Skywork UniPic 2.0: Building Kontext Model with Online RL for Unified Multimodal Model

Skywork Multimodality Team
multimodal@skywork.ai

Abstract

Recent advances in multimodal models have demonstrated impressive capabilities in unified image generation and editing. However, some prominent open-source models focus primarily on scaling model parameters to achieve better image generation and editing results, rather than focusing on optimal training strategies. In this technical report, we build upon SD3.5-Medium to explore how a 2B-parameter DiT model can achieve impressive image generation and editing capabilities, as well as how to extend it into a unified multimodal model that delivers leading performance in the open-source community. We first modify the architecture of SD3.5-Medium and pre-train it on large-scale, high-quality image generation and editing data to jointly support text-to-image generation and image editing tasks. To enhance instruction following and editing consistency, we propose a novel Progressive Dual-Task Reinforcement strategy(PDTR), which effectively strengthens both tasks in a staged manner. We empirically validate that the reinforcement phases for different tasks are mutually beneficial and do not induce negative interference. After pre-training and reinforcement strategies, we obtain the **UniPic2-SD3.5M-Kontext** model, which surpasses BAGEL (with 7B generation parameters) and Flux-Kontext (with 12B generation parameters) in image generation and editing performance. Furthermore, following the MetaQuery, we connect the UniPic2-SD3.5M-Kontext and Qwen2.5-VL-7B via a connector and perform joint training to launch a unified multimodal model **UniPic2-Metaquery**. UniPic2-Metaquery integrates understanding, generation, and editing, achieving top-tier performance across diverse tasks with a simple and scalable training paradigm. This consistently validates the effectiveness and generalizability of our proposed training paradigm, which we formalize as **Skywork UniPic 2.0**.

1 Introduction

Recent advances in multimodal generative models [2, 14, 59, 65, 30, 63] have demonstrated remarkable capabilities in unifying image generation and editing, offering both high visual fidelity and enhanced interactivity. Models such as BAGEL [14] and FLUX.1-Kontext-Dev [2] have pioneered unified architectures for text-to-image (T2I) synthesis and image editing tasks, advancing the development of intelligent multimodal systems. However, the image generation modules of these models typically contain tens of billions of parameters, leading to prohibitive computational costs and slow inference speeds. Moreover, significant challenges persist in ensuring precise adherence to instructions in image generation and maintaining consistency in editing behaviors. An exclusive emphasis on parameter scaling, without corresponding advancements in training strategies, may prove suboptimal.

To evaluate these conjectures, we adopt the SD3.5-Medium [17] architecture as our foundational model, which comprises a relatively compact 2B-parameter image generation module. We first introduce targeted architectural enhancements, followed by extensive pre-training on large-scale, high-fidelity datasets encompassing both image generation and editing tasks. This approach enables
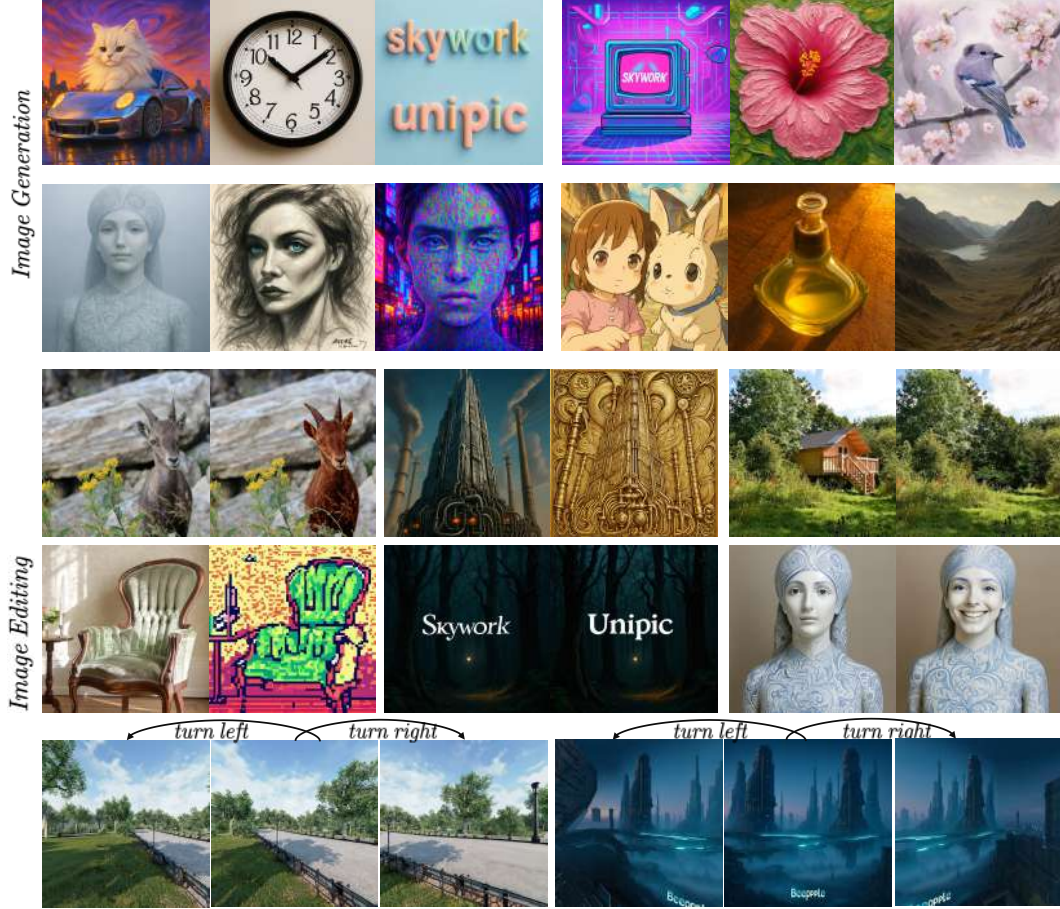
Figure 1: Showcase of the UniPic 2.0 in image generation and editing.

the resulting model to seamlessly support text-to-image (T2I) synthesis and image editing within a shared, integrated framework. Furthermore, by employing progressive resolution training in conjunction with balanced data sampling across a wide spectrum of aspect ratios and resolutions, the proposed model attains native support for dynamic resolution generation during both training and inference, which is pivotal for practical use in real-world scenarios.

