



# UniGen-1.5: Enhancing Image Generation and Editing through Reward Unification in Reinforcement Learning

Rui Tian<sup>1,2\*</sup>, Mingfei Gao<sup>2†‡</sup>, Haiming Gang<sup>2</sup>, Jiasen Lu<sup>2</sup>, Zhe Gan<sup>2</sup>, Yinfai Yang<sup>2</sup>, Zuxuan Wu<sup>1‡</sup>, Afshin Dehghan<sup>2</sup>

<sup>1</sup>Institute of Trustworthy Embodied AI, Fudan University    <sup>2</sup>Apple

†Project lead; ‡Corresponding authors

We present *UniGen-1.5*, a unified multimodal large language model (MLLM) for advanced image understanding, generation and editing. Building upon *UniGen*, we comprehensively enhance the model architecture and training pipeline to strengthen the image understanding and generation capabilities while unlocking strong image editing ability. Especially, we propose a unified Reinforcement Learning (RL) strategy that improves both image generation and image editing jointly via shared reward models. To further enhance image editing performance, we propose a light Edit Instruction Alignment stage that significantly improves the editing instruction comprehension that is essential for the success of the RL training. Experimental results show that *UniGen-1.5* demonstrates competitive understanding and generation performance. Specifically, *UniGen-1.5* achieves 0.89 and 4.31 overall scores on GenEval and ImgEdit that surpass the state-of-the-art models such as BAGEL and reaching performance comparable to proprietary models such as GPT-Image-1.

**Date:** November 19, 2025



**Figure 1** Examples of images generated by *UniGen-1.5*.

\*Work done while at Apple.

## 1 Introduction

Unified Multimodal Large Language Models (MLLMs) Chen et al. (2025a); Deng et al. (2025); Lin et al. (2025); Xie et al. (2025a); Jiao et al. (2025); AI et al. (2025); Li et al. (2025) have achieved promising performance across both visual understanding and generation domains. Taking the advantage of strong reasoning ability and knowledge-rich representations of Large Language Models (LLMs) Zhang et al. (2024); McKinzie et al. (2024); Yang et al. (2024, 2025a), unified MLLMs have demonstrated superiority in achieving better semantic consistency compared to vanilla image-generation-only models Betker et al. (2023); Han et al. (2024); Podell et al. (2024) that rely on conditions from text encoders.

Among recent unified MLLMs, *UniGen* Tian et al. (2025) introduced an effective data-centric pipeline for building a competitive model from pre-training to post-training stages. Specifically in post-training, it leverages its intrinsic understanding capability to enhance its generation performance via a chain-of-thought verification (CoT-V) strategy. While CoT-V effectively improves performance on text-to-image generation, it also introduces substantial inference overheads. Additionally, *UniGen* lacks the ability of image editing Wu et al. (2025b); Xiao et al. (2025); Liu et al. (2025b) that is considered as the core for measuring fine-grained controllability of content generation.

We introduce *UniGen-1.5*, that significantly improves *UniGen* with a focus on the post-training stages. We design an effective model architecture for *UniGen-1.5* that supports image understanding, generation as well as editing within a single model. Moreover, we observe that model remains inadequate in handling diverse editing scenarios after supervised fine-tuning due to its insufficient comprehension of the editing instructions. Therefore, we propose *Edit Instruction Alignment* as a light Post-SFT stage to enhance the alignment between editing instruction and the semantics of the target image. Specifically, it takes the condition image and the instruction as inputs and is optimized for predicting the semantic content of the target image via textual descriptions. Experimental results suggest that this stage is highly beneficial for boosting the editing performance.

RL has demonstrated great potential for improving text-to-image generation Jiang et al. (2025a); Wang et al. (2025a); Liu et al. (2025a); Guo et al. (2025b) by encouraging path exploration without incurring huge inference overheads compared to test-time scaling methods such as CoT-V in Tian et al. (2025). However, fewer works Wei et al. (2025) have demonstrated effective ways for elevating image editing using RL for unified MLLMs. We propose that image editing is a more challenging task that involves complicated variations ranging from very subtle changes such as removing/replacing small objects to substantial changes such as altering image style in the pixel space. This raises a big challenge for robust reward modeling. To relieve the issue, we propose to reformulate the image editing task as a general image generation task and optimize it together with the standard text-to-image task via shared reward models under the same schema of RL. Similar to text-to-image generation, we supervise image editing training using the RL reward signals built from directly measuring the alignment between the generated/edited image and its text description. This strategy unlocks us to use the stable text-to-image reward models Wu et al. (2023); Xu et al. (2023); Wang et al. (2025b) for jointly improving both tasks.

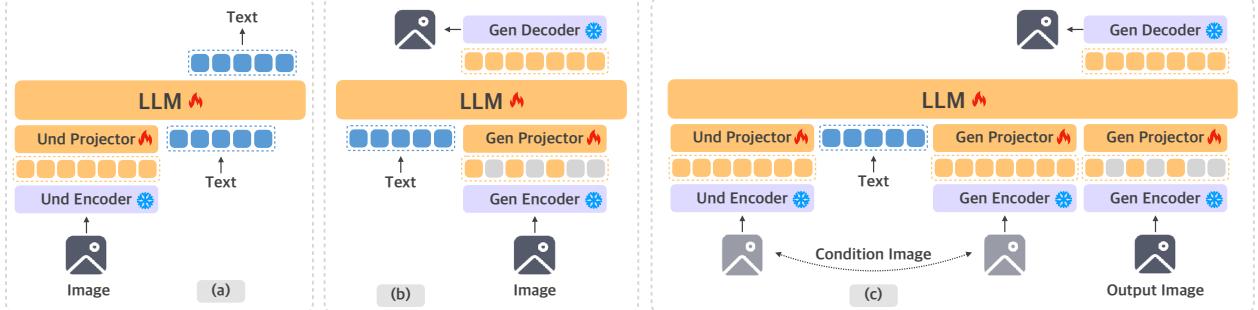
Through the efforts above, *UniGen-1.5* provides a stronger baseline for advancing research on unified MLLMs and establishes competitive performance across image understanding, generation, and editing benchmarks. The experimental results show that *UniGen-1.5* obtains 0.89 and 86.83 on GenEval and DPG-Bench, significantly outperforming recent methods such as BAGEL Deng et al. (2025) and BLIP3o Chen et al. (2025a). For image editing, *UniGen-1.5* achieves 4.31 overall scores on ImgEdit, surpassing recent open-sourced models such as OminiGen2 Wu et al. (2025b) and is comparable to proprietary models such as GPT-Image-1.

Our contributions are summarized as follows:

- We present *UniGen-1.5*, a unified MLLM equipped with an effective model architecture and training pipeline for advanced image understanding, generation and editing.
- We design a unified RL training schema that optimizes image editing and generation using shared reward models that significantly boosts the performance of both tasks.
- We propose *Edit Instruction Alignment* as a Post-SFT stage that significantly improves the editing

performance via enhancing edit instruction comprehension.

- *UniGen-1.5* achieves competitive performance against state-of-the-art unified MLLMs. As shown in Figure 1, we obtain competitive performance on image editing (on-par with GPT-Image-1 on ImgEdit benchmark) and image generation (significantly outperforming BLIP3o on GenEval and DPG-Bench). We also attain strong results on image understanding (comparable to Show-o2 Xie et al. (2025a)).



**Figure 2** The architecture of *UniGen-1.5* jointly optimized for (a) image understanding, (b) text-to-image generation and (c) image editing. See more details in Section 3.1.

## 2 Related Work

There is a growing research trend towards building unified multimodal models capable of image understanding and generation within a single model or framework. Existing approaches can be broadly categorized into three paradigms. First, the unified autoregressive (AR) approach encodes images as either discrete tokens Team (2024); Wang et al. (2024); Wu et al. (2024b); Chen et al. (2025d); Geng et al. (2025) or continuous visual embeddings Sun et al. (2023, 2024); Ge et al. (2024); Tong et al. (2025); Fan et al. (2025), allowing LLMs to treat vision and text as a unified sequence for joint autoregressive prediction. Second, the decoupled LLM-diffusion approach Pan et al. (2025); Wu et al. (2025a,b); Chen et al. (2025a) separates reasoning from image generation, using a frozen LLM for multimodal understanding while offloading image synthesis to a diffusion-based decoder. Third, the hybrid AR-diffusion approach Deng et al. (2025); Zhou et al. (2024); Liang et al. (2024); Xie et al. (2024) integrates both the AR and diffusion paradigms within a single transformer that autoregressively generates text while employing embedded diffusion for visual output. Orthogonal to these modeling strategies, visual tokenizers are central in supporting both semantic understanding and high-fidelity generation. Recent studies have explored both decoupled encoders Wu et al. (2024a) and unified tokenizers Jiao et al. (2025); Li et al. (2025); Ma et al. (2025) to achieve better task balancing. Additionally, emerging research investigates the use of reinforcement learning (RL) to enhance native image generation quality Geng et al. (2025); Chen et al. (2025b); Wei et al. (2025); Mao et al. (2025); Jiang et al. (2025b), which is also the focus of our work. Building upon the *UniGen* framework Tian et al. (2025) which uses masked token prediction for image generation, we effectively integrates image generation and editing within a single RL training framework and optimize for both tasks via shared reward models.

