized. The rubric structure also facilitates per-criterion diagnostics, providing transparency into model behavior and simplifying both evaluation and debugging.

RubricRL is architecture-agnostic and compatible with both diffusion and autoregressive text-to-image models. The rubric outputs further support variance-aware group advantages, leading to robust updates even under long-horizon rollouts. Its prompt-adaptive design ensures that each reward vector reflects the salient aspects of the input text, such as numerals, named entities, styles, or embedded text, without requiring manual tuning.

We validate this simple yet effective idea on an autoregressive (AR) text-to-image model, and further demonstrate its generalizability on diffusion-based generators. Experiments show that RubricRL improves prompt faithfulness, compositional accuracy, and visual realism, while maintaining high generalizability across architectures and datasets. Compared to prior multi-reward or unified-reward approaches, RubricRL achieves more consistent optimization behavior and enables controllable, interpretable reward shaping. Fig. 1 provides visualization samples of our method, illustrating high visual quality.

In summary, RubricRL contributes:

- A generalizable rubric-based reward design applicable to both AR and diffusion text-to-image models;
- A prompt-adaptive, decomposable supervision framework that enhances interpretability and composability;
- A developer-controllable and auditable framework that exposes rubric categories as prompt-level controls, allowing users to tailor reward dimensions via simple system-prompt edits.

By operationalizing alignment through dynamically generated rubrics of explicit visual criteria, RubricRL makes reinforcement learning for text-to-image generation more interpretable, extensible, and developer-guided, offering a unified foundation for aligning visual generation with human intent.

2 Related work

2.1 Text-to-Image Generation Methods.

Text-to-image (T2I) generation has seen significant progress through both diffusion based [6,18,29,33,48,50] and autoregressive (AR) architectures [12,51,55,59]. Diffusion models iteratively refine latent representations conditioned on text prompts, achieving high-quality and photorealistic images. Variants such as Stable Diffusion [33] and flow-based extensions [14,26,27] provide diverse styles, controllable generation, and strong fidelity at both global and local levels. Autoregressive approaches, on the other hand, represent images as sequences of discrete

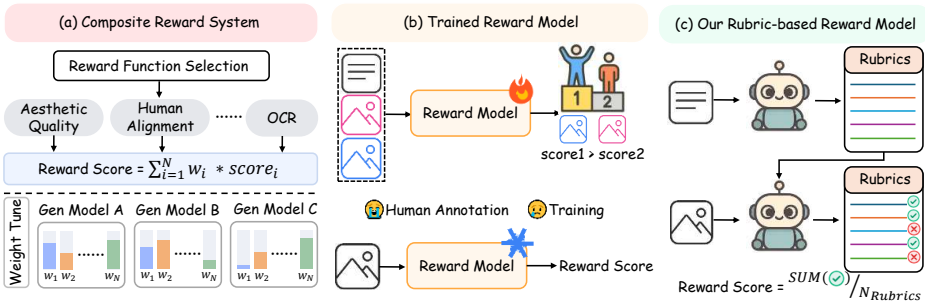

Fig. 2: Comparison of RubricRL with prior autoregressive (AR) reward formulations. (a) Multi-reward pipelines combine CLIP, OCR, alignment and realism metrics but require fragile weight tuning and often miss fine-grained attributes. (b) Unified scalar models collapse diverse objectives into a single learned score, simplifying optimization but reducing interpretability and adaptability. (c) RubricRL replaces both with a decomposable, prompt-adaptive rubric—an explicit checklist of visual criteria (counting, attributes, OCR/text fidelity, realism). Each criterion is scored independently and integrated into RL framework to provide interpretable, variance-aware supervision that improves detail, prompt faithfulness, and debuggability.

tokens and model the joint distribution of text and image tokens using a single transformer backbone. Early hybrid designs, such as DreamLLM [12], paired AR text encoders with separate diffusion decoders. More recent unified AR models, including Chameleon [25], Emu3 [44], TransFusion [59], and Janus [51], integrate visual tokenization and autoregressive modeling in one architecture. These models allow direct mapping between text tokens and visual outputs, enabling flexible control and fine-grained generation. In this paper, we propose a novel reward design for reinforcement learning in text-to-image models, and demonstrate their effectiveness using both unified AR and diffusion text-to-image models.

2.2 Reinforcement Learning for Text-to-Image Generation.

Maximum-likelihood training often under-optimizes user-salient qualities, such as semantic faithfulness, compositional accuracy, and aesthetics. Reinforcement learning (RL) offers task-aligned feedback that directly optimizes for human-relevant properties beyond likelihood. In diffusion-based text-to-image models, RL methods, such as Flow-GRPO [28], DanceGRPO [53], and DifusionNFT [58], have improved alignment by fine-tuning the velocity with preference or metric-based rewards. Recently, RL has also been applied to unified AR T2I models [43], where policy gradients act directly on next-token probabilities, enabling end-to-end credit assignment and fine-grained control over generated images.

The design of the reward function is central to effective reinforcement learning in text-to-image models. One line of work aggregates heterogeneous signals—such as CLIP-based image–text alignment [31], OCR/text correctness [10, 47], multimodal VLM judges (*e.g.*, Qwen2.5-VL-32B [4]), aesthetic and realism

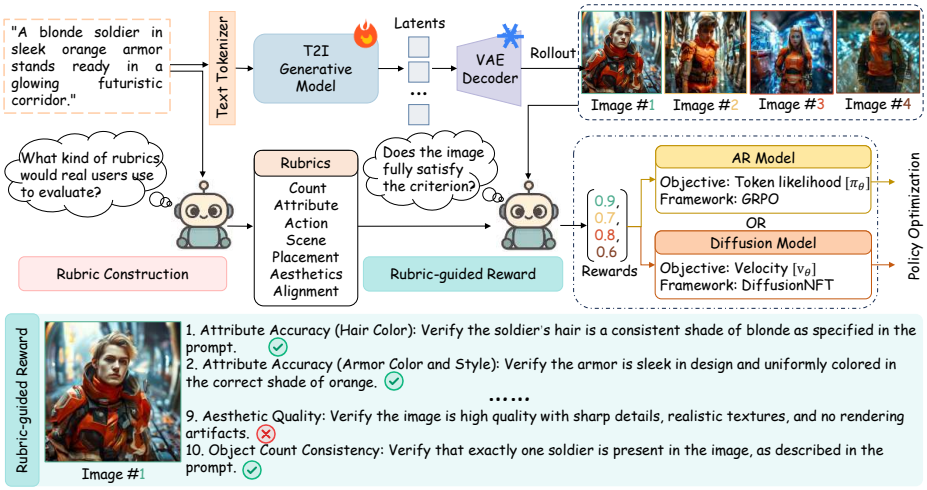

Fig. 3: Overview of the proposed method. We propose a simple, general rubric generation pipeline and rubric-based reward model for text-to-image generation.