Although reinforcement learning [42, 57, 16, 57, 4, 18] has been increasingly adopted to align text-to-image generation with human preferences, its application to jointly optimize generation and editing in a unified model remains unexplored. Due to significant differences in input modalities, output distributions, and evaluation criteria between the two tasks, naively optimizing them jointly often leads to gradient conflicts or performance degradation—a classic multitask learning dilemma where "optimizing one capability compromises the other." To overcome this challenge and enhance both instruction following in image generation and editing consistency, we follow Flow-GRPO [33] to propose Progressive Dual-Task Reinforcement (PDTR), a novel post-training paradigm that leverages Group Relative Policy Optimization (GRPO) [48] for online reinforcement learning. Specifically, PDTR first reinforces the image editing task independently, followed by a second phase focusing on T2I generation. Our experiments demonstrate that the two tasks can be iteratively improved in a synergistic manner without negative interference, effectively resolving the conflict in multi-task RL. In the image editing enhancement phase, we leveraged both the self-trained Skywork-EditReward model and the online GPT-4.1 [37] system as reward evaluators. For the image generation enhancement phase, we employed GenEval [19] as a verifiable reward signal, complemented by classical detector [12, 10] to automatically assess compositional accuracy and instruction adherence. This multi-faceted evaluation strategy reinforces the model's robustness and reliability in handling complex and fine-grained prompts.

Building upon large-scale pre-training and PDTR optimization on a modified SD3.5M architecture, we introduce UniPic2-SD3.5M-Kontext, a unified model that demonstrates strong capabilities in

both image generation and editing. Comprehensive evaluations indicate that UniPic2-SD3.5M-Kontext attains leading performance across multiple benchmarks: it outperforms most existing unified frameworks on GenEval in instruction-following accuracy and establishes new records in editing quality, thereby substantiating its superior generation fidelity, instruction comprehension, and editing consistency.

Motivated by the recent advances in MetaQuery [38], we investigate the integration of UniPic2-SD3.5M-Kontext, which has strong image generation and editing capabilities, with Qwen2.5-VL-7B [1], which excels at multimodal understanding, to construct a unified framework for multimodal understanding, generation, and editing. In the first stage, we freeze the parameters of both the SD3.5-Medium and Qwen2.5-VL models, and pre-train a 24-layer connector module using large-scale text-to-image datasets. In the second stage, we substitute SD3.5-Medium with UniPic2-SD3.5M-Kontext, unfreeze the connector parameters, and jointly fine-tune the connector and Kontext components on high-quality image generation and editing datasets. This yields UniPic2-MetaQuery, a unified multimodal model capable of robust understanding, high-fidelity image generation, and consistent image editing. The proposed training paradigm is simple yet highly scalable, and achieves state-of-the-art results across diverse benchmarks, highlighting its strong modularity and extensibility.

We integrate UniPic2-SD3.5M-Kontext and UniPic2-MetaQuery into Skywork UniPic 2.0, an efficient generative framework for unified multimodal modeling that is designed to enhance speed, efficiency, and generalization, with its capabilities illustrated in Fig. 1. All models and source code are publicly released to foster reproducibility and accelerate progress in efficient multimodal generation. Our results demonstrate that, through deliberate architectural design, targeted pretraining strategies, and coordinated reinforcement learning, lightweight models can outperform substantially larger counterparts in generation fidelity, instruction compliance, and inference efficiency. Skywork UniPic 2.0 thus offers a practical and scalable pathway toward deployable, high-performance multimodal intelligence.

Key contributions of this work are summarized as follows:

- We present **UniPic2-SD3.5M-Kontext**, a lightweight unified model for image generation and editing, enabled by large-scale pretraining, achieving leading performance under high inference speed.

- We introduce **Progressive Dual-Task Reinforcement** (**PDTR**), the first strategy to enable synergistic improvement of image generation and image editing through staged RL, without cross-task interference—significantly boosting instruction following of generation and editing consistency.

- We introduce **UniPic2-Metaquery**, a general and modular paradigm for unified multimodal modeling, which enables end-to-end integration of understanding, generation, and editing through a parameter-efficient connector-based training strategy, achieving SOTA performance and strong generalization across tasks.

2 Related Work

2.1 Image Generation

Diffusion models [46, 45, 40, 8, 44, 3, 39, 29] have emerged as the dominant paradigm for image generation, achieving high-quality visual synthesis conditioned on text prompts. These models are typically instantiated in the latent space of variational autoencoders [23, 55] (VAE), significantly reducing computational and memory overhead. Building upon the Denoising Diffusion Probabilistic Models (DDPMs)[20], recent advances [17, 69, 25, 35] adopt Flow Matching (FM), which replaces explicit diffusion simulation with an ODE-driven continuous transformation from noise to data, effectively reducing the diffusion processes and permitting more direct probability transport paths to improve efficiency. Meanwhile, diffusion transformers [56] (DiTs) have been adopted as the architectural backbone in state-of-the-art frameworks [39, 25, 35, 78, 70], surpassing U-Net backbone [47] in superior scalability and generation quality.
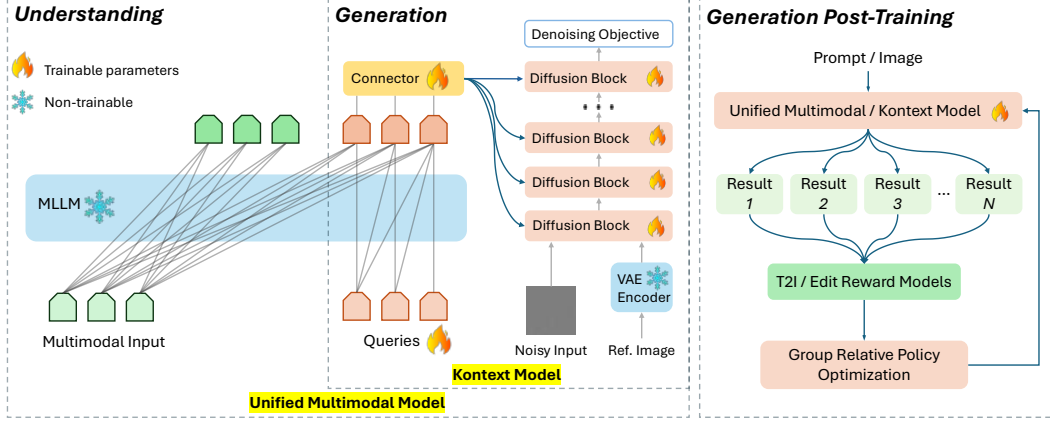
Figure 2: The overall pipeline of UniPic 2.0.