## 3 Methods

### 3.1 Architecture

We build *UniGen-1.5* upon a pre-trained LLM, i.e., Qwen2.5-7B Yang et al. (2025a), and leverage separate encoders for understanding and generation. As shown in Figure 2, we use the discrete visual tokenizer (MAGViTv2 Yu et al. (2024)) for visual generation and continuous visual encoder (SigLIP2 Tschannen et al. (2025)) for visual understanding.

**For image understanding**, we utilize SigLIP2 as our visual encoder  $\text{Enc}^U$ . Comparing to SigLIP with fixed input resolution, e.g.,  $384 \times 384$ , SigLIP2 can receive images with varying input sizes of arbitrary aspect

ratio that is important for maintaining images' native information. An input image  $X^U$  will be projected to a set of continuous tokens  $\mathcal{X}^U = \text{Enc}^U(X^U)$  dependent on its original size. Following the LLaVA Liu et al. (2023) workflow, an MLP-based projector is adopted to align the image and text embeddings into the same space and then the visual embedding together with text embedding are fed into the LLM for response generation via next-token prediction as shown in Figure 2 (a).

**For text-to-image generation**, we generally adopt the same setting as *UniGen* by using masked token prediction Chang et al. (2022) as our training objective. For each image  $X^G$ , we encode it into a sequence of discrete tokens with the generation tokenizer  $\mathcal{X}^G = \text{Enc}^G(X^G)$ . The model is trained to generate target image tokens conditioned on a text prompt  $\mathcal{T}_C$ . During training, we randomly sample a binary mask  $\in \{0, 1\}$  for each token, given a masking ratio  $\eta$  according to a masking scheduling function  $\gamma(\cdot)$ . For each token with mask equal to 1, we replace its corresponding discrete image token  $\mathcal{X}_i^G$  with a special mask token [MASK] to form the final input image sequence. As shown in Figure 2 (b), the LLM takes the text prompt and the masked image sequence tokens as inputs and optimizes for predicting the masked visual tokens back. During inference, the image generation starts with all masked tokens and perform masked token prediction in multiple turns. We set the image-generation resolution to  $384 \times 384$ .

**For image editing**, we unlock this capability during the supervised fine-tuning stage. Given a condition image  $X_C$ , and an editing text prompt  $\mathcal{T}_C$ , we leverage both the understanding encoder and the generation tokenizer, to obtain  $\mathcal{X}_C^U = \text{Enc}^U(X_C)$  and  $\mathcal{X}_C^G = \text{Enc}^G(X_C)$  that extract the continuous (semantic) and discrete (low-level) features from the condition image. We resize condition image to  $384 \times 384$  for feature extraction to ensure capturing substantial details. After projecting the features into a joint space via MLP layers, we sequentially concatenate the semantic visual embedding, the text embedding and the low-level visual embedding (see Figure 2 (c)). We then feed the assembled sequence as the condition for image editing into LLM. The goal is to generate discrete visual tokens  $\mathcal{X}_O^G$  of the output image  $\mathcal{X}_O$ , where  $\mathcal{X}_O^G = \text{Enc}^U(X_O)$ . Similar to the text-to-image generation, we utilize the masked token prediction strategy for image token prediction. The generation resolution for editing is set to  $384 \times 384$ .

### 3.2 Pre-training

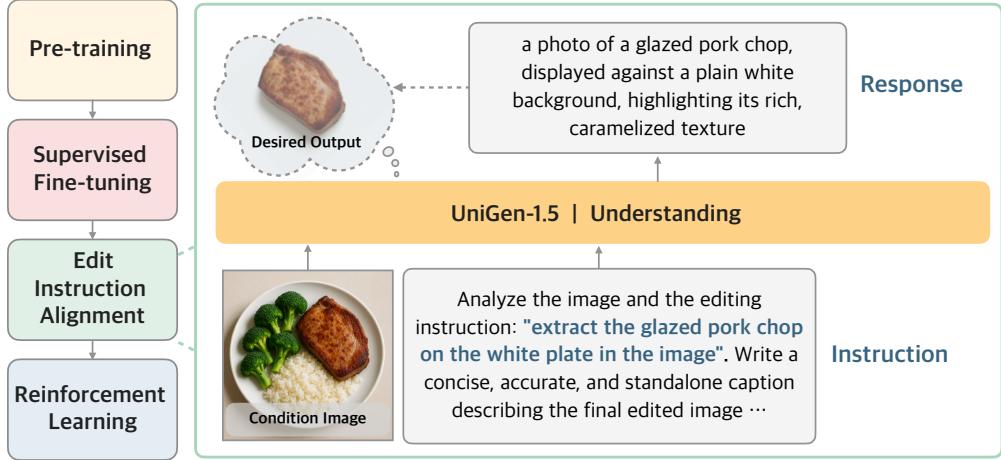
During the pre-training stage of *UniGen-1.5*, we aim to develop foundational visual captioning and generation capabilities with a large collection of well-aligned image–text pairs. Specifically, we employ the pre-training data with fine-grained captions from *UniGen*, composed of ImageNet Ridnik et al. (2021), CC-3M Sharma et al. (2018), CC-12M Changpinyo et al. (2021) and SAM-11M Kirillov et al. (2023). We also include a small portion of text-only training data from RefinedWeb Penedo et al. (2023) to maintain LLM's basic language ability. For simplicity, we adopt only one pre-training stage and unfreeze all the parameters except for the  $\text{Enc}^U$  and  $\text{Enc}^G$ . We include image understanding and text-to-image generation tasks in this stage and set both the resolution of image inputs for generation and understanding to  $384 \times 384$ . We construct each training batch by sampling data from image generation, image understanding and text understanding with a ratio of 3:2:1.

### 3.3 Supervised Fine-tuning

In supervised fine-tuning (SFT), we seek to push forward the generation and understanding performance of *UniGen-1.5* with stronger data mixtures and incentivize *UniGen-1.5*'s image editing ability with joint training.

**Image Generation and Editing.** We follow the architecture introduced in Section 3.1 for both image generation and editing. Inspired by works Chen et al. (2025c,a); Ye et al. (2025a) highlighting the advantage of synthetic data generated by GPT-4o Hurst et al. (2024), we expand our training mixture by adding the high-quality samples proposed in BLIP-3o Chen et al. (2025a) and ShareGPT-4o-Image Chen et al. (2025c). Meanwhile, we unlock image editing by enriching our mixture with image editing data sourced from ShareGPT-4o-Image and GPT-Image-Edit-1.5M Wang et al. (2025c).

**Image Understanding.** We employ the image mixture from SlowFast-LLaVA-1.5 Xu et al. (2025, 2024) to enhance the instruction following capability for image understanding. To encourage the model to perceive



**Figure 3** Illustration of *Edit Instruction Alignment* in the entire training pipeline of *UniGen-1.5*.

finer details of the input image while maintaining the training efficiency, we resize an input image according to the following rules: (1) its width and height must be a multiple of 16 to ensure the compatibility with the patch size of the encoder, (2) the resized image has the closest aspect ratio as its original one, and (3) we maximize the input resolution under the constraint that the number of visual tokens  $\leq 2,304$ . That is approximate to the number of tokens extracted from an image of  $768 \times 768$ .

**Joint SFT Training.** Similar to the pre-training stage, we optimize for three tasks in each training step including generation (either text-to-image generation or image editing), image understanding and text understanding. We use a ratio of 3:4:1 training samples from the above three tasks. In practice, we apply round-robin sampling of text-to-image generation and image editing in every other training batches to improve the training stability. After this joint SFT training, *UniGen-1.5* exhibits the new image editing capability.

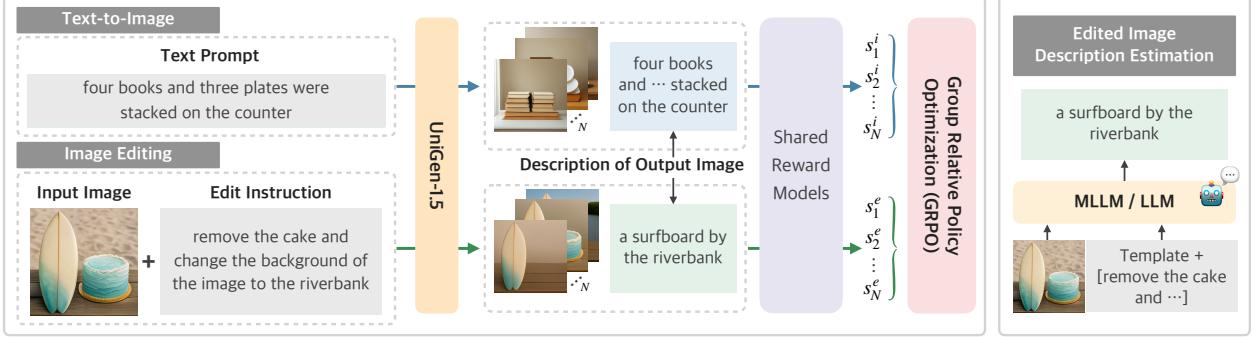
### 3.4 Editing Instruction Alignment

During preliminary experiments of RL, we found that for challenging editing instructions, our model often produced candidates that all failed to satisfy the instruction, resulting in small standard deviations in rewards. Under these circumstances, GRPO receives weak learning signals and struggles to effectively improve the policy. We attribute this challenge to the model’s insufficient ability to comprehend complicated editing instructions, therefore not able to accurately infer the semantic content of output images.

