Scaling Multi-Reference Image Generation with Dynamic Reward Optimization

Wenwang Huang[1*], Yusen Fu[1*], Junjie Wang[1], Mengfei Huang[1], Yulin Li[1], Gan Liu[2], Jing Cai[2], Yancheng He[2], and Zhuotao Tian[1,3†]

[1] Harbin Institute of Technology, Shenzhen, China
[2] Independent Researcher, China
[3] Shenzhen Loop Area Institute, China

* Equal contribution † Corresponding author: tianzhuotao@hit.edu.cn

Abstract. While personalized image generation has achieved remarkable progress, multi-reference image generation (MRIG) remains a challenging task. Most existing benchmarks fail to adequately evaluate complex MRIG scenarios, hindering further progress in this area. To better assess model performance on complex MRIG tasks, we introduce OmniRef-Bench, a benchmark that covers complex combinations of reference image types and a large number of reference images. Evaluations on OmniRef-Bench show that mainstream open-source models struggle in complex MRIG scenarios, and their performance deteriorates significantly as the number of mixed-type reference images increases. To address this issue, we propose **DyRef**, a two-stage training framework. In the first stage, supervised fine-tuning equips the model with the basic capability to handle complex MRIG tasks. In the second stage, we introduce Difficulty-aware Advantage Reweighting (DAR) and Discriminative Reward Scaling (DRS). DAR dynamically adjusts the optimization objective to improve performance when handling a large number of mixed-type reference images. DRS enlarges intra-group reward differences for more effective policy optimization. Experiments demonstrate that DyRef significantly improves the performance of open-source models on OmniRef-Bench and single-image editing benchmarks, demonstrating the effectiveness and generalization capability of our approach. Our code is available at https://github.com/Weistrass/DyRef.

Keywords: Personalized Image Generation · Reinforcement Learning

1 Introduction

The emergence of diffusion models [19, 24] has substantially advanced personalized image generation [9, 15, 16, 31, 37], enabling high-fidelity synthesis closely aligned with user instructions. However, among various personalization paradigms, Multi-Reference Image Generation (MRIG) [18,29,36,41,42] remains challenging due to the difficulty of effectively integrating multiple reference images of diverse

Fig. 1: Versatile samples of the proposed DyRef. Given complex samples that involve mixed reference image types and varying numbers of reference images, DyRef consistently generates high-quality results in accordance with user instructions.

types (e.g., subject, pose, and style). Addressing these challenges carries significant implications for fields like professional visual design [35] and advertising applications [3], where coherent integration of varied visual elements is crucial.

Insufficient Evaluation for Complex MRIG. Despite its potential, MRIG research remains constrained by insufficient benchmarks. Existing benchmarks [1, 2, 18, 43] typically involve simple tasks with limited numbers and types of reference images, neglecting complex scenarios that require integrating numerous and diverse references. Additionally, these benchmarks lack fine-grained and multi-dimensional evaluation metrics to accurately assess model performance across various reference types. These limitations hinder meaningful comparisons between methods and impede progress towards practical MRIG applications.

The Proposed OmniRef-Bench. To better assess MRIG performance, we introduce OmniRef-Bench, a personalized image generation benchmark comprising intricate combinations of diverse reference types as well as a large number of reference images across synthetic and real-world scenarios. Furthermore, we

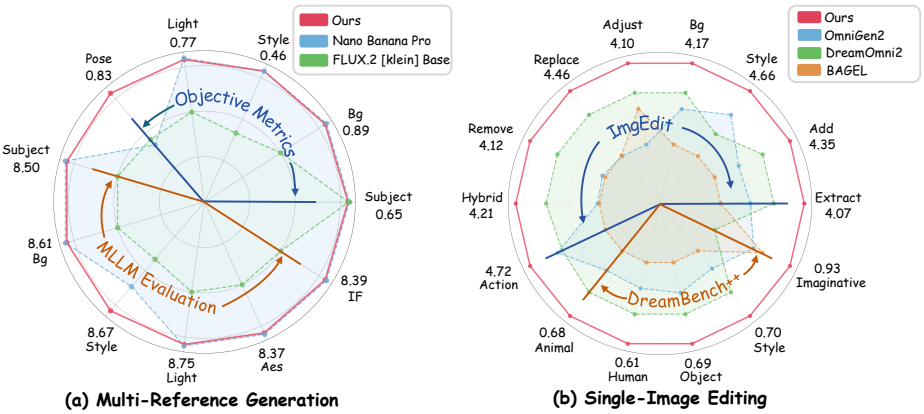

Fig. 2: Performance of DyRef. On both MRIG and single-image editing tasks, DyRef (Ours) consistently outperforms open-source state-of-the-art methods and achieves performance comparable to leading closed-source model Nano Banana Pro.

propose a hybrid evaluation strategy combining objective quantitative metrics with Multimodal Large Language Models (MLLMs)-based assessments, and design fine-grained evaluation criteria tailored to each reference type, facilitating a robust and multidimensional evaluation of complex MRIG tasks.

Key Observations. Using OmniRef-Bench (Tab. 2), we find that mainstream open-source MRIG models [6, 37, 38, 42] struggle in complex MRIG scenarios, exhibiting a huge gap compared with state-of-the-art closed-source models [5,26]. To bridge this disparity, we conduct further experiments and observe that while open-source models generate high-quality images with a few references, their performance deteriorates rapidly as the number of reference images involving mixed reference types increases, as illustrated in Fig. 3. This observation raises a critical research question: *Can mitigating this degradation effectively narrow the gap between open-source and closed-source models?*

Our Solution. To address this issue, we propose **DyRef**, a two-stage training framework. In the first stage, we use Supervised Fine-Tuning (SFT) to equip the model with preliminary capabilities to handle complex MRIG tasks. In the second stage, we introduce Difficulty-aware Advantage Reweighting (DAR) and Discriminative Reward Scaling (DRS). Specifically, DAR dynamically adjusts the optimization objective to encourage the model to focus more on underperforming samples with a larger number of mixed-type reference images, while preventing overfitting to well-performing samples with fewer references. DRS enlarges reward differences within each group for more effective policy optimization. The combination of DAR and DRS enhances the model's ability to learn from samples with a large number of diverse reference images, thereby improving performance on complex MRIG tasks.