2.2 Image Understanding

Multimodal large language models [76, 32, 31, 26, 77, 60, 28, 68] (MLLMs) combine visual encoders [41, 74, 54] with language models [72, 52, 53] to enable a wide range of image understanding tasks, including visual reasoning, dialogue, and instruction following. The visual encoders are typically pretrained through contrastive vision-language alignment, allowing the model to capture semantic concepts. While MLLMs demonstrated strong general and fine-grained image understanding capabilities by incorporating visual signals into LLMs, their output is primarily text, lacking the ability to produce visual content.

2.3 Unified Models

Efforts to unify multimodal understanding and generation generally follow two paradigms. The first builds native multimodals from scratch [50, 67, 71, 64, 11, 27, 14], and struggles to accommodate both tasks due to the inherently different granularity. To address this, some adopt disentangled visual encoders [64, 11] or mixture-of-experts architectures [27, 14, 49]. The more resource-efficient line of work connects pretrained LLMs and diffusion generators using intermediate ViT features [49, 58, 51, 66] or learnable queries [7, 38]. Among these works, MetaQuery [38]and BLIP3-o [7] build unified frameworks upon frozen multimodal LLMs, effectively transferring the knowledge learned in understanding tasks to visual generation.

3 Method

As shown in Fig. 2, we follow the resource-efficient design of MetaQuery [38] to build UniPic 2.0. The overall architecture mainly comprises an off-the-shelf mutlimodal large language model (MLLM) and a diffusion transformer (DiT), bridged by a set of learnable queries and a connector. To preserve the MLLM's ability in multimodal understanding, we freeze its parameters throughout the training process. We first pre-train UniPic 2.0 Kontext model on text-to-image and image editing datasets, and align the Kontext model and the MLLM model to achieve a unified multimodal model (Sec. 3.1). Then the models are further optimized via GRPO in a post-training process (Sec. 3.2). In our implementation, we choose Qwen2.5VL-7B [1] and SD3.5-Medium [17] as the instantiations of the MLLM and DiT.

3.1 Pre-training

In the pre-training stage, we extend the text-to-image DiT to support image editing, optimize the connector modules to align the MLLM with the DiT, and jointly fine-tune both the connector and DiT to enable unified multimodal modeling.

Kontext Model. Since the original DiT (i.e., SD3.5-Medium) is only trained for text-to-image generation, we retrain this model on image generation and editing datasets. Given that prior stud-

4

ies [30, 14, 65] have emphasized the importance of VAE latents in preserving image structure and texture details, we incorporate the reference images' VAE latents into DiT's self-attention modules to extend the editing capabilities. Specifically, the text instructions and reference images are used as conditional inputs at the same time. The text is encoded by the text encoder to obtain the instruction representation, and the reference image is encoded into a latent variable by VAE and mapped to the context token. The two are concatenated with the noise token of the target image in segments into a single sequence, and the existing positional encoding of the model is used to distinguish the reference image token from the target token. For clarity, the DiT (SD3.5-Medium) retrofitted for image generation and editing is referred to as UniPic2-SD3.5M-Kontext. Besides, to preserve the original generation ability of the DiT, we also include text-to-image data in the training of the Kontext model. In actual training, image generation and editing are performed alternately in batches to simultaneously train both tasks. Resolution design ensures that data covers common aspect ratios such as 1:1, 4:3, 3:2, and 16:9 to avoid distribution bias at a single scale or ratio, thereby improving generalization across different compositions and scenarios.

Align MLLM to Kontext Model.   Following the MetaQuery [38] paradigm, we align the pre-trained Kontext model with Qwen2.5-VL via a transformer-based connector, yielding a unified model that supports understanding, generation, and editing. In our implementation, we first align Qwen2.5-VL with the SD3.5-Medium image generation model, freeze the parameters of both the DiT and VL models, and pre-train the connector and learnable queries on large-scale text-to-image datasets. As emphasized in MetaQuery, a high-capacity connector is critical for effectively bridging the MLLM and DiT; accordingly, our design employs a 24-layer transformer with approximately 1 billion parameters. Empirically, we observe that the DiT module prompted by our learnable queries and connector exhibits even better text-to-image generation performance compared to using its original text encoders, e.g., T5 [43] in SD3.5.

Unified Multimodal Model.   With the connecting modules and the Kontext model obtained in the previous stages, we jointly fine-tune the learnable queries, connector and DiT while freezing the MLLM for unified image understanding, generation and editing. We refer to our unified model as UniPic2-MetaQuery. For image editing tasks, it is noteworthy that we harness the MLLM's image understanding ability to provide semantic-rich conditioning by feeding both text instructions and reference images to the MLLM, in addition to the VAE latents that are passed to the DiT's self-attention modules.

3.2   Post-training

Group Relative Policy Optimization (GRPO).   Given text hidden state h, the flow model samples a group of G images $\{x_0^i\}_{i=1}^G$ and the corresponding trajectories $\{x_T^i, x_{T-1}^i, \ldots, x_0^i\}_{i=1}^G$, following the Flow-GRPO [33]. Within each group, the advantage of the i-th image can be formulated as:

$$A_i^t = \frac{R\left(x_0^i, h\right) - \text{mean}\left(\left\{R\left(x_0^i, h\right)\right\}_{i=1}^G\right)}{\text{std}\left(\left\{R\left(x_0^i, h\right)\right\}_{i=1}^G\right)}, \tag{1}$$

where R denotes the reward model and GRPO employs a clipped objective. The GRPO training objective with a KL penalty term is given:

$$\mathcal{L}_{\text{GRPO}}(\theta) = \mathbb{E}_{h \sim \mathcal{D}, \left\{x_T^i, \ldots, x_0^i\right\}_{i=1}^G \sim \pi_\theta}$$
$$\frac{1}{G}\sum_{i=1}^G \frac{1}{T}\sum_{t=0}^{T-1}\left(\min\left(r_t^i(\theta)A_i^t, \text{clip}\left(r_t^i(\theta), 1-\epsilon, 1+\epsilon\right)A_i^t\right) - \beta D_{KL}\left(\pi_\theta \| \pi_{\text{ref}}\right)\right),$$
$$\tag{2}$$

where $r_t^i(\theta) = \frac{p_\theta\left(x_{t-1}^i | x_t^i, h\right)}{p_{\theta_{\text{old}}}\left(x_{t-1}^i | x_t^i, h\right)}$.