To mitigate this issue, we include *Editing Instruction Alignment* as a Post-SFT stage to enhance the alignment of editing instruction and semantic content of the desired output. As shown in Figure 3, *UniGen-1.5* takes condition image and editing instruction as inputs and is optimized for predicting the textual description of expected output image that forms a vital bridge to the final visual generation. Consequently, the model develops a more faithful understanding of editing intentions, enabling semantically coherent yet diverse candidate generation and providing informative learning signals during RL. See details of training data in Section A.

### 3.5 Reinforcement Learning

We improve the overall visual generation quality of *UniGen-1.5* via a RL stage empowered with group relative policy optimization (GRPO) Shao et al. (2024); Guo et al. (2025a). While a series of works highlight the effectiveness of GRPO over improving the performance of text-to-image generation Wang et al. (2025a); Jiang et al. (2025a); Huang et al. (2025), its impacts on more generalized forms of visual generation, e.g., image editing, remains underexplored. In *UniGen-1.5*, we propose to unify the RL training for both text-to-image generation and image editing as shown in Figure 4. Specifically, we propose that the output image’s quality from both tasks can be assessed by measuring the semantic alignment between the image and its corresponding text description.



**Figure 4 Left:** The pipeline of GRPO training in *UniGen-1.5*. We utilize shared reward models for both text-to-image generation and image editing. For the former, we directly input the generated image with the *text prompt* to obtain rewards. For the latter, we get reward signals by measuring the alignment between the *edited image description* and the generated image. **Right:** The pipeline of edited image description estimation. We leverage powerful external MLLMs and LLMs to generate the description of desired edited images.

**RL formulation.** Initialized from our Post-SFT model, *UniGen-1.5* acts as a policy model,  $\pi_\theta$ , that takes different conditions as inputs and generates the corresponding sequence of visual tokens  $\hat{\mathcal{X}}_O^G$ . For the text-to-image task, the condition is simply the text embedding of the prompt  $\mathcal{T}_C$ , while for editing task, *UniGen-1.5* conditions the image generation on  $\mathcal{X}_C^U$ , an editing text embedding  $\mathcal{T}_C$  and  $\mathcal{X}_C^G$ . During training, we sample  $N$  sequences  $\{\hat{\mathcal{X}}_{O1}^G, \dots, \hat{\mathcal{X}}_{ON}^G\}$  as output candidates from  $\pi_\theta$ , each of which will be assigned a scalar reward  $R_i$ . The rewards are used to calculate a group-normalized advantage as in Eq.3.1.

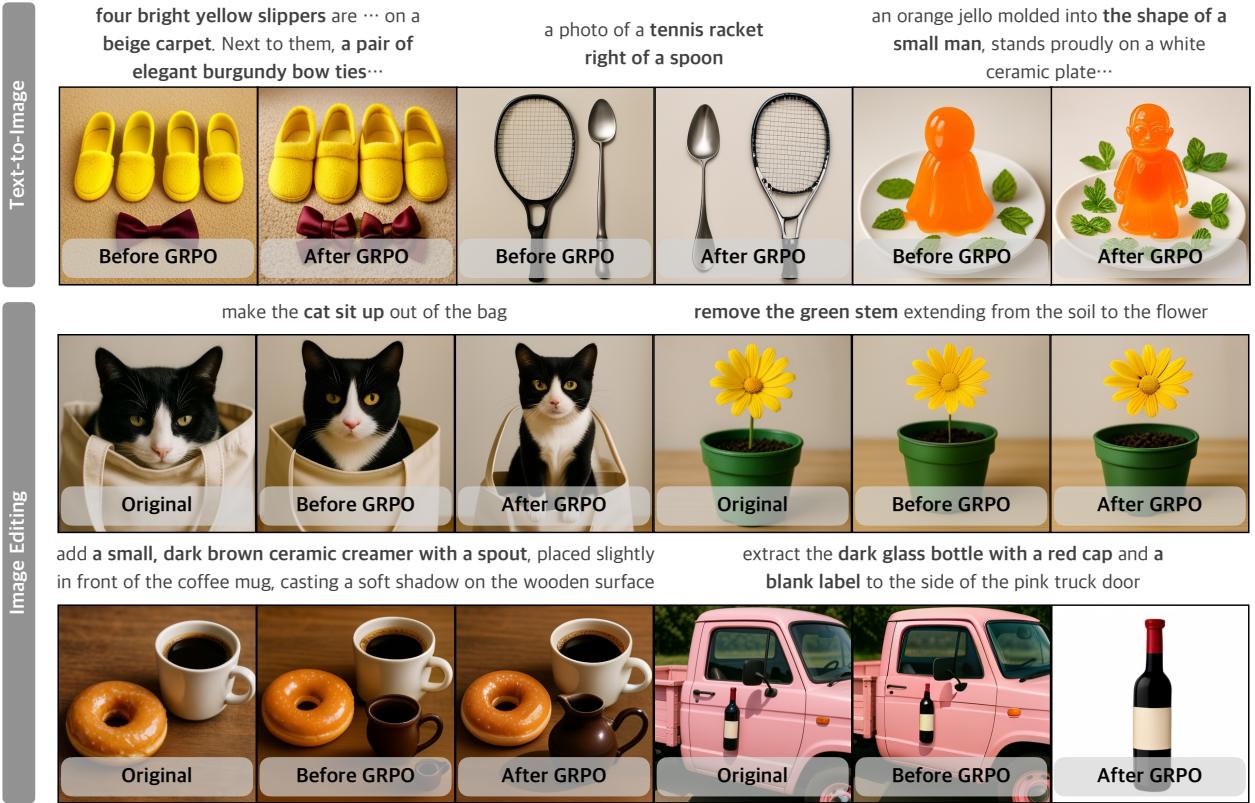
$$A_i = \frac{R_i - \text{mean}\{R_1, R_2, \dots, R_N\}}{\text{std}\{R_1, R_2, \dots, R_N\}} \quad (3.1)$$

The parameters of our policy model are updated by optimizing the training objective in Eq.3.2, where  $\pi_{ref}$  indicates reference policy (initial policy),  $\rho_i = \frac{\pi_\theta}{\pi_{\theta_{old}}}$  refers to the importance sampling ratio and  $\pi_{\theta_{old}}$  indicates the old policy before update.

$$\begin{aligned} \mathcal{J}(\theta) = & \frac{1}{N} \sum_{i=1}^N \min(\rho_i A_i, \text{clip}(\rho_i, 1 - \epsilon, 1 + \epsilon) A_i) \\ & - \beta D_{KL}(\pi_\theta \| \pi_{ref}) \end{aligned} \quad (3.2)$$

**Shared reward models.** Designing editing rewards is inherently more challenging than designing rewards for text-to-image generation, as visual edits are often subtle, localized, and highly context-dependent. Moreover, training editing-specific reward models Luo et al. (2025); Li et al. (2024) requires substantial manual annotation costs for collecting large-scale image editing data from diverse categories and obtaining high-quality labels that align with human preference. These challenges make it extremely difficult to construct reliable editing rewards at scale. To this end, we propose to leverage robust, well-developed text-to-image reward models to evaluate edited image.

Specifically, We introduce a unified RL formulation of image generation and editing by assessing the quality for both tasks using  $\mathbf{R}(\hat{\mathcal{X}}_O^G, \mathcal{T}_O)$ , where  $\mathbf{R}(\cdot)$  denotes the shared reward functions,  $\hat{\mathcal{X}}_O^G$  indicates the generated image in pixel space and  $\mathcal{T}_O$  refers to the text description of expected output. We directly use the ground-truth text prompt as  $\mathcal{T}_O$  for text-to-image generation and use the textual caption synthesized by Qwen2.5-72B for image editing (see Section A for details). We believe that a powerful LLM is capable of reliably reflecting the visual differences, capturing the details and layout of the edited image in its description, regardless of varying modification magnitude. Inspired by T2I-R1 Jiang et al. (2025a), we opt to implement  $\mathbf{R}(\cdot)$  with an ensemble of diverse vision experts to assign rewards for image candidates. Our reward models include CLIP-H Fang et al. (2023); Cherti et al. (2023), HPSv2 Wu et al. (2023), Unified-Reward-7B Wang et al. (2025b) and ORM Jiang et al. (2025a).



**Figure 5** Examples generated by *UniGen-1.5*, highlighting the contribution of GRPO training.

**RL Training Data.** For text-to-image generation, we use the training set from T2I-R1 [Jiang et al. \(2025a\)](#), including 6,786 prompts sourced from T2I-CompBench [Huang et al. \(2023\)](#) and PARM [Guo et al. \(2025b\)](#). For training samples of image editing, we collect the Edit-RL dataset with 10,568 samples. We generate the condition images using Qwen-Image [Wu et al. \(2025a\)](#), and construct versatile edit instructions with Qwen-2.5-VL-72B-instruct [Bai et al. \(2025\)](#) based on our designed templates. Moreover, we use Qwen-2.5-72B-instruct [Yang et al. \(2024\)](#) to synthesize output descriptions for the desired images. These pseudo labeled descriptions will be used to calculate the rewards of edited images during training. See details in [Section A](#).