metrics [54], and human-preference surrogates [52]. While comprehensive, these multi-reward mixtures demand careful weighting and tuning, which can destabilize optimization and obscure per-aspect failures. Another direction trains unified preference models [45, 46, 60] to predict a single scalar human-aligned score from paired image outputs, simplifying optimization but relying on costly human annotations and limited scalability. In this work, we propose a rubric-based reward that is simple, generalizable, and interpretable. For each prompt, a compact rubric defines aspect-wise criteria—such as text alignment/OCR accuracy, object count, spatial relations, and overall coherence/quality. Each criterion is scored independently by a dedicated evaluator, and a transparent aggregation produces the final reward. This design is more prompt-adaptive, decomposable, and interpretable, while providing user-controllable and auditable feedback. While several concurrent works [19, 24] investigate rubric-based rewards in natural language processing, to the best of our knowledge, we are the first to propose rubric based rewards in text-to-image RL.

3 Method

In this section, we present RubricRL as a general post-training framework for text-to-image generation that applies to both autoregressive (AR) and diffusion-based backbones (Fig. 3). The core components—rubric construction and rubric-guided reward—are shared across model families, while the RL optimizer differs (GRPO for AR and DiffusionNFT for diffusion). For clarity, we describe the method using an AR instantiation in the main text and defer diffusion-specific implementation details to the supplementary material.

3.1 Overall Architecture

As illustrated in Fig. 3, given an input text prompt p, we first tokenize it into a sequence of text tokens and feed them into a backbone text-to-image generator π_θ (either an AR token generator or a diffusion model) to produce a generated image I. In this paper, we primarily focus on post-RL fine-tuning of π_θ to further enhance its output quality, where designing an effective, reliable, and interpretable reward function is the key challenge. Existing methods typically employ one or multiple specialized models to evaluate different aspects of image quality, such as CLIP-based image–text semantic alignment reward [31] ($R_{\text{clip}}(I, p)$), OCR accuracy [10] ($R_{\text{ocr}}(I, p)$), and realism [52]. However, this approach has notable drawbacks: (1) deploying multiple specialized models is computationally expensive and difficult to scale to additional aspects; (2) it requires careful reward calibration and reweighting. Recent works have attempted to learn a single reward model from pairwise human preference data, simplifying optimization but offering limited extensibility due to high annotation costs and poor interpretability.

Motivated by the strong multimodal understanding capabilities of modern vision–language models (VLMs), we propose a simple and unified *rubric-based reward model*, denoted $R_{\text{rubric}}(I, p, \mathcal{C}(p))$. This model replaces the ensemble of task-specific evaluators with a single reasoning-capable grader. Rather than relying on fixed sub-models, our approach automatically constructs a set of interpretable, prompt-adaptive criteria termed *"rubrics"* that capture the essential quality requirements for each prompt p. In detail, given a prompt p, a *Rubric Generation Model* $\mathcal{G}$ generates a set of evaluation rubrics $\mathcal{C}(p) = \mathcal{G}(p)$, where $\mathcal{C}(p) = \{c_1, c_2, \ldots, c_M\}$ defines M prompt-specific criteria spanning dimensions such as object count, attribute accuracy, text/OCR fidelity, spatial relations, aesthetics, and style consistency.

In reinforcement learning (RL), the objective is to adjust the model parameters θ to maximize the expected rubric reward over the distribution of prompts:

$$\max_{\theta} \ \mathbb{E}_{p\sim\mathcal{D},\, I\sim\pi_\theta(\cdot|p)}\big[\, R_{\text{rubric}}(I, p, \mathcal{C}(p))\,\big], \tag{1}$$

where $\mathcal{D}$ denotes the set of prompts and a *rollout* corresponds to one sampled image from π_θ given p. Compared to composite reward systems, our rubric-based formulation offers three key advantages: (1) **Simplicity**: it eliminates the need for multiple task-specific graders; (2) **Adaptivity**: rubrics are dynamically generated for each prompt, ensuring relevance to diverse intents; and (3) **Interpretability**: each reward component aligns with a human-readable evaluation criterion, enabling transparent model diagnostics and controllable optimization.

3.2 Rubric-based Reward

The rubric based reward function proceeds in two stages. First, a *Rubric Generation Model* $\mathcal{G}$ interprets the user prompt p and produces a set of candidate evaluation rubrics $\mathcal{C}(p)$. Second, a multimodal LLM grader implements the *Rubric-Based Reward* $R_{\text{rubric}}(I, p, \mathcal{C}(p))$ that scores a generated image I against each

rubric in $\mathcal{C}(p)$. In this paper, we employ GPT-o4-mini to fulfill both roles, generating prompt-specific rubrics and providing per-criterion judgments that are aggregated into a scalar reward.

Rubric construction. Given a user prompt p, we ask GPT-o4-mini to generate a list of rubrics. Each rubric entry contains a short eval key that targets a specific aspect (*e.g.*, OCR alignment, object count, spatial relations, aesthetics) and a concise description of what to check in the image.

To promote diversity and reduce positional bias during rubric generation, we randomly permute the evaluation aspects in the rubric generation prompt and query GPT-o4-mini multiple times. In each round, the model produces a set of rubrics (we request 10 per query; because a prompt may describe multiple objects or attributes, the model may output multiple rubrics for one eval key to ensure adequate coverage). We aggregate all valid key–criterion pairs across runs into a unified rubric pool, discarding ambiguous or malformed entries. Finally, to remove redundancy and focus on the most important signals, we ask GPT-o4-mini to choose the top-10 most relevant and critical criteria for evaluating images generated from the user prompt p.