Experimental results (Fig. 2 and Fig. 3(b)) demonstrate that DyRef substantially improves the performance of open-source models on OmniRef-Bench,

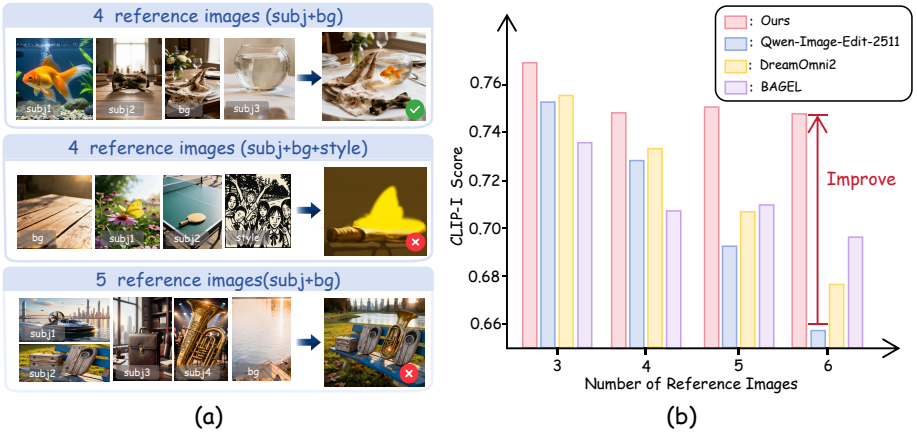

Fig. 3: Motivation of our DyRef. **(a)** Qualitative results: open source models yield high-quality results with a limited set of references, but suffer from significant artifacts and semantic loss as the number and complexity of reference images increase. **(b)** Quantitative results: on the OmniRef-Bench, CLIP-I scores between generated images and target images confirm that the performance of mainstream open-source models significantly drops as the number of mixed-type reference images increases.

achieving results comparable to state-of-the-art closed-source models. Moreover, DyRef further enhances the model performance on single-image editing benchmarks, highlighting its effectiveness and generalization ability.

In summary, the contributions of this work are as follows:

- To better assess model performance on complex MRIG tasks, we introduce OmniRef-Bench. Using this benchmark, we observe that open-source models struggle in complex multi-reference settings, and their performance declines rapidly as the number of mixed-type reference images increases.
- To address the issue, we propose DyRef. Building upon SFT, we employ DAR to encourage the model to focus more on underperforming samples with a larger number of mixed-type reference images, and utilize DRS to enhance reward differences across samples for better policy optimization.
- Experimental results show that DyRef not only substantially boosts performance on OmniRef-Bench, but also improves results on single-image editing benchmarks, demonstrating both its effectiveness and generalization.

2 Dataset and Benchmark

In this section, we first introduce the motivation of OmniRef-Bench in Sec. 2.1. Next, Sec. 2.2 describes the composition and construction pipeline of the benchmark. Finally, Sec. 2.3 outlines the corresponding evaluation strategy.

Table 1: Comparison of OmniRef-Bench with existing benchmarks. Eval Granularity indicates whether a benchmark conducts tailored assessment for every reference type (fine-grained), or simply conducts a general evaluation (coarse-grained). Max Combs denotes the maximum number of reference image type combinations. Obj refers to objective quantitative metrics. MLLM indicates evaluation conducted using the MLLM.

Benchmarks	Max Combs.	Expert-Annotated	Eval Granularity	Eval Perspective
XVerse [1]	1	✗	fine-grained	Obj
PSRBench [34]	1	✗	fine-grained	Obj + MLLM
MultiBanana [18]	3	✗	coarse-grained	MLLM
DreamOmni2 [42]	2	✓	coarse-grained	MLLM
OmniRef-Bench (Ours)	4	✓	fine-grained	Obj + MLLM

2.1 Motivation

As shown in Tab. 1, existing benchmarks exhibit limitations in evaluating complex MRIG tasks. On the one hand, most existing benchmarks [1, 20, 34] primarily focus on single-subject or multi-subject generation. Although some consider a large number of reference images, they lack combinations of multiple reference types, such as style, lighting, and pose. On the other hand, recent studies [2, 42, 43] have attempted to incorporate reference types beyond subject identity, but they typically only consider simple pairwise combinations with a small number of reference images, which fail to satisfy the complex semantic compositions required for artistic creation and advanced content generation.

Moreover, existing benchmarks [29, 44] typically rely exclusively on either objective quantitative metrics or MLLM-based scoring mechanisms. However, relying on only one of these approaches introduces inherent limitations. Detailed limitations are provided in Appendix B.5. Furthermore, some benchmarks [18,41] do not provide fine-grained evaluation criteria for each reference type, making it difficult to accurately measure model performance under combinations of diverse reference types. To address these limitations, we introduce OmniRef-Bench.

2.2 OmniRef-Bench

Benchmark Overview. OmniRef-Bench comprises a total of 395 expert-annotated benchmark instances. These instances cover five major reference types: subject, background, style, lighting, and pose. Each instance contains a target editing task together with multiple reference images, with the number of reference images ranging from 2 to 7. This design enables comprehensive evaluation from simple two-reference editing to complex MRIG. In addition, OmniRef-Bench is curated to include flexible and complex combinations of reference types, totaling 10 distinct combinations, with up to 4 reference types present simultaneously. More statistics about OmniRef-Bench can be found in Appendix B.1.

Benchmark Construction. OmniRef-Bench consists of both real-world and synthetic images. The real-world images are sourced from the LAION-5B dataset [25].