## 4 Experiments

### 4.1 Implementation Details

We initialize *UniGen-1.5* with pre-trained Qwen2.5-7B [Yang et al. \(2024\)](#) LLM, and adopt MAGVITv2 from Show-o [Xie et al. \(2024\)](#) as our discrete visual encoder with input resolution of  $384 \times 384$  and siglip2-so400m-patch16-naflex [Tschanne et al. \(2025\)](#) as our continuous visual encoder for native image resolution. For both image generation and editing, we leverage MAGVITv2’s decoder to project the visual tokens back to the image pixel space. Both discrete and continuous encoders are kept frozen across all the training stages.

During pre-training, we use 96 H100-80G GPUs, set the batch size to 576 and adopt the learning rate of  $1e^{-4}$ . For supervised fine-tuning, we use 64 H100-80G GPUs, set the batch size to 128 and the learning rate to  $2e^{-5}$ . For the *Edit Instruction Alignment* stage, we train on the collected Edit-Align dataset (See details in [Section A](#)) for 500 steps on 8 H100-80G GPUs with a batch size of 64. In this stage, we set the learning rate to  $1e^{-5}$  and adopt the cosine schedule. To accommodate the classifier-free guidance during inference, we randomly drop text prompts during training for both text-to-image and image editing tasks with a probability of 10%, while dropping  $\mathcal{X}_C^U$  and  $\mathcal{X}_C^G$  for image editing training samples with probabilities of 50% and 10%, respectively.

**Table 1** Comparison with baseline models on ImgEdit benchmark. The best and second-best results are highlighted in **bold** and underlined, respectively. *UniGen-1.5* achieves the best overall score against all the other models.

Model	#Params	Add	Adjust	Extract	Replace	Remove	Background	Style	Compose	Action	Overall
Step1X-Edit Liu et al. (2025b)	19B	3.88	3.14	1.76	3.40	2.41	3.16	4.63	2.64	2.52	3.06
UniWorld-V1 Lin et al. (2025)	7B + 12B	3.82	3.64	2.27	3.47	3.24	2.99	4.21	2.96	2.74	3.26
BAGEL Deng et al. (2025)	7B MoT	3.56	3.31	1.70	3.30	2.62	3.24	4.49	2.38	4.17	3.20
OmniGen Xiao et al. (2025)	7B	3.47	3.04	1.71	2.94	2.43	3.21	4.19	2.24	3.38	2.96
OmniGen2 Wu et al. (2025b)	7B	3.57	3.06	1.77	3.74	3.2	3.57	<u>4.81</u>	2.52	4.68	3.44
FLUX.1 Kontext [Pro] Batifol et al. (2025)	-	4.25	4.15	2.35	4.56	3.57	4.26	4.57	3.68	4.63	4.00
Qwen-Image Wu et al. (2025a)	7B	<u>4.38</u>	4.16	<u>3.43</u>	<u>4.66</u>	<u>4.14</u>	4.38	<u>4.81</u>	3.82	<u>4.69</u>	<u>4.27</u>
GPT Image 1 [High] Achiam et al. (2023)	-	<b>4.61</b>	<b>4.33</b>	2.9	4.35	3.66	<u>4.57</u>	<b>4.93</b>	<b>3.96</b>	<b>4.89</b>	4.20
<i>UniGen-1.5</i>	7B	4.31	<u>4.18</u>	<b>3.86</b>	<b>4.78</b>	<b>4.57</b>	<u>4.50</u>	4.69	<u>3.88</u>	4.00	<b>4.31</b>

**Table 2** Comparison with state-of-the-art models on GenEval and DPG-Bench. The best and second-best results are highlighted in **bold** and underlined, respectively. *UniGen-1.5* achieves the best performance on both benchmarks.

Model	# Params	GenEval↑					DPG-Bench↑		
		Two Obj.	Counting	Position	Color Attr.	Overall	Attribute	Entity	Overall
<i>Text-to-Image Generation Models</i>									
DALLE-3 Betker et al. (2023)	-	0.87	0.47	0.43	0.45	0.67	88.96	89.61	83.50
Emu3 Wang et al. (2024)	8B	0.71	0.34	0.17	0.21	0.54	88.39	86.68	80.60
Infinity Han et al. (2024)	2B	0.85	-	0.49	0.57	0.73	-	90.76	83.46
Lumia-Image-2.0 Qin et al. (2025)	2.6B	0.87	0.67	-	0.62	0.73	90.20	91.97	87.20
GPT Image 1 [High] Achiam et al. (2023)	-	0.92	0.85	0.75	0.61	0.84	89.84	88.94	85.15
<i>Unified MLLMs</i>									
Janus Wu et al. (2024a)	1.3B	0.68	0.30	0.46	0.42	0.61	87.70	87.38	79.68
<i>UniGen</i> Tian et al. (2025)	1.5B	<b>0.94</b>	<u>0.78</u>	<u>0.57</u>	<u>0.54</u>	<u>0.78</u>	<b>90.90</b>	<u>89.68</u>	<u>85.19</u>
Manzano Li et al. (2025)	3B	0.91	<b>0.82</b>	0.78	<u>0.71</u>	<u>0.85</u>	-	-	-
TokenFlow-XL Qu et al. (2024)	13B	0.72	0.45	0.45	0.42	0.63	81.29	79.22	73.38
Janus-Pro Chen et al. (2025d)	7B	0.89	0.59	<u>0.79</u>	0.66	0.80	89.40	88.90	84.19
BAGEL Deng et al. (2025)	7B MoT	<b>0.94</b>	<u>0.81</u>	0.64	0.63	0.82	-	-	-
Show-o2 Xie et al. (2025a)	7B	0.87	0.58	0.52	0.62	0.76	89.96	<u>91.78</u>	<u>86.14</u>
BLIP3-o Chen et al. (2025a)	8B	-	-	-	-	0.84	-	-	81.60
<i>UniGen-1.5</i>	7B	<u>0.93</u>	0.80	<b>0.92</b>	<u>0.81</u>	<b>0.89</b>	<u>90.55</u>	<b>92.64</b>	<b>86.83</b>

During GRPO, we follow T2I-R1 Jiang et al. (2025a) to remove the traditional ratio clipping and simply leverage an explicit KL-penalty regularization to constrain policy updates. We conduct GRPO training for 1500 steps with the learning rate set to  $3e^{-6}$  and the batch size set to 32 on 8 B200 GPUs. We set the KL-penalty coefficient  $\beta$  to 0.01 and generate  $N = 8$  image candidates for each input. To accelerate training with minimal impact on performance, we sample each image candidate using only 16 decoding steps and disable the classifier-free guidance.

During inference, we follow MaskGIT Chang et al. (2022) to use the cosine masking schedule and set the default number of generation steps to 50. Moreover, we follow the common practice to employ classifier free guidance scale. Specifically, the guidance scale for text-to-image generation is set to 5.0. As for image editing, we formulate the generation process with classifier free guidance as

$$\begin{aligned} \mathcal{X}_O &= \mathcal{P}_\theta(\emptyset, \emptyset, \emptyset) \\ &+ s_I \cdot (\mathcal{P}_\theta(\mathcal{X}_C^U, \emptyset, \mathcal{X}_C^G) - \mathcal{P}_\theta(\emptyset, \emptyset, \emptyset)) \\ &+ s_T \cdot (\mathcal{P}_\theta(\mathcal{X}_C^U, \mathcal{T}_C, \mathcal{X}_C^G) - \mathcal{P}_\theta(\mathcal{X}_C^U, \emptyset, \mathcal{X}_C^G)), \end{aligned}$$

where  $\mathcal{P}_\theta$  represents the parameter of *UniGen-1.5*,  $\emptyset$  denotes empty (drop the condition),  $s_T$  refers to the guidance scale of editing instruction and  $s_I$  refers to that of the condition image. For evaluation on the ImgEdit benchmark, we set  $s_T$  and  $s_I$  to 3 and 1.5, respectively.

## 4.2 Main Results

We compare *UniGen-1.5* with state-of-the-art unified MLLMs in Table 1, Table 2 and Table 3 and summarize the following findings based on the experimental results.

**First, *UniGen-1.5* obtains competitive performance on image editing benchmarks.** As shown in Table 1, *UniGen-1.5* demonstrates state-of-the-art performance on ImgEdit. Without leveraging external

**Table 3** Comparison with state-of-the-art models on image understanding benchmarks. \*denotes reproduced results. The best and second-best results are highlighted in **bold** and underlined, respectively.