Progressive Dual-Task Reinforcement (PDTR).   Currently, no effective paradigm exists for applying reinforcement learning (RL) to jointly optimize text-to-image (T2I) generation and image editing within a unified model. To address this challenge, we propose the Progressive Dual-Task

5

Reinforcement (PDTR), the first strategy to enable synergistic reinforcement learning of T2I and editing in a shared diffusion model. The core idea of PDTR is to decouple the optimization order of tasks through a staged, progressive training schedule, allowing the model to incrementally improve performance on a new task while preserving its existing capabilities, thereby avoiding cross-task interference. PDTR consists of two sequential stages as follows.

Stage 1: Image Editing Reinforcement. We first conduct independent reinforcement learning for image editing. Given an input image and a textual edit instruction, the goal is to generate outputs that are both semantically consistent and visually natural. Following Flow-GRPO, we adopt GRPO as the RL algorithm and design multi reward signals. We systematically validate the effectiveness of two editing reward signals: our self-trained Skywork-EditReward model and online GPT-4.1 evaluation. After this stage, the model progressively masters precise execution of complex edits while preserving source image structures.

Stage 2: Text-to-Image Reinforcement. Building upon the editing-enhanced model from Stage 1, we proceed to reinforce the T2I generation capability. Using GRPO with verifiable rewards, the model learns to generate images with richer semantic structures. Notably, we introduce compositionality metrics from the GenEval benchmark (e.g., attribute-object binding, spatial relation reasoning) as reward signals, combined with automated detection methods (e.g., object detection and layout analysis) for scalable online feedback. This stage significantly improves instruction adherence and semantic precision.

Critically, we conduct systematic evaluations across multiple training rounds to assess the interaction between the two stages. Experimental results show that reinforcing one task improves performance on the other: for instance, editing-focused RL enhances T2I quality, and vice versa. Moreover, reinforcing T2I after editing does not degrade editing performance. This positive cross-task transfer validates the effectiveness of PDTR: with proper task scheduling and optimization design, generation and editing can undergo non-adversarial co-evolution within a shared model architecture.

Skywork-EditReward vs. Online GPT-4.1. To enable more precise feedback for editing reinforcement learning, we develop Skywork-EditReward, a specialized reward model tailored for image editing quality assessment. Firstly, we leverage our pre-trained UniPic2-SD3.5M-Kontext model to generate 333k editing results. Through carefully designed stable and effective evaluation templates, we employ GPT-4.1 to provide multi-dimensional scoring aligned with human aesthetic standards, evaluating aspects such as instruction following accuracy and image quality. Then, based on Qwen2.5-VL-7B [60], we train Skywork-EditReward using supervised learning with regression loss, enabling the model to predict quality scores that are highly consistent with GPT-4.1 evaluations. In reinforcement learning, Skywork-EditReward captures subtle quality differences that traditional metrics cannot quantify, significantly improving the naturalness and consistency of editing results. This provides an effective technical pathway for future reinforcement learning post-training in image editing tasks.

Furthermore, we provide a scalable VLM evaluation approach. Specifically, following Step1X-Edit [34], we design stable and effective evaluation templates that assess model-generated results from multiple dimensions including editing instruction compliance and image quality, providing online scoring based on GPT-4.1 as reward signals for editing tasks. This multi-source feedback mechanism enables real-time quality assessment during training, ensuring that the model learns to optimize for both semantic accuracy and visual fidelity in editing operations.

4 Experiments

4.1 Setup

Pre-training. For aligning the MLLM and the DiT, we train the connecting modules on 150M image-text pairs for $500K$ steps with a global batch size of $1,024$ and a learning rate of $1e-4$. The training data includes 30M data samples from open-source datasets (CC12M [6], Megalith10M [36], RedCaps [15] and Laion-Aesthetics [13]) and 120M internal synthetic images. For UniPic2-SD3.5M-Kontext, we train the model on 5M image editing data samples and 6M text-to-image data samples for $200K$ steps with a global batch size of 128 and a learning rate of $4e-5$. The image editing

Ours	BAGEL	FLUX.1 Kontext	OmniGen2	Ovis-U1	SD-3.5 M	GPT-4o
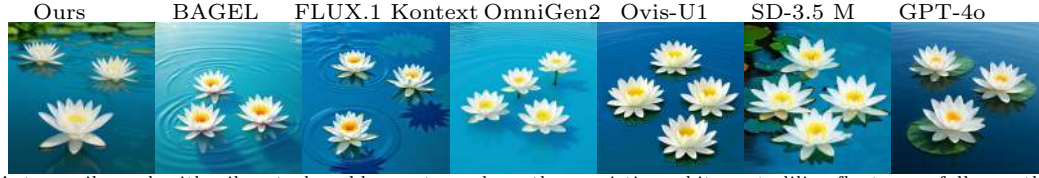
A tranquil pond with vibrant clear blue water, where three pristine white waterlilies float gracefully on the surface. Each waterlily has a perfectly formed shape with delicate petals radiating outwards from a yellow center. The smooth surface of the water reflects the sky above, enhancing the serenity of the scene, as gentle ripples emanate outward from the blossoms.

a detailed sketch of a space shuttle, rendered in the intricate, technical style reminiscent of Leonardo da Vinci's famous drawings. The shuttle is depicted with numerous annotations and measurements, showcasing its complex design and structure. The paper on which it is drawn has an aged, yellowed appearance, adding to the historical feel of the artwork.

An exquisite mahogany chair with intricate carvings stands prominently in the room. The chair features a high back with elaborate designs and a plush red cushion that provides a stark contrast to the dark wood. Its legs are elegantly curved, adding to the overall sophistication of the piece. The texture of the cushion looks soft and inviting, beckoning one to sit and enjoy the comfort it offers.

A striking orca whale is seen gliding through the blue waters of the Nile River, its black and white pattern contrasting vividly against the river's hues. In the background, the ancient silhouette of an Egyptian pyramid looms, its sandy beige stones bathed in the sunlight. The surface of the water ripples lightly as the majestic creature navigates the unusual setting, with palm trees and desert landscapes visible on the riverbanks.