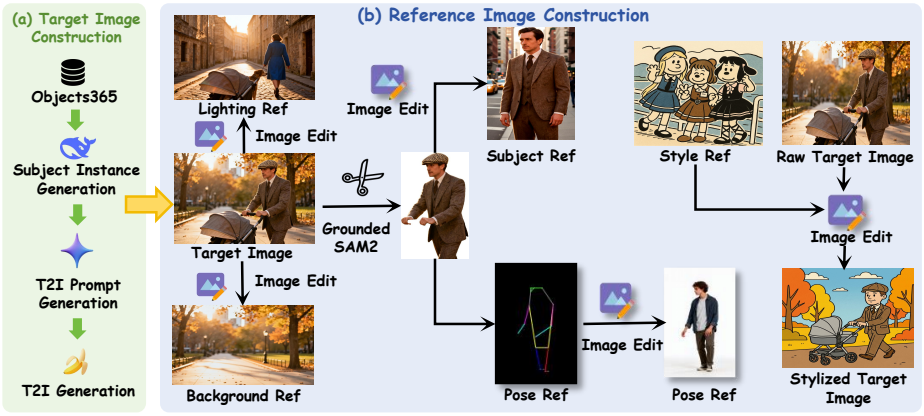

Fig. 4: Data construction pipeline. Our data construction pipeline is structured into two phases. **(a)** We obtain the target image using T2I generation. **(b)** We acquire the corresponding reference images of each type (except style) based on the target image, and we obtain the stylized target image by updating the original target image using an external style reference image from the OmniConsistency dataset [30].

For synthetic image generation, we design an efficient multi-reference data generation pipeline to obtain high-quality data for both training and benchmarking. As illustrated in Fig. 4, inspired by UNO [40], we obtain raw subject concepts from the Objects365 dataset [27], and then utilize LLM [12, 32] to generate subject instances and text-to-image (T2I) prompts sequentially. The generated prompts are then fed into the T2I generation models [5, 26] to obtain the target image. Based on the generated target image, we then utilize a segmentation model [14, 23] and a series of image editing models [37] to produce each type of reference image. For more details, please refer to Appendix B.2.

2.3 Evaluation Protocol

As shown in Tab. 1, OmniRef-Bench adopts both objective metrics and MLLM-based evaluation. The details are as follows.

Objective Metrics. We assess the consistency between the generated image and the reference images for each of the five reference types. For each type, we either leverage established metrics [4, 8, 17, 22, 28], or propose carefully-designed custom metrics for dimensions where prior evaluation methods are lacking. Please refer to Appendix B.3 for more details.

MLLM-based Evaluation. We use Gemini 3 Flash [32] to conduct the evaluation. Compared with objective metrics, we remove the pose consistency evaluation based on our observation in Appendix B.4 that MLLM cannot accurately measure the fine-grained angles of human limbs. Meanwhile, we introduce aesthetic and instruction following dimensions to fully leverage the high-level semantic understanding ability of MLLM. Refer to Appendix B.4 for more details.

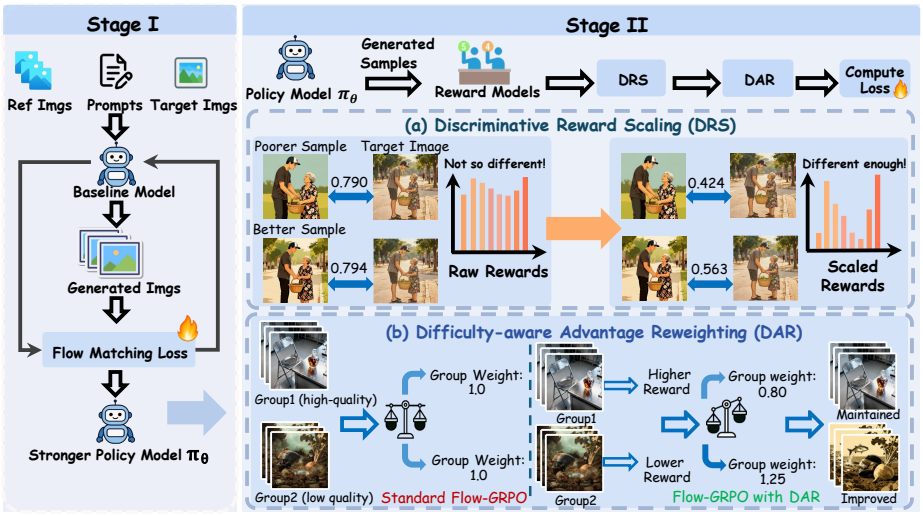

Fig. 5: Framework of DyRef. In Stage I, SFT equips the model with the basic capability to handle complex MRIG tasks. In Stage II, **(a)** DRS enlarges the reward differences across samples for better training. **(b)** DAR enhances the model's focus on samples with a large number of mixed-type reference images.

3 Methods

In Sec. 2, we introduce OmniRef-Bench to assess the complex multi-reference image generation (MRIG) task. Evaluation results (Tab. 2) indicate that mainstream open-source models perform poorly on this task. Further analysis (Fig. 3) shows that their performance declines rapidly as the number of reference images involving mixed reference types increases. To address this issue, we propose DyRef, a two-stage training framework. As shown in Fig. 5, in Stage I, we employ SFT to equip the model with the basic capability to handle complex MRIG tasks. In Stage II, **(a)** DRS enlarges the reward differences across samples for better training, and **(b)** DAR enhances the model's focus on samples with numerous mixed-type reference images. The combination of DAR and DRS enhances the model's ability to learn from samples with a large number of diverse reference images, thereby improving performance on complex MRIG tasks.

3.1 Stage I: Supervised Fine-Tuning

In the first stage, we adopt Qwen-Image-Edit-2511 [37] as our training baseline, which employs a standard dual-stream MMDiT backbone [9]. This dual-stream mechanism enables the editing module to achieve a robust balance between maintaining high-level semantic consistency and preserving low-level visual fidelity across multiple inputs. To equip the model with the basic capability to handle complex MRIG tasks, we construct approximately 14000 training samples using

the data construction pipeline outlined in Fig. 4. We then fine-tune the model using LoRA [7] with Flow Matching loss [11, 13, 21].

Formally, let $z_0 \sim p_0(z)$ denote samples drawn from a standard Gaussian distribution, and let $z_1 \sim p_{\text{data}}(z)$ denote samples drawn from the target data distribution. z_t represents the intermediate state along the transport trajectory that maps z_0 to z_1. Moreover, $v_\theta(z, t)$ denotes a time-dependent velocity field parameterized by a neural network, and θ denotes the learnable parameters of the model. The Flow Matching optimization objective is defined as:

$$\min_v \int_0^1 \mathbb{E}\Big[\|(z_1 - z_0) - v_\theta(z_t, t)\|^2\Big] \, \mathrm{d}t. \tag{1}$$

3.2 Stage II: Difficulty-aware Advantage Reweighting

As shown in Tab. 2 and Tab. 8, although the model fine-tuned with SFT has acquired preliminary capabilities, a noticeable gap remains compared with the SOTA closed-source model [5] on OmniRef-Bench, particularly in style-related tasks. Furthermore, we observe that the model performs poorly when processing complex samples containing numerous mixed-type reference images (Fig. 3). To address these limitations, we introduce a style-specific reward for targeted optimization of style-related tasks. In addition, we propose **D**ifficulty-aware **A**dvantage **R**eweighting (DAR) to enhance the model's focus on underperforming samples with a large number of diverse reference images. Notably, during the RL stage, we use data drawn from the same source as that used in the SFT stage, without introducing new datasets.