Model	#Params	AI2D	GQA	POPE	MMMU	MathVista	ScienceQA	Seedbench
Janus <a href="#">Wu et al. (2024a)</a>	1.3B	49.0*	59.1	87.0	30.5	33.7*	76.5*	63.7
OmniMamba <a href="#">Zou et al. (2025)</a>	1.3B	-	60.8	86.3	30.6	-	-	-
Janus-Pro <a href="#">Chen et al. (2025d)</a>	1.5B	63.7*	59.3	86.2	36.3	36.8*	75.5*	68.3
RecA <a href="#">Chen et al. (2025d)</a>	1.5B	-	58.4	83.2	35.7	-	-	65.3
ULM-R1 <a href="#">Jiang et al. (2025b)</a>	1.5B	-	-	<b>88.9</b>	<u>42.3</u>	42.5	-	-
Harmon <a href="#">Wu et al. (2025c)</a>	1.5B	-	58.9	87.6	38.9	-	-	67.1
<i>UniGen</i> <a href="#">Tian et al. (2025)</a>	1.5B	<b>67.4</b>	62.3	87.8	32.3	44.6	<b>79.4</b>	<b>70.8</b>
UniToken <a href="#">Jiao et al. (2025)</a>	7B	68.7	-	-	32.8	38.5	-	69.9
Show-o2 <a href="#">Xie et al. (2025a)</a>	7B	<b>78.6</b>	<u>63.1</u>	-	<b>48.9</b>	-	-	69.8
MUSE-VL <a href="#">Xie et al. (2025b)</a>	7B	69.8	-	-	39.7	<u>51.3</u>	-	69.1
MMaDA <a href="#">Yang et al. (2025b)</a>	8B	-	61.3	86.1	30.2	-	-	64.2
<i>UniGen-1.5</i>	7B	<u>77.4</u>	<b>63.7</b>	<u>88.3</u>	35.9	<b>51.9</b>	<b>86.3</b>	<b>76.5</b>

diffusion models, *UniGen-1.5* is leading the benchmark with an overall score significantly outperforming recent models of similar model size such as BAGEL and OmniGen2. Notably, *UniGen-1.5* even achieves slightly better performance than GPT-Image-1.

**Second, *UniGen-1.5* achieves promising performance on text-to-image generation benchmarks.** *UniGen-1.5* yields the final score of 0.89 and 86.83 on GenEval and DPG-Bench, respectively. Compared with *UniGen*, there's a growth of 0.11 on GenEval and 1.6 on DPG-Bench. *UniGen-1.5* also beats a range of state-of-the-art unified MLLMs on GenEval, especially on the “Position” category. For example, *UniGen-1.5* significantly outperforms Show-o2, BLIP3-o and BAGEL by 0.13, 0.05 and 0.07 points in overall score. On DPG-Bench, *UniGen-1.5* largely surpass BLIP3-o by more than 5 points.

**Third, *UniGen-1.5* effectively improves *UniGen* on understanding benchmarks.** As shown in [Table 3](#), *UniGen-1.5* significantly improves *UniGen* across all the benchmarks. We attribute the improvements to three aspects, 1) we scale up the model size to 7B, enhancing the overall capability of the unified MLLM, 2) we increase the resolution of input images and keep the original aspect ratio that are beneficial for maintaining images' native information and 3) we perform understanding-based pre-training, alleviating the mismatch between the training objective for generation and understanding. When compared to other strong unified MLLMs of similar size, *UniGen-1.5* still demonstrates competitive performance, achieving superior numbers than UniToken, MUSE-VL and MMaDA on most of the benchmarks and on-par results with Show-o2.

### 4.3 Ablation Results

#### 4.3.1 The impact of Unified RL

**The RL (GRPO) stage significantly improves both image generation and editing tasks.** Comparing first row and last row in [Table 4](#), we observe a considerable gain introduced by the RL stage, where all three benchmarks are improved with clear margin (from 0.85 to 0.89 in GenEval, from 84.19 to 86.83 in DPG-Bench and from 3.93 to 4.31 in ImgEdit). We also show the comparison qualitatively in [Figure 5](#). For text-to-image task, *UniGen-1.5* demonstrates better semantic alignment between text prompts and generated images in diverse scenarios including counting (1st example), position (2nd example) and shape (3rd example). For image editing, we observe that *UniGen-1.5* achieves finer-grained control over the condition images after GRPO. For example, it successfully makes the “cat sit up” (1st example) and the “glass bottle extracted” (last example) which failed before GRPO. Furthermore, we argue that GRPO introduces no performance drop in understanding (See [Section D.1](#)).

**Removing either text-to-image or image editing in RL results in significant performance drop.** When discarding image editing in the RL stage, the image generation benchmarks (GenEval and DPG-Bench) are comparable to *UniGen-1.5*, but there is large drop on ImgEdit benchmark (row 2 vs. row 4 in [Table 4](#)). When removing text-to-image in RL training, we observe significant performance degradation on text-to-image generation. Keeping both tasks leads to the best overall performance.

**Table 4 Ablation of Unified RL.** We train *UniGen-1.5* with different tasks during RL for same steps. T2I stands for text-to-image generation and I-Edit represents image editing. We report the overall score for GenEval, DPG-Bench and ImgEdit benchmarks. We highlight the default setting of *UniGen-1.5* in light blue.

T2I	I-Edit	GenEval	DPG-Bench	ImgEdit
		0.85	84.19	3.93
✓		0.90	86.62	4.01
	✓	0.85	86.39	4.32
✓	✓	0.89	86.83	4.31

**Table 5 Ablation of Edit Instruction Alignment.** We report the overall score for GenEval, DPG-Bench and ImgEdit benchmarks. We highlight the default setting of *UniGen-1.5* in light blue.

Edit Inst. alignment	Unified RL	GenEval	DPG-Bench	ImgEdit
		0.83	83.92	3.87
✓		0.85	84.19	3.93
	✓	0.90	86.96	4.08
✓	✓	0.89	86.83	4.31

#### 4.3.2 The impact of Edit Instruction Alignment

**Edit Instruction Alignment is important prior to RL.** We first evaluate the effect of adding this stage by comparing the results to those in the SFT stage. As shown in [Table 5](#) (row 1 vs. row 2), adding *Edit Instruction Alignment* boost performance of all the three benchmarks even before RL that suggests the advantage of this stage in general.

**The impact of Edit Instruction Alignment is amplified in RL.** As shown in [Table 5](#) (row 3 vs. row 4), adding the *Edit Instruction Alignment* stage is crucial for image editing after RL. Without this stage, *UniGen-1.5* improves ImgEdit by 0.21 points with RL (row 1 vs. row 3). Benefited from the refined semantic alignment introduced by this stage, RL achieves much larger gain by 0.38 points (row 2 vs. row 4). We also show in [Section D.1](#) that this stage does not sacrifice performance for image understanding.

## 5 Conclusion

We presented *UniGen-1.5*, a unified MLLM that achieves competitive performance across image understanding, generation and editing tasks. Building upon the *UniGen* framework, *UniGen-1.5* enhances the model architecture to extend its capabilities to support image editing and further improves it with our designed *Edit Instruction Alignment* stage. We also proposed a unified RL strategy that jointly optimizes generation and editing via shared reward models, leading to substantial gains in both fidelity and controllability. Extensive experiments demonstrate that *UniGen-1.5* achieves state-of-the-art results on a wide range of benchmarks for image understanding, text-to-image generation, and image editing, establishing a strong and extensible baseline to advance future research on unified MLLMs.

**Limitation.** First, *UniGen-1.5* is not proficient in rendering textual contents ([Figure A](#) 1st row). Our model focuses on improving semantic alignment between text instructions and discrete visual tokens and uses only a light-weight visual detokenizer for image reconstruction. This leads to a disadvantage in generating text, which critically relies on preserving fine-grained structural details. We believe that integrating a diffusion-based component into the framework can effectively address this limitation. Second, *UniGen-1.5* still suffers from visual inconsistency ([Figure A](#) last row), a key challenge in image editing tasks. A dedicated reward model is required to enforce visual consistency during RL. We leave this direction for future work.

**Acknowledgment.** We thank Mingze Xu, Peter Fu and Oğuzhan Fatih Kar for their kind help.

## References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- Inclusion AI, Biao Gong, Cheng Zou, Chuanyang Zheng, Chunluan Zhou, Canxiang Yan, Chunxiang Jin, Chunjie Shen, Dandan Zheng, Fudong Wang, et al. Ming-omni: A unified multimodal model for perception and generation. *arXiv preprint arXiv:2506.09344*, 2025.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv e-prints*, pages arXiv–2506, 2025.
- James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science.*, 2023.
- Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *CVPR*, 2022.
- Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021.
- Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025a.
- Jiuhai Chen, Le Xue, Zhiyang Xu, Xichen Pan, Shusheng Yang, Can Qin, An Yan, Honglu Zhou, Zeyuan Chen, Lifu Huang, et al. Blip3o-next: Next frontier of native image generation. *arXiv preprint arXiv:2510.15857*, 2025b.
- Junying Chen, Zhenyang Cai, Pengcheng Chen, Shunian Chen, Ke Ji, Xidong Wang, Yunjin Yang, and Benyou Wang. Sharegpt-4o-image: Aligning multimodal models with gpt-4o-level image generation. *arXiv preprint arXiv:2506.18095*, 2025c.
- Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv:2501.17811*, 2025d.
- Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *ICML*, 2023.
- Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.
- Lijie Fan, Luming Tang, Siyang Qin, Tianhong Li, Xuan Yang, Siyuan Qiao, Andreas Steiner, Chen Sun, Yuanzhen Li, Tao Zhu, et al. Unified autoregressive visual generation and understanding with continuous tokens. *arXiv preprint arXiv:2503.13436*, 2025.
- Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. *arXiv preprint arXiv:2309.17425*, 2023.
- Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*, 2024.
- Zigang Geng, Yibing Wang, Yeyao Ma, Chen Li, Yongming Rao, Shuyang Gu, Zhao Zhong, Qinglin Lu, Han Hu, Xiaosong Zhang, et al. X-omni: Reinforcement learning makes discrete autoregressive image generative models great again. *arXiv preprint arXiv:2507.22058*, 2025.
- Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In *NeurIPS*, 2023.

- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025a.
- Ziyu Guo, Renrui Zhang, Chengzhuo Tong, Zhizheng Zhao, Rui Huang, Haoquan Zhang, Manyuan Zhang, Jiaming Liu, Shanghang Zhang, Peng Gao, et al. Can we generate images with cot? let's verify and reinforce image generation step by step. *arXiv preprint arXiv:2501.13926*, 2025b.
- Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. *arXiv:2412.04431*, 2024.
- Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv:2403.05135*, 2024.
- Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. In *NeurIPS*, 2023.
- Wenxuan Huang, Shuang Chen, Zheyong Xie, Shaosheng Cao, Shixiang Tang, Yufan Shen, Qingyu Yin, Wenbo Hu, Xiaoman Wang, Yuntian Tang, et al. Interleaving reasoning for better text-to-image generation. *arXiv preprint arXiv:2509.06945*, 2025.
- Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- jackyhate. text-to-image-2m: A high-quality, diverse text-to-image training dataset. huggingface, 2024.
- Dongzhi Jiang, Ziyu Guo, Renrui Zhang, Zhuofan Zong, Hao Li, Le Zhuo, Shilin Yan, Pheng-Ann Heng, and Hongsheng Li. T2i-r1: Reinforcing image generation with collaborative semantic-level and token-level cot. *arXiv preprint arXiv:2505.00703*, 2025a.
- Jingjing Jiang, Chongjie Si, Jun Luo, Hanwang Zhang, and Chao Ma. Co-reinforcement learning for unified multimodal understanding and generation. *arXiv preprint arXiv:2505.17534*, 2025b.
- Yang Jiao, Haibo Qiu, Zequn Jie, Shaoxiang Chen, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. Unitoken: Harmonizing multimodal understanding and generation through unified visual encoding. *arXiv:2504.04423*, 2025.
- Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, 2016.
- Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023.
- Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv:2307.16125*, 2023a.
- Tiancheng Li, Jinxiu Liu, Huajun Chen, and Qi Liu. Instructrl4pix: Training diffusion for image editing by reinforcement learning. *arXiv preprint arXiv:2406.09973*, 2024.
- Yanghao Li, Rui Qian, Bowen Pan, Haotian Zhang, Haoshuo Huang, Bowen Zhang, Jialing Tong, Haoxuan You, Xianzhi Du, Zhe Gan, et al. Manzano: A simple and scalable unified multimodal model with a hybrid vision tokenizer. *arXiv preprint arXiv:2509.16197*, 2025.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, 2023b.
- Weixin Liang, Lili Yu, Liang Luo, Srinivasan Iyer, Ning Dong, Chunting Zhou, Gargi Ghosh, Mike Lewis, Wen-tau Yih, Luke Zettlemoyer, et al. Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models. *arXiv preprint arXiv:2411.04996*, 2024.
- Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, et al. Uniworld: High-resolution semantic encoders for unified visual understanding and generation. *arXiv preprint arXiv:2506.03147*, 2025.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2023.
- Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025a.
- Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025b.
- Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.
- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *ICLR*, 2024.
- Xin Luo, Jiahao Wang, Chenyuan Wu, Shitao Xiao, Xiyan Jiang, Defu Lian, Jiajun Zhang, Dong Liu, et al. Editscore: Unlocking online rl for image editing via high-fidelity reward modeling. *arXiv preprint arXiv:2509.23909*, 2025.
- Chuofan Ma, Yi Jiang, Junfeng Wu, Jihan Yang, Xin Yu, Zehuan Yuan, Bingyue Peng, and Xiaojuan Qi. Unitok: A unified tokenizer for visual generation and understanding. *arXiv preprint arXiv:2502.20321*, 2025.
- Weijia Mao, Zhenheng Yang, and Mike Zheng Shou. Unirl: Self-improving unified multimodal models via supervised and reinforcement learning. *arXiv preprint arXiv:2505.23380*, 2025.
- Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Anton Belyi, et al. Mm1: methods, analysis and insights from multimodal llm pre-training. In *ECCV*, 2024.
- Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Juhai Chen, Kunpeng Li, Felix Juefei-Xu, et al. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025.
- Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*, 2023.
- Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024.
- Qi Qin, Le Zhuo, Yi Xin, Ruoyi Du, Zhen Li, Bin Fu, Yiting Lu, Jiakang Yuan, Xinyue Li, Dongyang Liu, et al. Lumina-image 2.0: A unified and efficient image generative framework. *arXiv preprint arXiv:2503.21758*, 2025.
- Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K Du, Zehuan Yuan, and Xinglong Wu. Tokenflow: Unified image tokenizer for multimodal understanding and generation. *arXiv:2412.03069*, 2024.
- Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv:2104.10972*, 2021.
- Zhilong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
- Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018.
- Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Emu: Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*, 2023.
- Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *CVPR*, 2024.
- Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.
- Rui Tian, Mingfei Gao, Mingze Xu, Jiaming Hu, Jiasen Lu, Zuxuan Wu, Yinfei Yang, and Afshin Dehghan. Unigen: Enhanced training & test-time strategies for unified multimodal understanding and generation. *arXiv preprint arXiv:2505.14682*, 2025.

- Shengbang Tong, David Fan, Jiachen Li, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. Metamorph: Multimodal understanding and generation via instruction tuning. In *ICCV*, 2025.
- Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.
- Junke Wang, Zhi Tian, Xun Wang, Xinyu Zhang, Weilin Huang, Zuxuan Wu, and Yu-Gang Jiang. Simplear: Pushing the frontier of autoregressive visual generation through pretraining, sft, and rl. *arXiv preprint arXiv:2504.11455*, 2025a.
- Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv:2409.18869*, 2024.
- Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025b.
- Yuhan Wang, Siwei Yang, Bingchen Zhao, Letian Zhang, Qing Liu, Yuyin Zhou, and Cihang Xie. Gpt-image-edit-1.5 m: A million-scale, gpt-generated image dataset. *arXiv preprint arXiv:2507.21033*, 2025c.
- Hongyang Wei, Baixin Xu, Hongbo Liu, Cyrus Wu, Jie Liu, Yi Peng, Peiyu Wang, Zexiang Liu, Jingwen He, Yidan Xietian, et al. Skywork unipic 2.0: Building kontext model with online rl for unified multimodal model. *arXiv preprint arXiv:2509.04548*, 2025.
- Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025a.
- Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv:2410.13848*, 2024a.
- Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025b.
- Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Zhonghua Wu, Qingyi Tao, Wentao Liu, Wei Li, and Chen Change Loy. Harmonizing visual representations for unified multimodal understanding and generation. *arXiv preprint arXiv:2503.21979*, 2025c.
- Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.
- Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024b.
- Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuteng Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *CVPR*, 2025.
- Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv:2408.12528*, 2024.
- Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *arXiv preprint arXiv:2506.15564*, 2025a.
- Rongchang Xie, Chen Du, Ping Song, and Chang Liu. Muse-vl: Modeling unified vlm through semantic discrete encoding. In *ICCV*, 2025b.
- Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *NeurIPS*, 2023.
- Mingze Xu, Mingfei Gao, Zhe Gan, Hong-You Chen, Zhengfeng Lai, Haiming Gang, Kai Kang, and Afshin Dehghan. Slowfast-llava: A strong training-free baseline for video large language models. *arXiv:2407.15841*, 2024.

- Mingze Xu, Mingfei Gao, Shiyu Li, Jiasen Lu, Zhe Gan, Zhengfeng Lai, Meng Cao, Kai Kang, Yinfei Yang, and Afshin Dehghan. Slowfast-llava-1.5: A family of token-efficient video large language models for long-form video understanding. *arXiv:2503.18943*, 2025.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv:2412.15115*, 2024.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengan Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.
- Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. Mmada: Multimodal large diffusion language models. *arXiv preprint arXiv:2505.15809*, 2025b.
- Junyan Ye, Dongzhi Jiang, Zihao Wang, Leqi Zhu, Zhenghao Hu, Zilong Huang, Jun He, Zhiyuan Yan, Jinghua Yu, Hongsheng Li, et al. Echo-4o: Harnessing the power of gpt-4o synthetic images for improved image generation. *arXiv preprint arXiv:2508.09987*, 2025a.
- Yang Ye, Xianyi He, Zongjian Li, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, and Li Yuan. Imgedit: A unified image editing dataset and benchmark. *arXiv preprint arXiv:2505.20275*, 2025b.
- Lijun Yu, Jose Lezama, Nitesh Bharadwaj Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G Hauptmann, et al. Language model beats diffusion-tokenizer is key to visual generation. In *ICLR*, 2024.
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhua Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, 2024.
- Haotian Zhang, Mingfei Gao, Zhe Gan, Philipp Dufter, Nina Wenzel, Forrest Huang, Dhruti Shah, Xianzhi Du, Bowen Zhang, Yanghao Li, et al. Mm1. 5: Methods, analysis & insights from multimodal llm fine-tuning. *arXiv:2409.20566*, 2024.
- Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024.
- Jialv Zou, Bencheng Liao, Qian Zhang, Wenyu Liu, and Xinggang Wang. Omnimamba: Efficient and unified multimodal understanding and generation via state space models. *arXiv preprint arXiv:2503.08686*, 2025.