Rubric-guided reward. Given a generated image I, its corresponding text prompt p and the rubric pool $\mathcal{C}$, we simply ask GPT-o4-mini again to output a single score $y_i \in \{0, 1\}$ for each criterion to reflect whether the generated image fully satisfies this rubric ($y_i = 1$) or not ($y_i = 0$). The overall rubric reward is computed as the normalized mean of:

$$R(I, p, \mathcal{C}) = \frac{1}{M} \sum_{i=1}^{M} y_i, \quad M = 10 \tag{2}$$

3.3 Reinforcement Learning with GRPO

To align the autoregressive image generator with rubric-based rewards, we employ **Group Relative Policy Optimization (GRPO)** [35], a variant of PPO designed for stable optimization over grouped rollouts. For diffusion-based models, we adopt **DiffusionNFT** [58] as the RL strategy; details are provided in the supplementary material. For each prompt, the set of generated rollouts forms a group, and the reward of each rollout is normalized relative to the group to reduce variance and improve credit assignment. Concretely, let π_θ denote the current policy and R_i the rubric reward for rollout i in group g. GRPO computes the relative advantage

$$A_i = \frac{R_i - \bar{R}_g}{\sqrt{\frac{1}{|g|-1} \sum_{j \in g} \left(R_j - \bar{R}_g\right)^2}}, \ \bar{R}_g = \frac{1}{|g|} \sum_{k \in g} R_k, \tag{3}$$

and updates the policy by maximizing a clipped objective similar to PPO:

$$\mathcal{L}(\theta) = \mathbb{E}_i \Big[\min \left(r_i(\theta) A_i, \ \text{clip}(r_i(\theta), 1 - \epsilon, 1 + \epsilon) A_i \right) \Big], \tag{4}$$

where $r_i(\theta) = \frac{\pi_\theta(a_i|s_i)}{\pi_{\theta_{\text{old}}}(a_i|s_i)}$, a_i and s_i are the sampled action and state corresponding to rollout i, and ϵ is the PPO clipping parameter. By leveraging this group-relative advantage, GRPO stabilizes training across prompts, making the model robust to heterogeneous reward scales and noisy evaluations. Combined with our rubric-based reward and dynamic rollout selection strategy described below, we find GRPO can effectively guide the generative model toward images that are both human-aligned and high-quality.

3.4 Dynamic Rollout Sampling

As discussed above, the target policy model π_θ in GRPO explores the generation space by sampling multiple rollouts, each yielding a reward R_i used for advantage computation. In the original GRPO design, all N rollouts from a single prompt are grouped together for policy updates, *i.e.*, $|g| = N$. Subsequent works introduce over-sampling and filtering strategies to improve training efficiency. For instance, DAPO [56] adopts a *prompt-level* over-sampling approach: it generates N rollouts per prompt and discards prompts whose rollouts all have accuracy 1 or 0, thereby retaining only moderately difficult prompts for policy optimization. Formally, DAPO selectively sample prompts used in training while still using all rollouts from each retained prompt for RL updates.

In this paper, we propose a new *rollout-level* dynamic sampling mechanism, where selection occurs within the rollouts of a single prompt rather than filtering entire prompts. Specifically, given a text prompt, instead of sampling only N rollouts, we oversample N' rollouts ($N' > N$) and selectively use a subset of N representative rollouts for policy updates. To balance quality and diversity, we adopt a hybrid selection strategy: we take the top-K high-reward rollouts and randomly sample the remaining $N - K$ rollouts from the others to encourage diversity. Formally, the rollout group g is constructed as

$$g = \{\tau_{(1)}, \ldots, \tau_{(K)}\} \cup \text{RS}\big(\{\tau_{(K+1)}, \ldots, \tau_{(N')}\}, N - K\big), \tag{5}$$

where RS denotes random sampling. Empirically, we observe this hybrid design achieves a better balance between stability and diversity, achieving better model quality. As a result, the loss in Eq. 4 is computed over a more representative and informative subset of rollouts, leading to more consistent and efficient learning compared to both the original GRPO and the prompt-level filtering in DAPO.

4 Experiments

4.1 Implementation Details

Following SimpleAR [43], we construct a training set of 11,000 images sampled from JourneyDB [39] and Synthetic-1M [13]. We re-caption all images with GPT-o4-mini to obtain diverse prompt lengths, and randomly sample one caption per image during training. All experiments are conducted on 8 NVIDIA A100 GPUs.

Table 1: Evaluation of text-to-image generation trained on Phi3 (3.8B) and Qwen2.5 (0.5B) as AR backbone on the GenEval.

Method	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attr.	Overall
Phi3 (3.8B) [1]	0.9938	0.8939	0.4562	0.9255	0.7200	0.5850	0.7624
+ CLIPScore [31]	0.9938	0.9242	0.5250	0.9362	0.7725	0.7000	0.8086
+ HPSv2 [52]	0.9906	0.8813	0.5125	0.9441	0.7675	0.7205	0.8035
+ Unified Reward [46]	0.9969	0.9318	0.4156	0.9388	0.8150	0.7000	0.7997
+ LLaVA-Reward-Phi [60]	0.9844	0.8864	0.4719	0.9176	0.7250	0.5975	0.7638
+ AR-GRPO [57]	0.9938	0.8712	0.5406	**0.9574**	0.8075	0.6300	0.8001
+ X-Omni [15]	0.9969	0.9192	0.4719	0.9548	0.8175	0.6875	0.8080
+ RubricRL	**1.0000**	**0.9343**	**0.6125**	0.9415	**0.8275**	**0.7650**	**0.8468**
Qwen2.5 (0.5B) [41]	0.9625	0.5303	0.2500	0.7606	0.3575	0.2825	0.5239
+ CLIPScore [31]	0.9656	0.6162	0.2750	**0.8404**	0.3825	0.3250	0.5674
+ HPSv2 [52]	0.9750	0.6465	0.2438	0.8005	0.3875	0.2825	0.5560
+ Unified Reward [46]	0.9625	0.6288	0.2656	0.8191	0.4050	0.3550	0.5727
+ LLaVA-Reward-Phi [60]	0.9625	0.5303	0.2500	0.7606	0.3575	0.2825	0.5239
+ AR-GRPO [57]	0.9656	0.5682	**0.2969**	0.8378	0.3825	0.3050	0.5593
+ X-Omni [15]	0.9812	0.5960	0.2219	0.8085	0.4125	0.3200	0.5567
+ RubricRL	**0.9844**	**0.6616**	0.2469	0.8378	**0.4825**	**0.3950**	**0.6014**

Autoregressive (AR) backbones. We evaluate two AR backbones: Phi3-3.8B [1] and Qwen2.5-0.5B [41], both initialized from SFT checkpoints. For image decoding, we use LlamaGen's VQ decoder [40] for Phi3-3.8B and Cosmos-Tokenizer [2] for Qwen2.5-0.5B. The output resolutions are 512 and 1024, respectively. RL training uses the TRL framework [49] with learning rate 1e−5, warm-up ratio 0.1, batch size 28, and 3 epochs. We use dynamic rollout sampling by selecting 4 candidates from 16 rollouts per prompt. During inference, we apply classifier-free guidance (CFG) [22] with scale 7.5.