Rewards. Inspired by the USO [39] framework, we adopt CSD [28] as the reward model for style alignment to guide the model toward generating images that better match the style references. In addition, we employ CLIP [22] or SigLIPv2 [33] to measure the semantic consistency of training samples between generated images and their corresponding target images, thereby further reducing the distribution gap between generated outputs and the target data.

Motivation. For MRIG tasks, the amount of visual information increases as the number of mixed-type reference images grows, making these samples more complex and requiring greater focus during training. However, standard Flow-GRPO [12, 21] lacks the ability to dynamically optimize for samples of varying complexity. As a result, training tends to be dominated by simpler samplers with fewer reference images, while more complex samples with numerous mixed-type reference images receive insufficient learning signals during training.

In dense object detection, prior studies [10] have addressed sample imbalance by designing loss functions that adjust each sample's contribution to the overall loss, promoting targeted optimization. Furthermore, we observe that the reward scores of different samples also decrease as the number of mixed-type reference images increases. A detailed analysis is provided in Appendix C.3. Inspired by these observations, we propose DAR, an improved method based on Flow-GRPO. DAR dynamically adjusts the contribution of each sample group

of Flow-GRPO to the total loss based on the average reward score of samples within the group, thereby encouraging the model to prioritize underperforming samples with more mixed-type reference images.

Specifically, consider a single optimization step with a training set of samples denoted by $\mathcal{S}$. We partition $\mathcal{S}$ into groups $\mathcal{G}$, where each group $g \in \mathcal{G}$ corresponds to multiple samples generated from the same prompt. Let $I_g \subset \{1, \ldots, |\mathcal{S}|\}$ denote the index set of samples belonging to group g. For each sample $i \in \mathcal{S}$, we obtain a scalar reward score $p_i \in [0, 1]$. We first compute the mean reward of each group g as follows:

$$p_g = \frac{1}{|I_g|} \sum_{i \in I_g} p_i. \tag{2}$$

To emphasize groups with poor performance (i.e., those with lower average rewards), we construct a raw group weight based on the mean reward as follows:

$$w_g^{\text{raw}} = (1 - p_g + \epsilon)^{\gamma}, \tag{3}$$

where $\epsilon > 0$ is a small constant for numerical stability and $\gamma > 0$ is a hyperparameter that controls the degree of reweighting. Subsequently, the raw weight is clipped to the predefined range $[w_{\min}, w_{\max}]$ as follows:

$$w_g = \text{clip}\left(w_g^{\text{raw}},\, w_{\min},\, w_{\max}\right). \tag{4}$$

To avoid unintended shifts in overall loss scale, we normalize group weights to a batch-wise mean of approximately 1. Let μ denote the mean group weight within the current batch, and the final normalized group weight is defined as:

$$\mu = \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} w_g, \quad \tilde{w}_g = \text{clip}\left(\frac{w_g}{\mu},\, w_{\min},\, w_{\max}\right). \tag{5}$$

Then, each sample i inherits the weight of its corresponding group $g(i)$:

$$w_i = \tilde{w}_{g(i)}. \tag{6}$$

Finally, the computed w_i weights each sample's contribution to the final objective. Following Flow-GRPO [13], $r_i(\theta)$ denotes the probability ratio between the current policy and the behavior policy for sample i, and $\hat{A}_i$ represents the estimated advantage of each sample. The weighted objective is defined as:

$$L(\theta) = -\frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} w_i \min\left(r_i(\theta)\hat{A}_i, \text{clip}\left(r_i(\theta), 1 - \epsilon, 1 + \epsilon\right)\hat{A}_i\right). \tag{7}$$

Discriminative Reward Scaling. Although DAR is effective, we still face another challenge. As illustrated in Fig. 5(a), we observe that the reward scores computed using CLIP or SigLIPv2 tend to be overly concentrated, and even when generated samples exhibit substantial differences in visual quality, their assigned reward scores remain highly similar. Consequently, when calculating

Table 2: Results of various methods on the OmniRef-Bench. **Bold** and underlined indicate the best and second-best results, respectively. Gray background indicates implementations with our method. Bg, Aes, IF, and Avg are the abbreviations for background, aesthetic, instruction following, and average, respectively. Qwen-2511 and FLUX.2 [klein] denotes the Qwen-Image-Edit-2511 model and the FLUX.2 [klein] 9B Base model, respectively. All MLLM evaluations are conducted using Gemini 3 Flash.

Methods	Objective Metrics						MLLM Evaluation						
	Subject	Style	Bg	Light	Pose	Avg	Subject	Bg	Style	Light	Aes	IF	Avg
Closed-Source Models													
Seedream4.5 [26]	0.65	0.34	0.89	0.74	0.74	0.67	8.77	8.92	6.27	8.78	8.08	7.97	8.13
Nano Banana Pro [5]	0.64	0.46	0.89	0.77	0.72	0.70	8.50	8.61	8.37	8.75	8.37	8.39	8.50
Open-Source Models													
OmniGen2 [38]	0.59	0.11	0.85	0.65	0.65	0.57	4.47	2.31	0.37	4.68	6.73	4.16	3.79
DreamOmni2 [42]	0.57	0.28	0.85	0.71	0.67	0.62	6.75	3.43	3.45	6.07	6.65	4.71	5.18
BAGEL [6]	0.58	0.15	0.85	0.67	0.69	0.59	6.21	3.66	1.13	6.31	5.98	4.54	4.64
Qwen-2511 [37]	0.55	0.15	0.84	0.71	0.71	0.59	6.50	4.35	1.61	6.81	6.09	4.46	4.97
+ Ours	0.62	0.46	0.87	0.76	**0.83**	**0.71**	**8.34**	**8.51**	**8.67**	**8.54**	**8.15**	**8.08**	**8.38**
FLUX.2 [klein] [9]	**0.63**	0.21	0.87	0.69	0.74	0.63	8.25	6.32	2.18	7.69	7.95	6.53	6.49
+ Ours	**0.63**	**0.48**	**0.90**	**0.78**	0.76	**0.71**	8.07	8.12	8.06	8.35	8.11	7.48	8.03