## A Data Generation Details

### Prompt. 1: Generate Edit Instructions

{condition\_image} You are a creative assistant specializing in image editing. Your task is to analyze the provided image and generate {num} distinct prompts to modify the original images with different categories. Be concise yet cover detailed information. Specify the modification with fine-grained visual details related with appearance, action, attributes and location.

#### Output Format:

Your entire response must be a single block of text. Separate each distinct prompt with a semicolon (;). Do not use bullet points, numbering, or line breaks.

#### Example of Expected Output (3 edit instructions)

- “Add a vintage bicycle leaning against the brick wall; Change the color of the car to a deep emerald green; Change the plain brick wall behind the subject to a vibrant, colorful graffiti art wall”
- “Remove the briefcase from the man’s hand, and change the color of his tie to a deep crimson red; Make the woman turn her head to the left; Extract the pink top shirt the person wears in the image”
- “Replace the red car in the foreground with a classic vintage motorcycle; Remove the red car parked by the curb; Transfer the image into a dramatic, high-contrast chiaroscuro painting style”

We introduce the construction of **Edit-Align** dataset for the *Edit Instruction Alignment* stage and the **Edit-RL** dataset for the RL stage. Generally, the data source of Edit-RL is a subset of Edit-Align. For each dataset, we collect the condition images, the edit instructions and the description of desired output images. We describe the data generation pipeline in details as below.

### Prompt. 2: Generation of Desired Output Description

You are a precise and objective visual analyst. Your task is to provide a single, standalone, and very short description of an edited image. The description must be 60 words or less.

You will be given:

1. An Original Caption describing the original image.
2. An Editing Prompt describing the change.

#### Instructuons:

1. Be Standalone: Imagine you are describing this final image to someone who has never seen the original. Your description must stand on its own and cannot refer to any previous state or editing process. Avoid using comparative or editing-related words like: ‘changed’, ‘now’, ‘instead of’, ‘no longer’, ‘modified’, ‘edited’, ‘replaced’, ‘added’, ‘removed’, ‘unlike the original’, ‘previously’.
2. Be Faithful: Your description must be a factual account of the final image. Specifically describe the edited element as dictated by the Editing Prompt along with the other key visual elements.
3. Be Concise: Your entire response must be 60 words or less. Focus only on the most essential information.
4. Be Faithful: Do not invent details, objects, or attributes that are not visually present.

Please analyze the following inputs and generate your objective, faithful, and very short description of the final, edited image below.

Original Caption: {original\_caption}

Editing Prompt: {editing\_prompt}

**Sourcing condition images.** For Edit-RL dataset, we employ Qwen-Image Wu et al. (2025a) to generate 5,000 synthetic images based on text prompts collected from COCO val2017 Lin et al. (2014) and segment anything dataset (SAM) Kirillov et al. (2023). Furthermore, we supplement Edit-Align datasets with around 7,000 images sampled from the BLIP-3o SFT image source.

**Preparing edit instructions.** Given each image in Edit-RL, we leverage Qwen2.5-VL-72B-instruct Bai et al. (2025) to construct 10 different edit instructions (see [Prompt. 1](#)). As for image collected from BLIP-3o-SFT in Edit-Align, we generate versatile edit prompts using hand-crafted templates.

### Prompt. 3: Assessing Edit and Description Quality

{condition\_image} You are a meticulous AI Data Quality Analyst. Your task is to score the sample on two criteria: overall description quality and editing quality.

#### 1. Overall Description Quality Score (0-5):

This is a holistic score measuring the quality of the edited image description. A high score requires strength in all of the following areas: (1) The description must accurately reflect the edit requested in the Editing Prompt; (2) It must remain faithful to the unedited parts of the image, without imaginary details, subjective opinions, or hallucinations; (3) It must be a standalone description that does not compare edited elements to their original, unseen state.

#### 2. Editing Quality Score (0-5):

This score evaluates the overall quality of the edit instruction and its outcome based on two factors: (1) the instruction should be plausible given the original image, and (2) it should clearly and unambiguously specify the content to be edited.

Your response must use the following format: “Reasoning: {A brief justification};{Overall description quality score};{Editing quality score}”

Analyze the provided condition image, and the following data. Provide your evaluation in the specified single-line, semicolon-separated format.

Editing Instruction: {edit\_instruction}

Edited Image Caption: {edited\_caption}

### Prompt. 4: Conversation in Edit Instruction Alignment

#### User:

{condition\_image} Analyze the image and the edit instruction: {edit\_instruction}. Write a concise, accurate, and standalone caption describing the final edited image. Focus only on the result of the edit, without referring to or comparing with the original image.

#### Assistant:

{output\_description}

**Generating output descriptions.** As shown in [Prompt. 2](#), we use a Qwen-2.5-72B-instruct [Yang et al. \(2024\)](#) model to generate the description for desired output images for training data in Edit-RL. For additional samples in Edit-Align, we use hand-crafted rules to obtain the edited images descriptions. Next, we filter out edit prompts or captions with poor quality by assessing them with Qwen2.5-VL-72B-instruct [Bai et al. \(2025\)](#) (see [Prompt. 3](#)). Consequently, we collect 17,663 triplets for Edit-Align and 10,568 triplets for Edit-RL.

**Constructing conversations for instruction alignment.** To perform Post-SFT training with *Edit Instruction Alignment*, we assemble the 17,663 data samples of condition images, edit instructions and output descriptions into instructional conversations, as displayed in [Prompt. 4](#)

## B Benchmarks and Evaluation Protocol

**For image understanding**, we include (i) general VQA benchmarks, such as GQA [Hudson and Manning \(2019\)](#) and Seedbench [Li et al. \(2023a\)](#), (ii) knowledge-based benchmarks, such as AI2D [Kembhavi et al. \(2016\)](#), ScienceQA [Lu et al. \(2022\)](#), MMMU [Yue et al. \(2024\)](#), and MathVista [Lu et al. \(2024\)](#), and (iii) hallucination benchmarks, such as POPE [Li et al. \(2023b\)](#). We leverage the lmms-eval toolkit to compute the results for the above benchmarks.

**For text-to-image generation benchmarks**, we report results on GenEval [Ghosh et al. \(2023\)](#) and DPG-bench [Hu et al. \(2024\)](#) to comprehensively evaluate the semantic alignment between a text prompt and the generated images. We report our results using the official evaluation repository of GenEval and DPG-bench, respectively.

**For image editing**, we report results on ImgEdit [Ye et al. \(2025b\)](#) benchmark using the official evaluation repository. All results are evaluated with GPT-4o [Hurst et al. \(2024\)](#).

**Table A** Hyperparameter setup for different training stages of *UniGen-1.5*. Data ratio refers to the ratio of image understanding data, text understanding data, image generation data and image editing data.

Hyperparameters	PT	SFT	Edit-Inst-Align	RL
Learning rate	$1e^{-4}$	$2e^{-5}$	$1e^{-5}$	$3e^{-6}$
LR scheduler	constant	cosine	cosine	cosine
Gradient clip	1.0	1.0	1.0	1.0
Warm-up steps	0	5000	0	0
Training steps	300k	73k	0.5k	1.5k
Batch size	576	128	64	32
Data ratio	2:1:3:-	8:2:3:3	1:-:-:-	-:-:1:1

**Table B** Training data overview of different stages. CC, SA, IMN, T2I-2M, SI, BLIP-3o, GIE stands for CC-3M [Sharma et al. \(2018\)](#) & CC-12M [Changpinyo et al. \(2021\)](#), SAM-11M [Kirillov et al. \(2023\)](#), ImageNet [Ridnik et al. \(2021\)](#), Text-2-Image-2M [jackyhate \(2024\)](#), ShareGPT-4o-Image [Chen et al. \(2025c\)](#), text-to-image data from BLIP-3o-SFT [Chen et al. \(2025a\)](#) and GPT-Image-Edit-1.5M [Wang et al. \(2025c\)](#), respectively.

Stage	Image Gen Data	Image Edit Data	Und Data	Text-only
Pre-training	(CC+SA+IMN) (Recap)	-	(CC+SA+IMN) (Recap)	RefinedWeb
Supervised Fine-tuning	BLIP-3o+T2I-2M+SI	SI+GIE	SF-LLaVA1.5 (Image Mixture) <a href="#">Xu et al. (2025)</a>	RefinedWeb
Edit-Inst-Align	-	-	Edit-Align	-
RL	T2I-R1 <a href="#">Jiang et al. (2025a)</a>	Edit-RL	-	-

## C Training Details

### C.1 Hyper-parameters and Datasets

An overview of training hyper-parameters is shown in [Table A](#). We also list our training datasets for each training stage in [Table B](#).