A bustling city street lined with various shops and cafes, leading the eye towards a distinctive building in the distance. The building is notable for its maroon steeple, which houses clocks just beneath a prominent standing platform. The architecture of the building stands out against the urban landscape, hinting at historical significance amidst the modern city life.

A dog on the left and a cat on the right.

A close-up shot of a vibrant sunflower in full bloom, with a honeybee perched on its petals, its delicate wings catching the sunlight.

A steaming bowl of chicken noodle soup on a worn, wooden table, surrounded by soft, golden lighting, with a shallow depth of field, capturing the gentle steam rising from the bowl.

Figure 3: Qualitative comparison of text-to-image generation results.

includes public data collected from UniWorld-V1 [30], OmniGen2 [65], NHR-Edit [24], ShareGPT4o-Image [9] and GPT-Image-Edit-1.5M [62] as well as 450K internal data samples. The text-to-image data mainly comprises 5.6M internal high-quality real images and 400K public data from BLIP3o [7] and ShareGPT4o-Image [9]. It is noteworthy that the same data and hyperparameters are used for training our unified model in the last stage. For all stages in pre-training, we use cosine learning rate annealing and AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 1 \times 10^{-8}$ and weight decay of 0.05.

Post-Training. For post-training, all hyperparameters are kept fixed across different reinforcement learning stages. We use a sampling timestep of $T = 10$, a group size $G = 16$, noise level $a = 0.7$, and image resolution of 512×512. The KL ratio β is set to 0.04. We employ LoRA for parameter-efficient fine-tuning, with rank $r = 32$ and scaling factor $\alpha = 64$. Training uses a learning rate of 3×10^{-4} and a global batch size of 256. We use the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1 \times 10^{-8}$.

4.2 Main results.

Tab. 1 summarizes our main comparative results, evaluating our models against other advanced methods on image generation, image editing, and multimodal understanding. We compare with GPT-4o [22], Emu3 [61], Janus-Pro [11], Blip3-o [7], BAGEL [14], UniWorld-V1 [30], OmniGen2 [65], and Ovis-U1 [59]. The results clearly demonstrate the exceptional performance of our approach across major benchmarks, highlighting its powerful unified capabilities. Notably, we achieve a leading performance with fewer parameters. Detailed comparisons for text-to-image generation, image editing, and multimodal understanding are provided in subsequent subsections. We further present ablation studies in Sec. 4.3 and analyze failure cases in Sec. 4.4.

Text-to-Image Generation. To evaluate the text-to-image generation capabilities of our method, we employ two established benchmarks: GenEval [19], which assesses the ability to generate images with accurate object attributes, counts, positions, and colors; and DPG-Bench [21], which evaluates fine-grained semantic alignment using long and dense prompts.

As shown in Tab. 1, our lightweight model UniPic2-SD3.5M-Kontext (0.89) outperforms both specialized image generation models like FLUX.1-dev (0.67) and SD3-Medium (0.74), and larger unified models including BAGEL (0.88), Blip3-o-8B (0.84), and MetaQuery-XL (0.80). UniPic2-Metaquery archives an overall score of 0.90. On the DPG-Bench benchmark, which contains lengthy and free-form text prompts, UniPic2-SD3.5M-Kontext (84.23) also surpasses specialized generation methods and achieves performance comparable to much larger models, such as BAGEL with 7 billion generation parameters. We further conduct a qualitative comparison between BAGEL, FLUX.1 Kontext, OmniGen2, Ovis-U1, SD3.5-Medium and GPT-4o. As illustrated in Fig. 3, our method delivers faithful results that capture the specified counts, positions, colors, and styles described in the prompts.

Image Editing. Our UniPic2-SD3.5M-Kontext, with only 2B parameters in the generation module, achieve scores of (6.59, 4.00) on the GEdit and ImgEdit editing benchmarks, which outperforms the 12B Flux-Kontext (6.26, 3.52) and the 7B Bagel (6.52, 3.20). Furthermore, as shown in Tab. 1, benefiting from the internal knowledge and reasoning abilities of Qwen2.5-VL [1], UniPic2-Metaquery achieves a comparable score (6.87) to Step1X-Edit (6.97) on the GEdit-En benchmark, and even surpasses a range of specialized models, including FLUX.1 Kontext (3.52) and Step1X-Edit (3.06), as well as unified models such as Ovis-U1 (4.00) and OmniGen2 (3.44) on ImgEdit, achieving 4.03. As illustrated in Fig. 4, our model produces more consistent results in removing, replacing, or adding objects and applying style changes, while preserving the rest of the reference image. 'N.A.' indicates GPT-4o declined to produce results as the specified image editing involved brand or logo modifications may violate its content policies.

Multimodal Understanding. The Qwen2.5-VL branch remains frozen, thereby preserving its multimodal understanding capabilities. As shown in Fig. 5, we evaluate the proposed method on six representative tasks to demonstrate its generalization ability across visual understanding domains: (a) Universal Recognition – Given an image of a bird, the model correctly classifies it as a Blue-throated Bee-eater (Merops superciliosus) and describes distinctive visual traits (e.g., bright blue throat, yellow-green-orange plumage) along with its habitat distribution. (b) Multi-instance Recognition

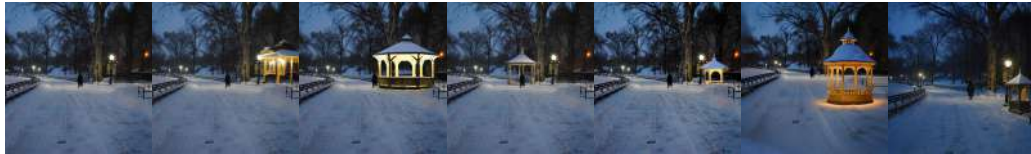| Ref.	Ours	BAGEL	FLUX.1 Kontext	OmniGen2	Ovis-U1	GPT-4o

Change the luxury boat color to red.

Replace the blue bird in the image with a red fox.

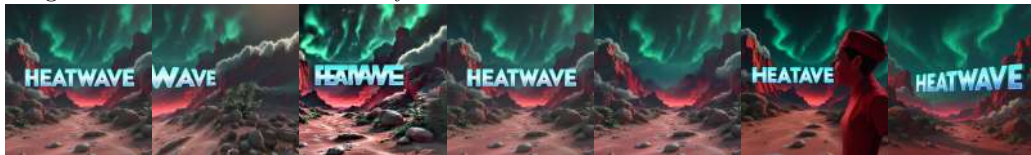Remove the helicopter in the foreground.