intra-group advantages, insufficient numerical contrast between superior and inferior samples weakens the gradient signals for policy updates. This phenomenon may stem from intrinsic biases in the pre-trained evaluation models. Detailed quantitative analysis is provided in Appendix C.3. To address this issue, we propose **D**iscriminative **R**eward **S**caling (DRS), which aims to enhance the discriminability of reward values through a transformation function. By enlarging reward differences between high-quality and low-quality samples, DRS preserves robust gradient signals for policy updates, improving overall model performance.

Specifically, at each optimization step of DAR, for every generated sample, let F_{gen} denote the CLIP or SigLIPv2 feature of the generated image and F_{target} denote the CLIP or SigLIPv2 feature of the target image, respectively. We first compute the cosine similarity $p = \cos(F_{\text{gen}}, F_{\text{target}})$ as the raw reward. Subsequently, we amplify the reward differences among samples within the same group using a function F:

$$\hat{p} = F(p, t), \tag{8}$$

where $0 < t < 1$ is a hyperparameter that controls the distribution of rewards and $\hat{p}$ is a reward transformed by DRS. A detailed analysis of the implementation of the function F is provided in Appendix C.4.

4 Experiments

In this section, we present the main results of our proposed method in Sec. 4.1 and analyze our design choices via ablation studies in Sec. 4.2. Details about our

Table 3: Performance evaluation on the ImgEdit Benchmark. All results are evaluated by GPT-4o following the default settings. Results on closed-source models are sourced from UniRef-Image-Edit [36].

Methods	Extract	Add	Style	Bg	Adjust	Replace	Remove	Hybrid	Action	Overall
Closed-Source Models										
Seedream4.0 [26]	2.39	4.52	4.76	4.30	4.41	4.56	4.44	3.33	4.36	4.18
GPT Image 1 [High]	2.90	4.61	4.93	4.57	4.33	4.35	3.66	3.96	4.89	4.20
Open-Source Models										
OmniGen2 [38]	2.43	3.72	4.63	3.72	3.11	3.75	2.62	2.60	4.52	3.46
DreamOmni2 [42]	2.95	3.83	4.43	3.90	3.53	4.27	3.09	2.91	4.50	3.71
BAGEL [6]	1.77	3.61	4.20	3.33	3.40	3.76	2.59	2.59	4.32	3.29
Qwen-2511 [37]	**4.23**	4.40	4.61	**4.23**	3.91	4.21	3.76	2.72	4.52	4.07
+ Ours	4.07	4.35	4.66	4.17	**4.10**	4.46	4.12	**4.21**	4.72	**4.32**
FLUX.2 [klein] [9]	2.89	4.36	**4.72**	3.99	4.02	4.47	3.95	3.50	4.72	4.07
+ Ours	2.81	**4.43**	4.69	4.08	3.84	**4.48**	**4.27**	3.65	**4.77**	4.11

Table 4: Performance on the DreamBench++. Photo., Style., Imag. are the abbreviations for Photorealistic, Style Transfer, and Imaginative, respectively. All results are evaluated by GPT-4o.

Methods	**Concept Preservation (CP)**					**Prompt Following (PF)**				**CP*PF**
	Animal	Human	Object	Style	Overall	Photo.	Style.	Imag.	Overall	
OmniGen2 [38]	0.54	0.41	0.55	0.49	0.52	0.95	0.93	0.87	0.92	0.48
DreamOmni2 [42]	0.64	0.59	0.67	0.53	0.63	0.81	0.80	0.7	0.78	0.49
BAGEL [6]	0.45	0.23	0.38	0.46	0.39	**0.97**	**0.97**	0.91	0.95	0.37
Qwen-2511	0.58	0.55	0.59	0.54	0.58	0.90	0.94	0.93	0.92	0.53
+ Ours	**0.68**	**0.61**	**0.69**	**0.70**	**0.68**	0.90	0.94	0.93	0.92	**0.62**
FLUX.2 [klein] [9]	0.59	0.53	0.61	0.54	0.58	**0.97**	**0.97**	**0.94**	**0.97**	0.56
+ Ours	0.65	0.58	0.66	0.61	0.64	0.94	**0.97**	0.93	0.95	0.61

training and experiment settings can be found in Appendix D and E, respectively. Fig. 6 presents qualitative comparisons between DyRef and existing methods on several representative examples. More qualitative analysis and failure case study can be found in Appendix F.

4.1 Main Results

First, we conduct extensive experiments on our proposed OmniRef-Bench and two other single-image editing benchmarks [20,44]. To verify the generalizability of our method on models of different parameter scales, we select Qwen-Image-Edit-2511 (Qwen-2511, 20B) and FLUX.2 [klein] 9B Base (FLUX.2 [klein]) as our base models and implement our method on them. We compare the results of our method with SOTA closed-source models [5, 26] and mainstream open-source models [6,38,42]. Subsequently, we analyze the alignment of our evaluation

Table 5: Performance of various methods on OmniContext. "Char.+Obj." indicates Character+Object. The best and second best results are shown in **bold** and underline, respectively. Qwen-2511 denotes the Qwen-Image-Edit-2511.

Methods	SINGLE		MULTIPLE			SCENE			Average
	Character	Object	Character	Object	Char.+Obj.	Character	Object	Char.+Obj.	
BAGEL [6]	6.99	5.94	5.20	6.74	7.02	4.75	4.59	5.55	5.85
OmniGen2 [38]	8.16	8.18	7.43	7.29	7.88	7.34	6.62	7.37	7.54
DreamOmni2 [42]	7.56	7.49	6.53	6.53	6.69	6.05	5.86	6.09	6.60
Qwen-2511 [37]	8.56	8.44	**8.38**	**8.88**	7.91	6.84	**8.04**	**8.10**	8.14
+Ours	**8.59**	**8.80**	8.24	8.80	**8.09**	**7.59**	7.93	7.89	**8.24**

Table 6: Performance of various methods on MultiBanana (10-point scale; the higher the better). Following the official evaluation protocol provided by the benchmark authors, we use Gemini-2.5 and GPT-5 as evaluators.