Diffusion backbones. For diffusion-based text-to-image models, we additionally fine-tune Qwen-Image [50] and FLUX-1.0 [dev] [5] using DiffusionNFT [58]. We adopt LoRA with rank r=32 and scaling factor α=64. Each epoch consists of 48 groups with group size 24. We use 12 rollout sampling steps, and 40 denoising steps. We set the learning rate to 3e−4 and batch size to 3. In inference stage, CFG is applied to Qwen-Image with a scale of 7.5, and we use the default guidance scale of 3.5 for FLUX.

4.2 Comparing with State-of-the-arts

We compare RubricRL with representative reward modeling strategies for RL post-training on two widely used benchmarks, DPG-Bench [23] and GenEval [47], covering both autoregressive (AR) and diffusion-based text-to-image generators. The compared methods can be grouped by reward design: (i) *single* specialized reward models, including CLIPScore [31], HPSv2 [52], Unified Reward [46], and LLaVA-Reward-Phi [60]; and (ii) *composite* reward metrics with fixed weights, such as AR-GRPO [57] and X-Omni [15]. For fair comparison, we implement these baselines under the same RL framework and hyperparameters as ours, and only vary the reward function. We also report the initial SFT models to isolate the gains brought by RL.

Table 2: Evaluation of text-to-image generation trained on Phi3 (3.8B) and Qwen2.5 (0.5B) as AR backbone on the DPG-Bench.

Method	Global	Entity	Attribute	Relation	Other	Overall
Phi3 (3.8B) [1]	**84.80**	87.90	88.18	93.30	82.00	81.25
+ CLIPScore [31]	81.76	89.95	89.42	93.50	86.00	84.15
+ HPSv2 [52]	82.98	90.71	89.94	93.19	87.60	84.85
+ Unified Reward [46]	82.37	89.94	89.50	93.93	85.20	84.06
+ LLaVA-Reward-Phi [60]	82.98	88.06	87.83	92.50	79.20	81.51
+ AR-GRPO [57]	82.37	89.05	90.00	93.08	**88.00**	83.81
+ X-Omni [15]	84.19	89.66	89.12	93.69	86.00	84.05
+ RubricRL (Ours)	83.28	**91.88**	**90.07**	**94.73**	85.20	**86.07**
Qwen2.5 (0.5B) [41]	84.78	84.74	86.41	87.34	84.27	78.02
+ CLIPScore [31]	80.55	87.22	86.19	91.33	67.60	79.78
+ HPSv2 [52]	78.42	87.29	85.04	91.45	68.80	80.23
+ Unified Reward [46]	79.03	87.09	85.24	90.68	69.20	79.69
+ LLaVA-Reward-Phi [60]	**84.78**	84.74	86.41	87.34	**84.27**	78.02
+ AR-GRPO [57]	80.24	86.75	85.95	**92.02**	69.35	79.74
+ X-Omni [15]	79.33	87.03	85.34	91.72	72.40	79.92
+ RubricRL (Ours)	79.33	**88.48**	**86.55**	91.37	68.00	**81.43**

Table 3: Comparison on diffusion-based generators. RubricRL consistently improves the diffusion-based Qwen-Image (4B) and FLUX (12B) baselines under the same training setting in DiffusionNFT.

Method	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attr.	Overall
Qwen-Image (4B) [50]	1.0000	0.9242	0.7125	0.9521	0.7750	0.6800	0.8406
+ CLIPScore	0.9938	0.9394	0.7219	0.9628	0.7625	0.7325	0.8521
+ HPSv2	0.9594	0.9293	0.6500	0.9122	0.6550	0.6500	0.7937
+ UnifiedReward	0.9688	0.9217	0.7094	0.9229	0.6750	0.7025	0.8167
+ LLaVA-Reward-Phi	0.9969	0.9268	0.7344	0.9574	0.7250	0.7075	0.8405
+ AR-GRPO	0.9844	0.9369	0.7719	0.9441	0.7450	0.7275	0.8516
+ X-Omni	0.9938	0.9369	0.7438	0.9441	0.7725	0.7300	0.8535
+ RubricRL (Ours)	**1.0000**	**0.9646**	**0.8562**	**0.9734**	**0.8275**	**0.8325**	**0.9091**
FLUX.1 Dev (12B) [5]	0.9844	0.9293	0.7531	0.9255	0.6900	0.5900	0.8121
+ CLIPScore	0.9875	0.9697	0.8062	0.9521	0.7250	0.6700	0.8518
+ HPSv2	0.9875	0.9444	0.6906	0.9096	0.6325	0.6525	0.8024
+ UnifiedReward	0.9906	0.9571	0.7688	0.9282	0.6700	0.6525	0.8279
+ LLaVA-Reward-Phi	0.9719	0.9697	0.7094	0.9362	0.6750	0.6850	0.8245
+ AR-GRPO	0.9750	0.9672	0.7188	0.9255	0.6850	0.7075	0.8298
+ X-Omni	0.9844	0.9495	0.7656	0.9335	0.7050	0.6625	0.8334
+ RubricRL (Ours)	**0.9969**	**0.9747**	**0.8438**	**0.9707**	**0.7800**	**0.7975**	**0.8969**

Autoregressive backbones. We evaluate the above reward baselines on two AR SFT backbones (Phi3 and Qwen2.5). Quantitative results are reported in Table 1 (GenEval) and Table 2 (DPG-Bench). For GenEval, we apply prompt rewriting following [11] to ensure evaluation consistency. For DPG-Bench, since LLM-based generators are trained with long prompts, we also rewrite prompts for a consistent evaluation interface. From the results, all RL post-trained methods consistently outperform the SFT baseline, confirming the benefit of reinforcement learning in enhancing image generation quality. And RubricRL achieves the best performance, surpassing X-Omni by approximately 4% on GenEval on both LLM backbones, highlighting the effectiveness and generalization of our rubric-based reward.