Add a small, charming gazebo in the foreground, slightly to the right of the path, near the park benches. The gazebo should match the wintery, serene atmosphere of the scene and be illuminated by a soft, warm light from the nearby street lamps.

Add a painting to the easel.

change the word "dreamshot" to "skywork"

Rotate the camera to the right by 40 degrees.

Change the style of this image to Ice Age.

Figure 4: Qualitative comparison of image editing results.

Table 1: **Comparisons on image understanding, generation, and editing.** 'Generation' and 'Editing' refer to models specialized in image generation and image editing, respectively, while 'Unified' denotes a model that has both understanding and generation capabilities. '×' indicates the model is incapable of performing the task. '†' indicates results obtained using the official OpenAI API. '*' denotes generation results for SD3.5-Medium are reported based on our evaluation.

Type	Model	# Params.	Generation		Editing		Understanding		
			GenEval	DPG	GEdit-En	Imgedit	MMBench	MMMU	MM-Vet
Generation	Emu3-Gen [61]	-	0.54	80.60	×	×	×	×	×
	SDXL [40]	-	0.55	74.56	×	×	×	×	×
	DALL-E 3 [3]	-	0.67	83.50	×	×	×	×	×
	FLUX.1-dev [25]	-	0.67	84.00	×	×	×	×	×
	FLUX.1 Kontext [2]	-	-	-	6.26	3.52	×	×	×
	SD3.5-Medium* [17]	-	0.65	83.86	×	×	×	×	×
	SD3-Medium [17]	-	0.74	84.08	×	×	×	×	×
Editing	AnyEdit [73]	-	×	×	3.21	2.45	×	×	×
	Instructuct-P2P [5]	-	×	×	3.68	1.88	×	×	×
	MagicBrush [75]	-	×	×	4.52	1.90	×	×	×
	Step1X-Edit [34]	-	×	×	6.97	3.06	×	×	×
Unified	Emu3 [61]	8B	0.66	80.60	-	-	58.5	31.6	37.2
	Janus-Pro [11]	7B	0.80	84.19	×	×	75.5	36.3	39.8
	MetaQuery-XL [38]	7B + 1.6B	0.80	82.05	-	-	83.5	58.6	66.6
	UniWorld-V1 [30]	7B + 12B	0.84	81.38	4.85	3.26	83.5	58.6	67.1
	Blip3-o-8B [7]	7B + 1.4B	0.84	81.60	×	×	83.5	58.6	66.6
	OmniGen2 [65]	3B + 4B	0.86	83.57	6.42	3.44	79.1	53.1	61.8
	BAGEL [14]	7B + 7B	0.88	85.07	6.52	3.20	85.0	55.3	67.2
	Ovis-U1 [59]	2.4B + 1.2B	0.89	83.72	6.42	4.00	77.8	51.1	66.7
	GPT-4o [22]	-	0.84	85.15	7.53	4.20	86.0	72.9	76.9
Ours	UniPic2-SD3.5M-Kontext	2B	0.89	84.23	6.59	4.00	×	×	×
	UniPic2-SD3.5M-Kontext †	2B	0.89	84.23	6.74	4.02	×	×	×
	UniPic2-Metaquery	7B + 2B	0.90	83.79	6.87	4.03	83.5	58.6	67.1
	UniPic2-Metaquery †	7B + 2B	0.90	83.79	7.10	4.06	83.5	58.6	67.1

Table 2: Different image conditioning strategies.

#	Model	Image Condition		GEdit-EN
		MLLM	DiT	
1	UniPic2-SD3.5M-Kontext	✗	✓	6.31
2		✓	✗	5.00
3	UniPic2-MetaQuery	✗	✓	6.40
4		✓	✓	**6.90**

Table 3: Freezing or fine-tuning (FT) the connector and DiT for UniPic2-MetaQuery.

#	Connector	DiT	GenEval	GEdit-EN
1	Freeze	FT	0.84	6.75
2	FT	Freeze	0.85	6.46
3	FT	FT	**0.86**	**6.90**

– For a set of attraction images, the model identifies The Great Wall of China, Eiffel Tower, Statue of Liberty, and Terracotta Army, providing their geographic locations and cultural significance. (c) Scene Unserstanding– In a Venice canal scene, the model generates a detailed description capturing global layout (historic buildings flanking the canal), architectural details (ornate facades, balconies, arched windows), illumination (warm golden street lights), and dynamic elements (boats, gondolas, outdoor seating), reflecting ambient atmosphere. (d) Object Grounding with structured output – The model localizes specific cartoon animals (elephant, lion) and outputs their bounding boxes in JSON format, demonstrating the ability to produce machine-readable structured predictions. (e) OCR – The model transcribes stylized text from the image, yielding "Specters of Obsidian Twilight: A Veil Between Realms" with high fidelity. These results highlight the model's capacity to handle heterogeneous input–output formats, perform fine-grained and scene-level reasoning, and maintain high accuracy in both naturalistic and stylized visual contexts.

Figure 5: Qualitative examples illustrating the capabilities of UniPic2-Metaquery across diverse multimodal tasks.

4.3 Ablation Studies

Pre-Training. In Tab. 2, we compare different image conditioning strategies for image editing. The results on GEdit-EN are used as the performance indicator. As illustrated in Tab. 2(#2), the image editing performance decreases drastically when reference images are only fed to the MLLM, highlighting the importance of image conditions in the diffusion process to preserve image structure and texture details. Besides, by comparing results in Tab. 2(#1 & #3), we observe that using MLLM as the text encoder slightly improves the performance on GEdit-EN when reference images are only passed to the DiT. Finally, we obtain the best performance by passing reference images to both the MLLM and the DiT as shown in Tab. 2(#4). In Tab. 3, we study the impact of freezing or fine-tuning the connector and DiT in the training of UniPic2-MetaQuery. As shown in Tab. 3(#3), the best performance is achieved when unlocking both the connector and DiT.

Post-Training. We present the ablation study of different reward signals used in the post-training of UniPic2-SD3.5M-Kontext, as shown in Tab. 4. The results include the individual and combined use of GenEval reward for text-to-image reinforcement, and Skywork-EditReward and online GPT-4.1 evaluation for image editing, providing empirical insights into effective reward design for unified generation and editing models.