Methods	Single	Two	X Objects	X-1 + Local	X-1 + Global	X-1 + Background	Average
DreamOmni2	**6.52**	4.07	2.80	3.04	2.87	2.59	3.65
OmniGen2	5.92	3.44	3.26	3.60	3.37	3.02	3.77
Qwen-2511	6.37	**4.54**	3.01	2.93	3.18	2.65	3.78
+Ours	6.34	4.42	**3.96**	**3.65**	**4.04**	**3.63**	**4.34**

results with human preference through qualitative analysis and a user study. More experimental results can be found in Appendix C.

Results on OmniRef-Bench. To verify the effectiveness of our method on the complex multi-reference generation task, we conduct experiments on the proposed OmniRef-Bench. As shown in Tab. 2, Qwen-2511 with our method significantly outperforms all open-source models by at least 12% (+0.08 compared with FLUX.2 [klein]) and 29% (+1.89 compared with FLUX.2 [klein]) in objective metrics and MLLM evaluation, respectively, with a particular enhancement in style and pose-related metrics. Meanwhile, it also surpasses the two closed-source models in objective metrics and remains comparable with Nano Banana Pro in MLLM evaluation. These results demonstrate that our method can handle diverse reference combinations with varying reference counts and achieve competitive performance with closed-source models.

Results on Single-Image Editing Benchmarks. To ensure our method enhances the base model's multi-reference generation ability without compromising its basic image editing capability, we conduct experiments on two single-image editing benchmarks, ImgEdit [44] and DreamBench++ [20]. The experimental results are reported in Tab. 3 and Tab. 4, respectively. Results on both benchmarks demonstrate that our method achieves superior performance over both the base models and other open-source models. Specifically, on ImgEdit, Qwen-2511 with our method not only significantly surpasses all open-source models with the highest score of 4.32, but also outperforms closed-source models such as Seedream4.0 [26] and GPT Image 1 [high]. On DreamBench++, Qwen-2511 with

Table 7: User Study. (1-10) and (1-4) denote the range of the scores. Nano denotes Nano Banana Pro. PLCC, SROCC, and KROCC denote Pearson, Spearman, and Kendall correlation coefficients, respectively.

Evaluation Setup	Evaluation Results				Correlation Analysis		
	Ours	Nano	Seedream4.5	Qwen-2511	PLCC	SROCC	KROCC
Gemini3 Flash (1-10)	8.38	**8.50**	8.13	4.97	0.902	0.800	0.667
Human (1-4)	**3.42**	3.09	2.26	1.22			

Table 8: Ablation study on the OmniRef-Bench.

Settings	Objective Metrics						MLLM Evaluation						
	Subject	Style	Bg	Light	Pose	Avg	Subject	Bg	Style	Light	Aes	IF	Avg
SFT	0.62	0.38	**0.88**	0.72	0.82	0.68	8.20	8.04	7.57	8.13	7.75	7.72	7.90
Ours	0.62	**0.46**	0.87	**0.76**	0.83	**0.71**	8.34	8.51	**8.67**	8.54	**8.14**	**8.08**	**8.38**
w/o CSD	**0.64**	0.36	**0.88**	0.72	0.83	0.69	8.52	**8.79**	6.33	8.42	8.05	7.91	8.00
w/o DAR	0.63	0.39	**0.88**	0.74	0.81	0.69	**8.54**	8.39	6.98	**8.58**	7.89	7.82	8.03
w/o DRS	0.63	0.45	**0.88**	0.75	**0.84**	**0.71**	8.53	8.41	8.12	8.55	8.03	8.01	8.28

our method achieves the highest overall score of 0.62. For the results of the objective metrics on DreamBench++, please refer to Appendix C.1. These results indicate that our method not only improves the base model's multi-reference generation ability, but also enhances its single-image editing ability.

Results on other MRIG Benchmark. To further verify that our method remains effective beyond the complex MRIG scenarios introduced in OmniRef-Bench, we evaluate it on two simpler MRIG benchmarks, OmniContext [38] and MultiBanana [18]. As shown in Tab. 5 and Tab. 6, compared with the baseline Qwen-Image-Edit-2511 (Qwen-2511), our method improves the average performance across all metrics by 1.2% and 14.8% on OmniContext and MultiBanana, respectively, demonstrating its effectiveness and strong generalization ability. In addition, we evaluate our method on reference types that are not included in the training data. Detailed results are provided in Appendix C.6. The results demonstrate that our method generalizes effectively to unseen reference types beyond the training distribution.

Generalizability of Our Method. As shown in Tab. 2, Tab. 3, and Tab. 4, integrating our method with FLUX.2 [klein] not only significantly outperforms the base model on the multi-reference generation task (8.03 vs. 6.49), but also enhances its single-image editing capabilities (e.g., improving from 4.07 to 4.11 on ImgEdit and 0.56 to 0.61 on DreamBench++). Furthermore, despite having a smaller parameter scale (9B) compared to Qwen-2511 (20B), our approach achieves competitive performance, underscoring its effectiveness and generalizability across models of varying scales.