Diffusion backbones. Beyond the AR setting, we further evaluate whether RubricRL generalizes to diffusion-based generators under the same benchmarks.

Table 4: Evaluation of text-to-image generation trained on Qwen-Image (4B) and FLUX.1 Dev (12B) as diffusion-based flow matching backbone on the DPG-Bench.

Method	Global	Entity	Attribute	Relation	Other	Overall
Qwen-Image (4B) [50]	78.72	86.47	88.68	93.00	81.60	82.40
+ CLIPScore [31]	80.55	92.16	**90.40**	93.81	86.00	85.09
+ HPSv2 [52]	78.72	92.23	87.85	93.73	85.60	85.26
+ Unified Reward [46]	79.94	92.15	90.30	93.93	**86.40**	85.96
+ LLaVA-Reward-Phi [60]	82.07	89.31	88.88	92.69	86.00	81.29
+ AR-GRPO [57]	81.54	91.10	89.47	93.31	83.30	84.93
+ X-Omni [15]	83.13	91.83	89.80	93.97	84.30	85.80
+ RubricRL (Ours)	**82.37**	**92.63**	89.88	**94.82**	82.80	**86.23**
FLUX.1 Dev (12B) [5]	81.46	76.91	79.32	84.91	50.00	63.99
+ CLIPScore [31]	79.03	76.06	80.06	86.69	49.20	65.82
+ HPSv2 [52]	79.94	**79.19**	**80.55**	86.07	51.60	66.65
+ Unified Reward [46]	82.07	78.79	80.33	87.35	53.20	66.40
+ LLaVA-Reward-Phi [60]	79.03	77.35	78.90	85.07	48.40	65.20
+ AR-GRPO [57]	80.55	78.89	80.55	85.88	53.20	66.56
+ X-Omni [15]	82.07	77.79	79.44	85.53	50.80	65.24
+ RubricRL (Ours)	**82.98**	77.93	79.74	**86.92**	**54.80**	**67.70**

Table 5: Comparison of dynamic rollout-sampling strategies used in GRPO.

Method	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attr.	Overall
Vanilla	0.9906	0.9217	0.6031	0.9441	0.7975	0.7525	0.8349
FFKC-1D [17]	1.0000	0.9268	0.5656	**0.9441**	0.8250	0.7500	0.8353
DAPO [56]	0.9969	0.9293	0.6125	0.9362	0.7975	0.7275	0.8333
Hybrid (Ours)	**1.0000**	**0.9343**	**0.6125**	0.9415	**0.8275**	**0.7650**	**0.8468**

As shown in Table 3, RubricRL yields consistent gains for both diffusion backbones, improving over the Qwen-Image baseline by +6.85% and over the FLUX baseline by +8.48%. We further report DPG-Bench results in Table 4, where RubricRL brings smaller but still consistent improvements. These results indicate that RubricRL is a simple and effective plug-and-play component for reward-driven optimization across model families.

4.3 Ablation Study

In this section, we present ablation studies and additional analyses conducted under the autoregressive (AR) setting. Unless otherwise specified, all experiments are performed on the Phi3 backbone and evaluated on GenEval.

Strategies for dynamic rollout sampling. To investigate the impact of different selection strategies used by dynamic rollout sampling, we compare four methods, *i.e.*, RubricRL without dynamic rollout sampling (Vanilla), FFKC-1D [17], DAPO [56], and our proposed hybrid strategy, and report the results in Table 5. Specifically, FFKC-1D also oversamples more rollouts and then keeps a diverse subset by first selecting a medoid (the rollout with reward closest to the median) and then greedily adding samples that maximize reward distance from already chosen ones. Compared to our hybrid strategy, FFKC-1D focuses too much on the diversity and ignore the importance of high quality rollouts. As

Table 6: Comparison of advantage computation in GRPO, with performance measured on GenEval.

Method	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attr.	Overall
RubricRL_GN	0.9938	0.9268	**0.6781**	0.9362	0.7850	0.6825	0.8337
RubricRL_LN	**1.0000**	**0.9343**	0.6125	**0.9415**	**0.8275**	**0.7650**	**0.8468**

shown in Table 5, our hybrid sampling strategy consistently achieves the best performance, surpassing both FFKC-1D and DAPO as well as the Vanilla baseline that directly uses four rollouts without any dynamics. Interestingly, FFKC-1D and DAPO do not outperform the vanilla baseline, suggesting that their dynamic prompt sampling and pure rollout diversity-driven sampling strategies fail to provide additional informative signals for RL. In contrast, our hybrid strategy effectively balances exploitation of high-reward rollouts and exploration of diverse candidates, enabling the policy model to leverage both higher-quality and diverse samples, resulting in more effective RL signal.

Normalization scope for advantages. In Eq. 3, the advantage used in GRPO is computed by normalizing rewards—using the mean and standard deviation—within a group of rollouts. Under our dynamic sampling strategy, only N rollouts are retained out of N' candidates. This raises an important design choice: should the normalization statistics (mean and standard deviation) be computed using all N' rollouts or only the retained N? We denote these two norm variants as "RubricRL_GN" and "RubricRL_LN", respectively. In Table 6, it reflects that "RubricRL_LN" yields better performance. This is because normalizing within the retained subset better reflects the actual reward distribution that guides learning, preventing high-variance or low-quality rollouts from distorting the gradient direction.

RubricRL v.s. SFT with Best-of-N sampling. We further compare the proposed RubricRL with the SFT model equipped with a Best-of-N sampling strategy during inference($N = 8$), which has been observed in prior work [15] to form an 'upper bound' for RL methods in language tasks. Specifically, for each prompt in GenEval, we first generate a rubric, then sample 8 rollouts from the SFT model. Each rollout is scored using the rubric-based reward, and the top 4 are selected for evaluation on GenEval. As shown in Table 7, although Best-of-N sampling significantly can achieve higher scores, RubricRL still achieves a notable improvement, exceeding it by over 5%. This result aligns with the observations in X-Omni [15], reconfirming that reinforcement learning provides a more effective optimization paradigm.