We use "w/o GRPO" as the baseline, corresponding to the UniPic2-SD3.5M-Kontext model before reinforcement learning is applied. When reward signals for image editing are introduced separately, the model's editing performance improves significantly. Using Online GPT-4.1 Evaluation as the editing reward boosts GEditBench-EN to 6.55 and ImgEdit to 4.04, demonstrating that multi-dimensional assessment based on vision-language models (VLMs) effectively guides fine-grained editing. Furthermore, when replacing GPT-4.1 with our self-trained Skywork-EditReward model, GEditBench-EN increases to 6.59, indicating its stronger discriminative capability in evaluating editing fidelity. Notably, when only the editing task is reinforced, the T2I generation performance remains stable at 0.83 in terms of GenEval, while DPG-Bench even shows a slight improvement. This confirms the absence of negative transfer between tasks and suggests potential positive cross-task generalization.

For T2I reinforcement, introducing the GenEval reward raises the GenEval score from 0.83 to 0.87, with a modest gain on DPG-Bench, validating the effectiveness of verifiable rewards in generating images with complex semantic structures. Most importantly, when both editing and T2I tasks are reinforced, the model achieves the best overall performance: GenEval reaches 0.89, GEditBench-EN peaks at 6.59, and ImgEdit attains 4.00. This result demonstrates that our proposed PDTR strategy enables positive cross-task synergy rather than mutual interference—confirming the feasibility of joint optimization in a unified architecture. Furthermore, when comparing the two editing reward modules within PDTR, Skywork-EditReward consistently outperforms online GPT-4.1 evaluation across all benchmarks. This highlights its superior stability, higher training efficiency, and elimination of costly API calls, making it a more suitable and scalable reward module for post-training in unified multimodal systems.

In summary, the ablation study fully validates the effectiveness of our reward design and the PDTR strategy: each reward component contributes significantly to its target task, and under progressive training, the two capabilities co-evolve without conflict—providing a reliable and scalable pathway for unified image generation and editing.

Table 4: Ablation Study of Reward Signals on UniPic2-SD3.5M-Kontext.

| Model | RL | | Generation | | Editing | |
	T2I	Edit	GenEval	DPG	GEdit-EN	ImgEdit
w/o GRPO	✗	✗	0.83	83.68	6.31	3.95
w/ GPT-4.1	✗	✓	0.83	84.15	6.55	4.04
w/ EditReward	✗	✓	0.83	83.99	6.59	4.00
w/ GenEval	✓	✗	0.87	83.97	6.40	3.95
w/ GPT-4.1 + GenEval	✓	✓	0.89	84.23	6.54	3.99
w/ EditReward + GenEval	✓	✓	0.89	84.36	6.59	4.00

4.4 Failure Cases

Fig. 6 illustrates several scenarios where our model struggles to precisely follow user instructions: (a) Complex textual rendering – For the fantasy book cover "Legends of the Enchanted Realm", while the style and composition match the description, the text suffers from character substitution and distortion, a common limitation in text rendering. (b) Special object substitution – In the "tiger with a dog reflection" task, although the reflection effect is plausible, semantic consistency and precision in depicting the reflected dog remain imperfect. For image editing, operations such as object extraction, complex text modification and precise attribute conversion are also shortcomings of our model. (a) Object extraction – When asked to extract the sliced steak, the model fails to cleanly isolate the target region, producing blurred artifacts instead of a precise cutout. (b) Complex text modification – Changing the clock numbers to "s-k-y-w-o-r-k-u-n-i-p-i" yields partial success in replacing digits with text but with distortion and incompleteness issues, indicating difficulties in high-fidelity text insertion. These cases highlight persistent challenges for contemporary text-to-image and editing models, particularly in specific count generation, complex text synthesis and simultaneous precise modifications mentioned above. As noted, the performance in such tasks could be improved through data scaling.

5 Conclusions

This work presents Skywork UniPic 2.0, an efficient unified multimodal understanding, generation and editing framework built upon the SD3.5-Medium and Qwen2.5-VL architectures. Through architectural modifications, large-scale pretraining, joint training, and the novel Progressive Dual-Task Reinforcement (PDTR) strategy, our approach achieves synergistic breakout in both image generation and image editing. We first develop UniPic2-SD3.5M-Kontext, a lightweight model with only 2B generation parameters, which surpasses the performance of much larger models while offering significantly faster inference. We then introduce UniPic2-Metaquery, which enables seamless integration of understanding, generation, and editing via a parameter-efficient connector-based training strategy, demonstrating strong extensibility and generalization across diverse multimodal
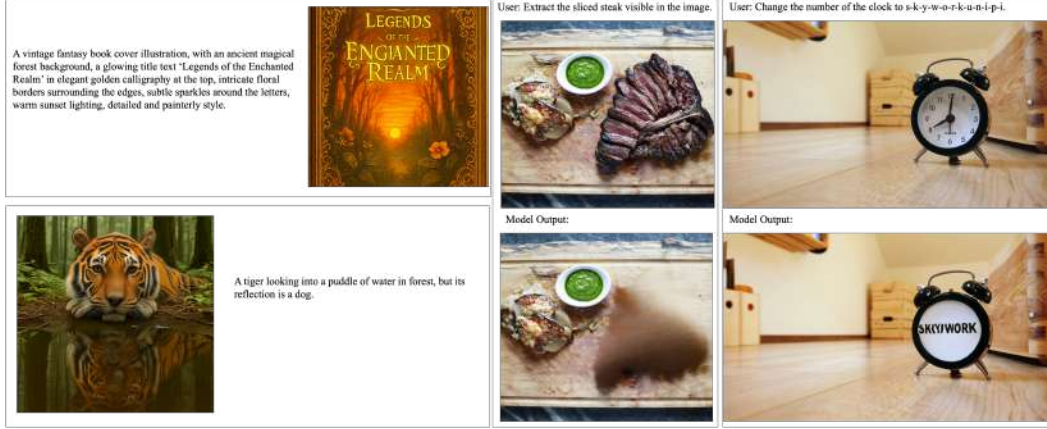
Figure 6: Failure cases on lengthy prompts with intricate semantics, complicated text and detailed object count requirements, often in long and descriptive text.

tasks. Experimental results show that Skywork UniPic 2.0 achieves state-of-the-art performance across multiple benchmarks in instruction following, editing consistency, and generation stability, while maintaining high efficiency and resource friendliness. This work provides a practical, scalable, and reproducible paradigm for advancing efficient, deployable multimodal intelligence.