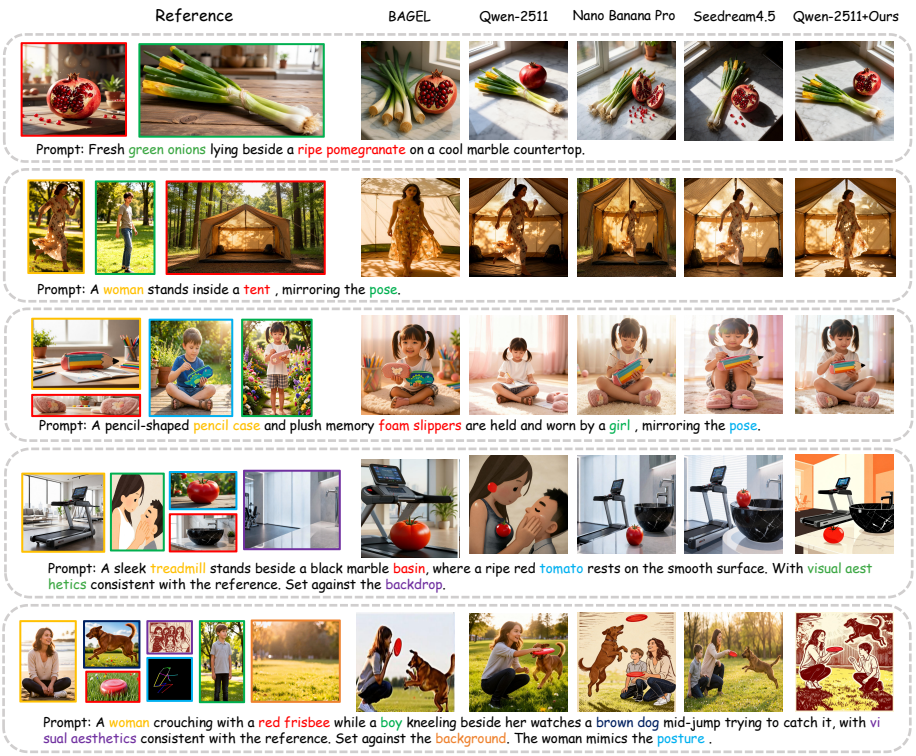

Fig. 6: Qualitative comparison with different models or methods on OmniRef-Bench. Each reference image is framed with a distinct color. Its corresponding phrase in the text prompt is also displayed in the same color.

User Study. To evaluate the consistency between MLLM-based scoring and human preferences, we conduct a user study. We invite 30 experts to rank 20 randomly sampled examples. The evaluation considers text fidelity, reference similarity, composition quality, visual appeal, and task completion, with detailed settings provided in Appendix E. As shown in Tab. 7, ours and Nano Banana Pro rank in the top two, while Seedream 4.5 and Qwen-Image-Edit-2511 rank third and fourth, respectively. Furthermore, the Pearson correlation coefficient between the two is 0.9, indicating strong linear agreement, while the Spearman correlation coefficient is 0.8, suggesting high consistency in ranking. These results show that LLM-based scoring aligns well with human preferences.

4.2 Ablation

We conduct ablation studies on Qwen-Image-Edit-2511 to verify the effectiveness of three of our key training components. The experimental results are reported in Tab. 8, where SFT denotes the results after the first stage of supervised fine-

tuning. Comparisons are made with SFT as the baseline. More ablation studies on hyperparameter sensitivity analysis are provided in Appendix C.2.

Effect of CSD. CSD score is added to our reward to improve the stylization ability of the model. As shown in the third row, removing the CSD reward substantially impairs the model's performance in style-related aspects. Specifically, the objective style metric drops from 0.46 to 0.36, and the style metric in MLLM evaluation drops from 8.67 to 6.33, which confirms that the CSD reward is an effective way to improve the model's stylization ability.

Effect of DAR. Removing DAR degrades the training process into standard GRPO, which has limited discrimination between simple and difficult samples. As shown in the fourth row, both objective metrics and MLLM evaluation results suffer a noticeable decline (0.71 to 0.69 and 8.38 to 8.03, respectively). Notably, even with CSD reward, the style-related score is low compared with our complete training settings (second row), emphasizing DAR's important role to tackle difficult combinations of reference types.

Effect of DRS. As stated in Sec. 3.2, DRS is introduced to magnify the inherent insufficient numerical contrast of objective reward scores. The last row of Tab. 8 shows that while objective metrics remain relatively stable after removing DRS, several MLLM evaluation metrics experience a nontrivial decline. For example, the Style score drops from 8.67 to 8.12, Aes score drops from 8.14 to 8.03. This verifies that DRS can improve the model's performance in high-level semantic aspects, which cannot be effectively captured by low-level objective scores.

5 Conclusion

We propose DyRef to advance multi-reference image generation (MRIG) by addressing the critical performance degradation observed as reference complexity and counts increase. By incorporating the Difficulty-aware Advantage Reweighting and Discriminative Reward Scaling, our approach effectively balances optimization across varying reference scales to mitigate performance variance. Experiments demonstrate that DyRef substantially improves performance on complex MRIG and single-image editing tasks, demonstrating the effectiveness and generalization capability of our approach.

Acknowledgements

This work was supported by the Guangdong Basic and Applied Basic Research Foundation (2025A1515011546) and by the Shenzhen Science and Technology Program (KJZD20240903102901003).

References

1. Chen, B., Zhao, M., Sun, H., Chen, L., Wang, X., Du, K., Wu, X.: Xverse: Consistent multi-subject control of identity and semantic attributes via dit modulation. arXiv preprint arXiv:2506.21416 (2025)
2. Chen, R., Chen, D., Wu, S., Wang, S., Lang, S., Sushko, P., Jiang, G., Wan, Y., Krishna, R.: Multiref: Controllable image generation with multiple visual references. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 13325–13331 (2025)
3. Chen, X., Feng, W., Du, Z., Wang, W., Chen, Y., Wang, H., Liu, L., Li, Y., Zhao, J., Li, Y., et al.: Ctr-driven advertising image generation with multimodal large language models. In: Proceedings of the ACM on Web Conference 2025. pp. 2262–2275 (2025)
4. Chen, Y., Tian, Y., He, M.: Monocular human pose estimation: A survey of deep learning-based methods. Computer vision and image understanding **192**, 102897 (2020)
5. DeepMind, G.: Nanobanana: Gemini 2.5 flash image model. `https://developers.googleblog.com/en/introducing-gemini-2-5-flash-image/` (2025)
6. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., Shi, G., Fan, H.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025)
7. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022)
8. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21807–21818 (2024)
9. Labs, B.F.: FLUX.2: Frontier Visual Intelligence. `https://bfl.ai/blog/flux-2` (2025)
10. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
11. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
12. Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al.: Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437 (2024)
13. 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)
14. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023)
15. Liu, S., Han, Y., Xing, P., Yin, F., Wang, R., Cheng, W., Liao, J., Wang, Y., Fu, H., Han, C., et al.: Step1x-edit: A practical framework for general image editing. arXiv preprint arXiv:2504.17761 (2025)
16. Mou, C., Wu, Y., Wu, W., Guo, Z., Zhang, P., Cheng, Y., Luo, Y., Ding, F., Zhang, S., Li, X., et al.: Dreamo: A unified framework for image customization. In: Proceedings of the SIGGRAPH Asia 2025 Conference Papers. pp. 1–12 (2025)

17. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
18. Oshima, Y., Miyake, D., Matsutani, K., Iwasawa, Y., Suzuki, M., Matsuo, Y., Furuta, H.: Multibanana: A challenging benchmark for multi-reference text-to-image generation. arXiv preprint arXiv:2511.22989 (2025)
19. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
20. Peng, Y., Cui, Y., Tang, H., Qi, Z., Dong, R., Bai, J., Han, C., Ge, Z., Zhang, X., Xia, S.T.: Dreambench++: A human-aligned benchmark for personalized image generation. In: The Thirteenth International Conference on Learning Representations (2025), https://dreambenchplus.github.io/
21. Ping, B., Jia, C., Luo, M., Qian, H., Tsang, I.: Flow-factory: A unified framework for reinforcement learning in flow-matching models. arXiv preprint arXiv:2602.12529 (2026), https://arxiv.org/abs/2602.12529
22. 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)
23. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.Y., Girshick, R., Dollár, P., Feichtenhofer, C.: Sam 2: Segment anything in images and videos (2024), https://arxiv.org/abs/2408.00714
24. 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)
25. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in neural information processing systems **35**, 25278–25294 (2022)
26. Seedream, T., Chen, Y., Gao, Y., Gong, L., Guo, M., Guo, Q., Guo, Z., Hou, X., Huang, W., Huang, Y., et al.: Seedream 4.0: Toward next-generation multimodal image generation. arXiv preprint arXiv:2509.20427 (2025)
27. Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., Sun, J.: Objects365: A large-scale, high-quality dataset for object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8430–8439 (2019)
28. Somepalli, G., Gupta, A., Gupta, K., Palta, S., Goldblum, M., Geiping, J., Shrivastava, A., Goldstein, T.: Measuring style similarity in diffusion models. arXiv preprint arXiv:2404.01292 (2024)
29. Song, X., Wang, L., Wang, W., Li, Z., Sun, J., Zheng, D., Chen, J., Li, Q., Sun, Z.: 3sgen: Unified subject, style, and structure-driven image generation with adaptive task-specific memory. arXiv preprint arXiv:2512.19271 (2025)
30. Song, Y., Liu, C., Shou, M.Z.: Omniconsistency: Learning style-agnostic consistency from paired stylization data (2025), https://api.semanticscholar.org/CorpusID:278905729
31. Tan, Z., Liu, S., Yang, X., Xue, Q., Wang, X.: Ominicontrol: Minimal and universal control for diffusion transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14940–14950 (2025)
32. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)

33. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)
34. Wang, S., Wei, L., He, X., Ouyang, J., Lu, H., Zhao, Z., Tian, Q.: Psr: Scaling multi-subject personalized image generation with pairwise subject-consistency rewards. arXiv preprint arXiv:2512.01236 (2025)
35. Wang, Y., Zhang, J., Shen, C., Yu, H., Luo, S.: Generative ai aids personalized product aesthetic generation and evaluation based on style themes. Advanced Engineering Informatics **68**, 103756 (2025)
36. Wei, H., Wen, B., Long, Y., Yang, Y., Hu, Y., Zhang, T., Chen, W., Fan, H., Jiang, K., Chen, J., et al.: Uniref-image-edit: Towards scalable and consistent multi-reference image editing. arXiv preprint arXiv:2602.14186 (2026)
37. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., ming Yin, S., Bai, S., Xu, X., Chen, Y., Chen, Y., Tang, Z., Zhang, Z., Wang, Z., Yang, A., Yu, B., Cheng, C., Liu, D., Li, D., Zhang, H., Meng, H., Wei, H., Ni, J., Chen, K., Cao, K., Peng, L., Qu, L., Wu, M., Wang, P., Yu, S., Wen, T., Feng, W., Xu, X., Wang, Y., Zhang, Y., Zhu, Y., Wu, Y., Cai, Y., Liu, Z.: Qwen-image technical report (2025), https://arxiv.org/abs/2508.02324
38. Wu, C., Zheng, P., Yan, R., Xiao, S., Luo, X., Wang, Y., Li, W., Jiang, X., Liu, Y., Zhou, J., Liu, Z., Xia, Z., Li, C., Deng, H., Wang, J., Luo, K., Zhang, B., Lian, D., Wang, X., Wang, Z., Huang, T., Liu, Z.: Omnigen2: Exploration to advanced multimodal generation. arXiv preprint arXiv:2506.18871 (2025)
39. Wu, S., Huang, M., Cheng, Y., Wu, W., Tian, J., Luo, Y., Ding, F., He, Q.: Uso: Unified style and subject-driven generation via disentangled and reward learning. arXiv preprint arXiv:2508.18966 (2025)
40. Wu, S., Huang, M., Wu, W., Cheng, Y., Ding, F., He, Q.: Less-to-more generalization: Unlocking more controllability by in-context generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18682–18692 (2025)
41. Xia, B., Peng, B., Liu, J., Wu, S., Li, J., Huang, J., Zhao, X., Wang, Y., Chu, R., Yu, B., et al.: Dreamomni3: Scribble-based editing and generation. arXiv preprint arXiv:2512.22525 (2025)
42. Xia, B., Peng, B., Zhang, Y., Huang, J., Liu, J., Li, J., Tan, H., Wu, S., Wang, C., Wang, Y., et al.: Dreamomni2: Multimodal instruction-based editing and generation. arXiv preprint arXiv:2510.06679 (2025)
43. Xu, R., Zhou, D., Ma, F., Yang, Y.: Contextgen: Contextual layout anchoring for identity-consistent multi-instance generation. arXiv preprint arXiv:2510.11000 (2025)
44. Ye, Y., He, X., Li, Z., Lin, B., Yuan, S., Yan, Z., Hou, B., Yuan, L.: Imgedit: A unified image editing dataset and benchmark. arXiv preprint arXiv:2505.20275 (2025)