Failure case analysis. As the grader, although GPT-o4-mini is highly general and powerful in evaluating the quality of generated images, we observe that it can assign incorrect scores—*e.g.*, underestimating or overestimating object counts, particularly when the base model's generation quality is poor. Fig. 4 illustrates several typical failure cases in the counting subcategory of GenEval, such as redundant poles near traffic lights, intertwined bicycles, and overlapping zebras. These challenging scenarios often mislead GPT-o4-mini, resulting in inaccurate

Table 7: Comparison on GenEval of Best-of-N ($N = 8$) and our RL training for Phi3-3.8B (P-*) and Qwen2.5-0.5B (Q-*).

Method	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attr.	Overall
P-SFT model	0.9938	0.8939	0.4562	0.9255	0.7200	0.5850	0.7624
P-Best-of-N	0.9812	0.9091	0.6000	0.9309	0.7450	0.5900	0.7927
P-RubricRL	**1.0000**	**0.9343**	**0.6125**	**0.9415**	**0.8275**	**0.7650**	**0.8468**
Q-SFT model	0.9625	0.5303	0.2500	0.7314	0.3575	0.2825	0.5239
Q-Best-of-N	0.9562	0.6566	**0.3750**	0.8138	0.4100	0.3600	0.5953
Q-RubricRL	**0.9844**	**0.6616**	0.2469	**0.8378**	**0.4825**	**0.3950**	**0.6014**

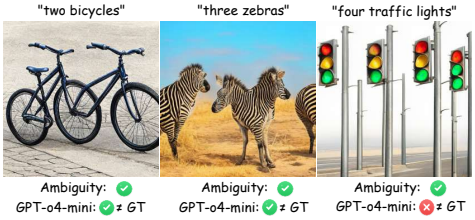

Fig. 4: Failure cases of GPT-o4-mini when grading counting on GenEval. The model misjudges instance counts under ambiguity. The grader's score deviates from the GT and thus from common-sense human judgment.

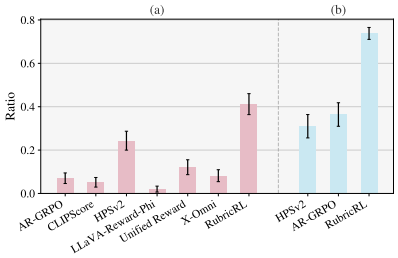

Fig. 5: Visualization of human preference and judge alignment analysis. (a) Top-1 preference ratio. (b) Human preference alignment ratio.

counts. However, when the base model generates higher-quality images, this issue is less pronounced. This explains why RubricRL's performance on the "counting" subcategory in GenEval and the "Other" subcategory in DPG-Bench—both containing many counting cases—is worse than the baseline SFT model when using Qwen2.5-0.5B as the base model. In contrast, with Phi3-3.8B, this issue almost disappears, allowing RubricRL to substantially improve performance in counting-related categories.

Rubric reliability and grader correctness. To validate our rubric-based training signal, we run a human study on rubric quality and grader reliability. We sample 100 prompts, generate 10 rubrics per prompt, and evaluate 1,000 prompt–rubric pairs with ten annotators for (i) prompt–rubric consistency, (ii) rubric hallucination, and (iii) agreement between GPT-o4-mini and human judgments. We find that 98.1% of rubrics are consistent with the prompt and only 0.95% contain hallucinated constraints. Moreover, GPT-o4-mini matches human judgments in 94.8% of cases, supporting its use as a reliable grader.

Human preference and judge alignment analysis. We conduct two user studies: (1) human preference among final models trained with different rewards, and (2) agreement between human and reward-model judgments. For (1), we sample 100 prompts, generate outputs from each method, and ask users to pick the best image. Fig. 5(a) shows RubricRL is chosen most often, indicating higher human preference. For (2), we compare RubricRL with HPSv2 (unified) and AR-GRPO (scalar): each method produces two images per prompt, users select the

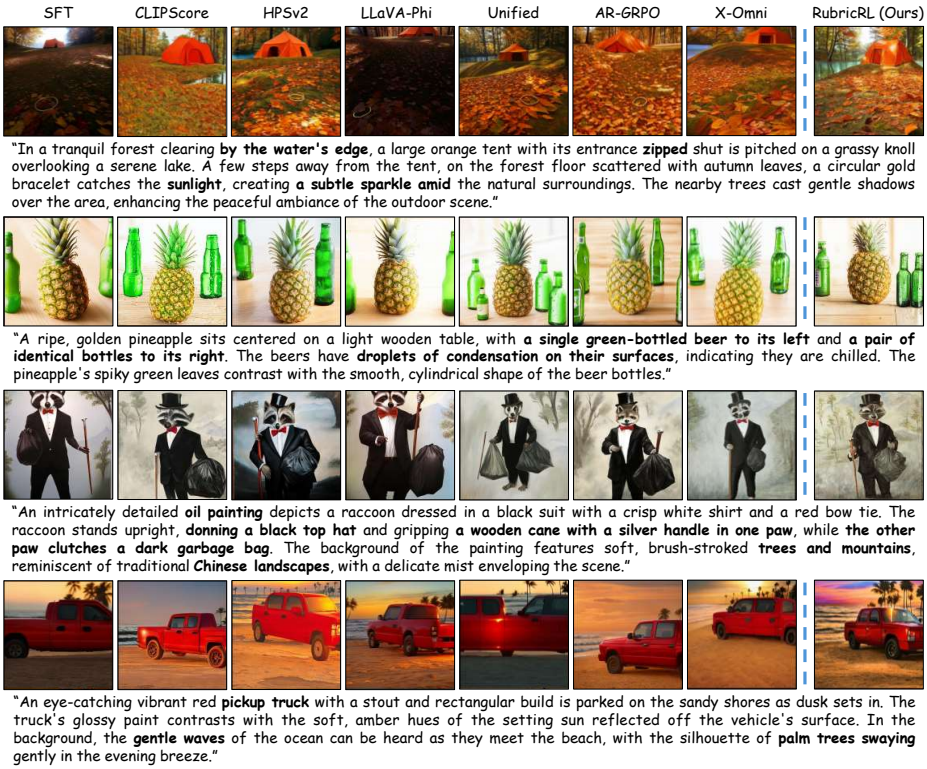

Fig. 6: Qualitative comparison: we visualize RubricRL and baseline models using prompts from DPG. RubricRL shows superior image quality that is both aesthetically pleasing and better aligned with the prompt. The **bold text** highlights key elements that RubricRL successfully captures, while baseline models often fail to generate these details accurately.

better one, and we compute agreement with the model judge. Fig. 5(b) shows RubricRL achieves the highest agreement with lower variance.