### C.2 Reward Functions

**Image-text Alignment Model.** We employ DFN5B-CLIP-ViT-H-14 [Fang et al. \(2023\)](#) to model the similarity between a given image and its text prompt. We feed the generated image and the output description into the visual encoder and the text encoder, respectively. Then, we compute the cosine similarity as  $\mathbf{R}_C$  between the image and text embeddings for RL training.

**Human Preference Model.** We use HPSv2 [Wu et al. \(2023\)](#) to evaluate the aesthetic appeal and the alignment between the text description and the generated image. Similarly, we also obtain the cosine similarity between the visual and text features as the reward  $\mathbf{R}_H$ .

**Semantic Consistency Model.** We leverage the UnifiedReward-7B [Wang et al. \(2025b\)](#) model to measure the fine-grained consistency (e.g., objects, attributes and relationship) between the text description and the output image. The model outputs a simple reasoning process followed by a final score ranging from 1-5, which is normalized to 0-1 as the reward  $\mathbf{R}_U$ .

**Outcome Reward Model.** We take advantage of ORM from [Guo et al. \(2025b\)](#) to judge whether the generated images correctly represent the given text description. The model is trained to output ‘Yes’ for aligned image-text pairs and yield ‘No’ otherwise. We compute the probability as the reward based on the model’s first-token distribution. Given  $p_y$  denoting the probability of first token assigned to ‘Yes’ and  $p_n$  indicating the probability for ‘No’, we define the scalar reward as  $\mathbf{R}_O = \frac{p_y}{p_y + p_n}$ .

**Ensemble of Reward Models.** We simply average all reward scores by  $\mathbf{R} = \text{mean}(\mathbf{R}_C + \mathbf{R}_H + \mathbf{R}_U + \mathbf{R}_O)$ . We then perform the advantage computation based on the averaged reward score within each group.

## D More Results

### D.1 Understanding Metrics across Training Stages

Although, both *Edit Instruction Alignment* (Post-SFT) and RL stages are designed to improve image generation and editing, it is also valuable to learn how do they impact *UniGen-1.5*'s image understanding capability. As shown in [Table C](#), there is no performance drop after Post-SFT and RL, suggesting that these stages effectively maintain strong understanding capability while improving image generation and editing.

**Table C** Performance on image understanding benchmarks across different training stages. *Und Avg.* denotes the average score over all the benchmarks.

Model	AI2D	GQA	POPE	MMMU	MathVista	ScienceQA	Seedbench	Und Avg.
SFT	77.6	64.0	89.2	35.7	52.3	85.9	76.4	68.7
Edit Inst. Align	77.6	63.8	88.6	35.7	51.7	86.3	76.5	68.6
RL	77.4	63.7	88.3	35.9	51.9	86.3	76.5	68.6

**Table D** Ablation of condition designs for image editing. We report performance of *UniGen-1.5* after RL with different sequences of condition in image editing task.  $\rightarrow$  denotes the order when concatenating different embeddings. We highlight the default setting of *UniGen-1.5* in light blue.

Condition Input	GenEval	DPG-Bench	ImgEdit
$\mathcal{X}_C^G \rightarrow \mathcal{X}_C^U \rightarrow \mathcal{T}_C$	0.89	87.47	4.05
$\mathcal{X}_C^U \rightarrow \mathcal{X}_C^G \rightarrow \mathcal{T}_C$	0.89	86.97	3.98
$\mathcal{X}_C^U \rightarrow \mathcal{T}_C \rightarrow \mathcal{X}_C^G$	0.89	86.83	4.31

### D.2 Ablation of Condition Design in Image Editing

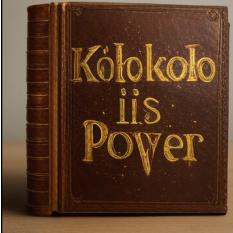
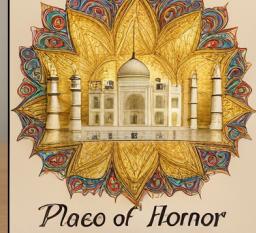
Inspired by prior works [Chen et al. \(2025c\)](#); [Wu et al. \(2025b\)](#), we utilize semantic ( $\mathcal{X}_C^U$ ) and low-level ( $\mathcal{X}_C^G$ ) visual embeddings as well as text embedding  $\mathcal{T}_C$  as conditions for image editing. When constructing the representation of the conditions, the order of these embeddings is very important. We start with  $\mathcal{X}_C^U \rightarrow \mathcal{T}_C$  since this arrangement aligns better with how our model perceives the tokens from visual and text modalities during pre-training, as compared to  $\mathcal{T}_C \rightarrow \mathcal{X}_C^U$ . Then, there are three options for inserting the  $\mathcal{X}_C^G$  as shown in [Table D](#). From the comparison results, we observe that appending  $\mathcal{X}_C^G$  at the end introduces the best overall result. This is because during pre-training, *UniGen-1.5* is optimized with  $\mathcal{X}_C^U \rightarrow \mathcal{T}_C$  for image understanding and  $\mathcal{T}_C \rightarrow \mathcal{X}_C^G$  for image generation so that the design of  $\mathcal{X}_C^U \rightarrow \mathcal{T}_C \rightarrow \mathcal{X}_C^G$  maximally leverages the training momentum.

### D.3 The Impact of Different Reward Models

**Table E Ablation of reward models.** We train *UniGen-1.5* with unified RL using the same training recipe except for the different sets of reward models in each setting. UniR refers to UnifiedReward and DPG stands for the DPG-Bench. We report the overall score for GenEval, DPG-Bench ImgEdit benchmarks. We highlight the default setting of *UniGen-1.5* in light blue.

HPS	Reward model				GenEval	DPG.	ImgEdit
	CLIP	ORM	UniR				
✓					0.76	87.22	4.21
✓	✓				0.85	87.13	4.28
✓	✓	✓			0.88	87.03	4.28
✓	✓	✓	✓		0.89	86.83	4.31

Generally, using the ensemble of multiple vision experts strikes a better trade-off between text-to-image and image editing benchmarks. As shown in [Table E](#), using UnifiedReward-7B slightly benefits the ImgEdit benchmark (row 4 vs. row 3). The removal of ORM leads to a drop of 0.03 in the overall score on GenEval (row 2 vs. row 3). Using HPSv2 alone underperforms the combination of HPSv2 and CLIP-H by 0.09 and 0.07 on GenEval and ImgEdit benchmarks, respectively (row 2 vs. row 1).

<p><b>Text-to-Image</b></p> <table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="text-align: center; padding: 5px;">A beautifully aged antique book ... words "Knowledge is Power" are prominently featured ...</td><td style="text-align: center; padding: 5px;">An ornate representation of the Taj Mahal ... the words "Place of Honor" are inscribed ...</td><td style="text-align: center; padding: 5px;">A creative studio photograph featuring tactile <b>text spelling</b> 'hello' with vibrant, <b>multicolored fur</b>...</td><td style="text-align: center; padding: 5px;">A collection of nine vibrant, assorted speech bubble stickers, ... the acronym 'LOL' in <b>bold</b>, playful lettering ...</td></tr> </table>	A beautifully aged antique book ... words "Knowledge is Power" are prominently featured ...	An ornate representation of the Taj Mahal ... the words "Place of Honor" are inscribed ...	A creative studio photograph featuring tactile <b>text spelling</b> 'hello' with vibrant, <b>multicolored fur</b> ...	A collection of nine vibrant, assorted speech bubble stickers, ... the acronym 'LOL' in <b>bold</b> , playful lettering ...	   
A beautifully aged antique book ... words "Knowledge is Power" are prominently featured ...	An ornate representation of the Taj Mahal ... the words "Place of Honor" are inscribed ...	A creative studio photograph featuring tactile <b>text spelling</b> 'hello' with vibrant, <b>multicolored fur</b> ...	A collection of nine vibrant, assorted speech bubble stickers, ... the acronym 'LOL' in <b>bold</b> , playful lettering ...		
<p><b>Image Editing</b></p> <table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="text-align: center; padding: 5px;">make the cat sit up and look directly at the viewer</td> <td style="text-align: center; padding: 5px;">extract the blue bird from the image</td> </tr> </table>	make the cat sit up and look directly at the viewer	extract the blue bird from the image	   		
make the cat sit up and look directly at the viewer	extract the blue bird from the image				

**Figure A** Failure cases of *UniGen-1.5*.

#### D.4 More Visualization

Failure cases of *UniGen-1.5* in both text-to-image generation and image editing tasks are illustrated in **Figure A**. In the first row, we present the instances where *UniGen-1.5* fails to accurately render text characters, as the light-weight discrete detokenizer struggles to control the fine-grained structural details required for text generation. In the second row, we display two examples with visible identity shifts highlighted by the circle, e.g., the changes in cat's facial fur texture and shape, and the differences in color of the bird's feather. *UniGen-1.5* needs further improvement to address these limitations.