4.4 Visual results

We further present comprehensive visual comparisons between RubricRL and other baseline approaches in Fig. 6. As illustrated, models trained with RubricRL consistently produce images that are not only more aesthetically appealing but also demonstrate superior semantic alignment with the given input prompts. To aid interpretation, any misaligned or missing elements in the generated images are emphasized using bold text. In particular, LLaVA-Reward-Phi [60] and UnifiedReward [46] generate outputs where the black bag is not properly held in hand, and in some cases, depict two bags in both paws while omitting the wooden cane altogether. These qualitative observations underscore the effec-

tiveness of RubricRL in enhancing the model's capability to follow complex, fine-grained instructions and produce high-quality, prompt-consistent images.

5 Conclusion

In this paper, we introduce RubricRL, a rubric-based reward RL framework that provides prompt-adaptive, decomposable supervision for text-to-image generation. By explicitly creating configurable visual criteria (*e.g.*, counting, attributes, OCR fidelity, realism) and scoring them independently, RubricRL produces interpretable and modular signals that integrate seamlessly with standard policy optimization in RL. Experimental results demonstrate that RubricRL outperforms existing RL-based approaches for enhancing text-to-image generation. We hope this work provides new insights into applying reinforcement learning to visual generation models.

6 Acknowledgment

Xuelu Feng and Chunming Qiao are supported in part by NSF CNS-2413876 and CNS-2120369.

References

1. Abdin, M., Aneja, J., Awadalla, H., Awadallah, A., Awan, A.A., Bach, N., Bahree, A., Bakhtiari, A., Bao, J., Behl, H., et al.: Phi-3 technical report: A highly capable language model locally on your phone. arXiv e-prints pp. arXiv–2404 (2024)
2. Agarwal, N., Ali, A., Bala, M., Balaji, Y., Barker, E., Cai, T., Chattopadhyay, P., Chen, Y., Cui, Y., Ding, Y., et al.: Cosmos world foundation model platform for physical ai. arXiv preprint arXiv:2501.03575 (2025)
3. Avrahami, O., Lischinski, D., Fried, O.: Blended diffusion for text-driven editing of natural images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18208–18218 (2022)
4. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
5. Black Forest Labs: Flux.1 [dev]. `https://huggingface.co/black-forest-labs/FLUX.1-dev` (2024), accessed: 2026-02-24
6. Chang, H., Zhang, H., Barber, J., Maschinot, A., Lezama, J., Jiang, L., Yang, M.H., Murphy, K., Freeman, W.T., Rubinstein, M., et al.: Muse: Text-to-image generation via masked generative transformers. arXiv preprint arXiv:2301.00704 (2023)
7. Chen, X., Mishra, N., Rohaninejad, M., Abbeel, P.: Pixelsnail: An improved autoregressive generative model. In: International conference on machine learning. pp. 864–872. PMLR (2018)
8. Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C.: Janus-pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811 (2025)

9. Choi, J., Kim, S., Jeong, Y., Gwon, Y., Yoon, S.: Ilvr: Conditioning method for denoising diffusion probabilistic models. arXiv preprint arXiv:2108.02938 (2021)
10. Cui, C., Sun, T., Lin, M., Gao, T., Zhang, Y., Liu, J., Wang, X., Zhang, Z., Zhou, C., Liu, H., et al.: Paddleocr 3.0 technical report. arXiv preprint arXiv:2507.05595 (2025)
11. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., et al.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025)
12. Dong, R., Han, C., Peng, Y., Qi, Z., Ge, Z., Yang, J., Zhao, L., Sun, J., Zhou, H., Wei, H., et al.: Dreamllm: Synergistic multimodal comprehension and creation. arXiv preprint arXiv:2309.11499 (2023)
13. Egan, B., Redden, A., XWAVE, SilentAntagonist: Dalle3 1 Million+ High Quality Captions (May 2024), `https://huggingface.co/datasets/ProGamerGov/synthetic-dataset-1m-dalle3-high-quality-captions`
14. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)
15. Geng, Z., Wang, Y., Ma, Y., Li, C., Rao, Y., Gu, S., Zhong, Z., Lu, Q., Hu, H., Zhang, X., et al.: X-omni: Reinforcement learning makes discrete autoregressive image generative models great again. arXiv preprint arXiv:2507.22058 (2025)
16. Gong, Y., Wang, X., Wu, J., Wang, S., Wang, Y., Wu, X.: Onereward: Unified mask-guided image generation via multi-task human preference learning. arXiv preprint arXiv:2508.21066 (2025)
17. Gonzalez, T.F.: Clustering to minimize the maximum intercluster distance. Theoretical computer science **38**, 293–306 (1985)
18. Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., Yuan, L., Guo, B.: Vector quantized diffusion model for text-to-image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10696–10706 (2022)
19. Gunjal, A., Wang, A., Lau, E., Nath, V., He, Y., Liu, B., Hendryx, S.: Rubrics as rewards: Reinforcement learning beyond verifiable domains. arXiv preprint arXiv:2507.17746 (2025)
20. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
21. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
22. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
23. Hu, X., Wang, R., Fang, Y., Fu, B., Cheng, P., Yu, G.: Ella: Equip diffusion models with llm for enhanced semantic alignment. arXiv preprint arXiv:2403.05135 (2024)
24. Huang, Z., Zhuang, Y., Lu, G., Qin, Z., Xu, H., Zhao, T., Peng, R., Hu, J., Shen, Z., Hu, X., et al.: Reinforcement learning with rubric anchors. arXiv preprint arXiv:2508.12790 (2025)
25. Karypis, G., Han, E.H., Kumar, V.: Chameleon: Hierarchical clustering using dynamic modeling. computer **32**(8), 68–75 (1999)
26. Labs, B.F.: Flux. `https://github.com/black-forest-labs/flux` (2024)
27. Li, S., Kallidromitis, K., Gokul, A., Liao, Z., Kato, Y., Kozuka, K., Grover, A.: Omniflow: Any-to-any generation with multi-modal rectified flows. In: Proceedings

of the Computer Vision and Pattern Recognition Conference. pp. 13178–13188 (2025)
28. Liu, J., Liu, G., Liang, J., Li, Y., Liu, J., Wang, X., Wan, P., Zhang, D., Ouyang, W.: Flow-grpo: Training flow matching models via online rl. arXiv preprint arXiv:2505.05470 (2025)
29. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
30. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. Advances in neural information processing systems **35**, 27730–27744 (2022)
31. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
32. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International conference on machine learning. pp. 8821–8831. Pmlr (2021)
33. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
34. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
35. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
36. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. pmlr (2015)
37. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. Advances in neural information processing systems **32** (2019)
38. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
39. Sun, K., Pan, J., Ge, Y., Li, H., Duan, H., Wu, X., Zhang, R., Zhou, A., Qin, Z., Wang, Y., et al.: Journeydb: A benchmark for generative image understanding. Advances in neural information processing systems **36**, 49659–49678 (2023)
40. Sun, P., Jiang, Y., Chen, S., Zhang, S., Peng, B., Luo, P., Yuan, Z.: Autoregressive model beats diffusion: Llama for scalable image generation. arXiv preprint arXiv:2406.06525 (2024)
41. Team, Q., et al.: Qwen2.5 technical report. arXiv preprint arXiv:2407.10671 **2**(3) (2024)
42. Tian, K., Jiang, Y., Yuan, Z., Peng, B., Wang, L.: Visual autoregressive modeling: Scalable image generation via next-scale prediction. Advances in neural information processing systems **37**, 84839–84865 (2024)
43. Wang, J., Tian, Z., Wang, X., Zhang, X., Huang, W., Wu, Z., Jiang, Y.G.: Simplear: Pushing the frontier of autoregressive visual generation through pretraining, sft, and rl. arXiv preprint arXiv:2504.11455 (2025)

44. Wang, X., Zhang, X., Luo, Z., Sun, Q., Cui, Y., Wang, J., Zhang, F., Wang, Y., Li, Z., Yu, Q., et al.: Emu3: Next-token prediction is all you need. arXiv preprint arXiv:2409.18869 (2024)
45. Wang, Y., Li, Z., Zang, Y., Zhou, Y., Bu, J., Wang, C., Lu, Q., Jin, C., Wang, J.: Pref-grpo: Pairwise preference reward-based grpo for stable text-to-image reinforcement learning. arXiv preprint arXiv:2508.20751 (2025)
46. Wang, Y., Zang, Y., Li, H., Jin, C., Wang, J.: Unified reward model for multimodal understanding and generation. arXiv preprint arXiv:2503.05236 (2025)
47. Wei, H., Liu, C., Chen, J., Wang, J., Kong, L., Xu, Y., Ge, Z., Zhao, L., Sun, J., Peng, Y., et al.: General ocr theory: Towards ocr-2.0 via a unified end-to-end model. arXiv preprint arXiv:2409.01704 (2024)
48. Wei, T., Chen, D., Zhou, Y., Pan, X.: Enhancing mmdit-based text-to-image models for similar subject generation. arXiv preprint arXiv:2411.18301 (2024)
49. von Werra, L., Belkada, Y., Tunstall, L., Beeching, E., Thrush, T., Lambert, N., Huang, S., Rasul, K., Gallouédec, Q.: Trl: Transformer reinforcement learning. `https://github.com/huggingface/trl` (2020)
50. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
51. Wu, C., Chen, X., Wu, Z., Ma, Y., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C., et al.: Janus: Decoupling visual encoding for unified multimodal understanding and generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12966–12977 (2025)
52. Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., Li, H.: Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. arXiv preprint arXiv:2306.09341 (2023)
53. Xue, Z., Wu, J., Gao, Y., Kong, F., Zhu, L., Chen, M., Liu, Z., Liu, W., Guo, Q., Huang, W., et al.: Dancegrpo: Unleashing grpo on visual generation. arXiv preprint arXiv:2505.07818 (2025)
54. Yang, S., Wu, T., Shi, S., Lao, S., Gong, Y., Cao, M., Wang, J., Yang, Y.: Maniqa: Multi-dimension attention network for no-reference image quality assessment. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1191–1200 (2022)
55. Yu, J., Xu, Y., Koh, J.Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B.K., et al.: Scaling autoregressive models for content-rich text-to-image generation. arXiv preprint arXiv:2206.10789 **2**(3), 5 (2022)
56. Yu, Q., Zhang, Z., Zhu, R., Yuan, Y., Zuo, X., Yue, Y., Dai, W., Fan, T., Liu, G., Liu, L., et al.: Dapo: An open-source llm reinforcement learning system at scale. arXiv preprint arXiv:2503.14476 (2025)
57. Yuan, S., Liu, Y., Yue, Y., Zhang, J., Zuo, W., Wang, Q., Zhang, F., Zhou, G.: Ar-grpo: Training autoregressive image generation models via reinforcement learning. arXiv preprint arXiv:2508.06924 (2025)
58. Zheng, K., Chen, H., Ye, H., Wang, H., Zhang, Q., Jiang, K., Su, H., Ermon, S., Zhu, J., Liu, M.Y.: Diffusionnft: Online diffusion reinforcement with forward process. arXiv preprint arXiv:2509.16117 (2025)
59. Zhou, C., Yu, L., Babu, A., Tirumala, K., Yasunaga, M., Shamis, L., Kahn, J., Ma, X., Zettlemoyer, L., Levy, O.: Transfusion: Predict the next token and diffuse images with one multi-modal model. arXiv preprint arXiv:2408.11039 (2024)
60. Zhou, S., Zhang, R., Zhu, H., Kveton, B., Zhou, Y., Gu, J., Chen, J., Chen, C.: Multimodal llms as customized reward models for text-to-image generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19638–19648 (2025)