6 Contributors

Core contributors: Hongyang Wei*, Baixin Xu*, Hongbo Liu*, Cyrus Wu[†], Jie Liu, Yi Peng, Peiyu Wang, Zexiang Liu, Jingwen He, Yang Liu[‡], Xuchen Song[‡], Eric Li[‡]

Contributors: Yidan Xietian, Chuanxin Tang, Zidong Wang, Yichen Wei, Liang Hu, Boyi Jiang, William Li, Ying He, Yahui Zhou

* Equal contribution.
† Project Lead.
‡ Corresponding author.

References

[1] 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. 3, 4, 8

[2] 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. 1, 10

[3] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf, 2(3):8, 2023. 3, 10

[4] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. arXiv preprint arXiv:2305.13301, 2023. 2

[5] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 18392–18402, 2023. 10

[6] 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 Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 3558–3568, 2021. 6

[7] Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, Le Xue, Caiming Xiong, and Ran Xu. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset, 2025. 4, 8, 10

[8] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. PixArt-Sigma: Weak-to-strong training of diffusion transformer for 4K text-to-image generation. arXiv preprint arXiv:2403.04692, 2024. 3

[9] 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, 2025. 8

[10] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155, 2019. 2

[11] 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, 2025. 4, 8, 10

[12] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 1290–1299, 2022. 2

[13] dclure. Laion-aesthetics-umap. https://huggingface.co/datasets/dclure/laion-aesthetics-12m-umap, 2022. 6

[14] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683, 2025. 1, 4, 5, 8, 10

[15] Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. Redcaps: Web-curated image-text data created by the people, for the people. arXiv preprint arXiv:2111.11431, 2021. 6

[16] Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. Raft: Reward ranked finetuning for generative foundation model alignment. arXiv preprint arXiv:2304.06767, 2023. 2

[17] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. 1, 3, 4, 10

[18] Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Reinforcement learning for fine-tuning text-to-image diffusion models. In Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS) 2023. Neural Information Processing Systems Foundation, 2023. 2

[19] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. Advances in Neural Information Processing Systems, 36:52132–52152, 2023. 2, 8

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. Advances in neural information processing systems, 33:6840–6851, 2020. 3

[21] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. arXiv preprint arXiv:2403.05135, 2024. 8

[22] 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. 8, 10

[23] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013. 3

[24] Maksim Kuprashevich, Grigorii Alekseenko, Irina Tolstykh, Georgii Fedorov, Bulat Suleimanov, Vladimir Dokholyan, and Aleksandr Gordeev. Nohumansrequired: Autonomous high-quality image editing triplet mining. Available at SSRN 5381374, 2025. 8

[25] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 3, 10

[26] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326, 2024. 4

[27] Hao Li, Changyao Tian, Jie Shao, Xizhou Zhu, Zhaokai Wang, Jinguo Zhu, Wenhan Dou, Xiaogang Wang, Hongsheng Li, Lewei Lu, et al. Synergen-vl: Towards synergistic image understanding and generation with vision experts and token folding. arXiv preprint arXiv:2412.09604, 2024. 4

[28] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. arXiv preprint arXiv:2403.18814, 2024. 4

[29] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al. Hunyuan-DiT: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. arXiv preprint arXiv:2405.08748, 2024. 3

[30] 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. 1, 5, 8, 10

[31] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 26296–26306, 2024. 4

[32] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 4

[33] Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. arXiv preprint arXiv:2505.05470, 2025. 2, 5

[34] 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, 2025. 6, 10

[35] Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models wfith scalable interpolant transformers. In European Conference on Computer Vision, pages 23–40. Springer, 2024. 3

[36] madebyollin. Megalith-huggingface. https://huggingface.co/datasets/madebyollin/megalith-10m, 2024. 6

[37] OpenAI. Gpt-4-1. https://openai.com/index/gpt-4-1, 2025. 2

[38] Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, Ji Hou, and Saining Xie. Transfer between modalities with metaqueries, 2025. 3, 4, 5, 10

[39] William Peebles and Saining Xie. Scalable diffusion models with transformers. In Proceedings of the IEEE/CVF international conference on computer vision, pages 4195–4205, 2023. 3

[40] 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. 3, 10

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 4

[42] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. Advances in neural information processing systems, 36:53728–53741, 2023. 2

[43] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research, 21(140):1–67, 2020. 5

[44] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125, 1(2):3, 2022. 3

[45] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In International conference on machine learning, pages 8821–8831. Pmlr, 2021. 3

[46] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 3

[47] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18, pages 234–241. Springer, 2015. 3

[48] Zhihong 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. 2

[49] Mustafa Shukor, Enrico Fini, Victor Guilherme Turrisi da Costa, Matthieu Cord, Joshua Susskind, and Alaaeldin El-Nouby. Scaling laws for native multimodal models. arXiv preprint arXiv:2504.07951, 2025. 4

[50] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. arXiv preprint arXiv:2405.09818, 2024. 4

[51] Shengbang Tong, David Fan, Jiachen Zhu, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. Metamorph: Multimodal understanding and generation via instruction tuning. arXiv preprint arXiv:2412.14164, 2024. 4

[52] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023. 4

[53] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288, 2023. 4

[54] 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. 4

[55] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. Advances in neural information processing systems, 30, 2017. 3

[56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017. 3

[57] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8228–8238, 2024. 2

[58] Chunwei Wang, Guansong Lu, Junwei Yang, Runhui Huang, Jianhua Han, Lu Hou, Wei Zhang, and Hang Xu. Illume: Illuminating your llms to see, draw, and self-enhance. arXiv preprint arXiv:2412.06673, 2024. 4

[59] Guo-Hua Wang, Shanshan Zhao, Xinjie Zhang, Liangfu Cao, Pengxin Zhan, Lunhao Duan, Shiyin Lu, Minghao Fu, Xiaohao Chen, Jianshan Zhao, et al. Ovis-u1 technical report. arXiv preprint arXiv:2506.23044, 2025. 1, 8, 10

[60] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191, 2024. 4, 6

[61] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. arXiv preprint arXiv:2409.18869, 2024. 8, 10

[62] 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, 2025. 8

[63] 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, 2025. 1

[64] 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 preprint arXiv:2410.13848, 2024. 4

[65] Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. arXiv preprint arXiv:2506.18871, 2025. 1, 5, 8, 10

[66] 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, 2025. 4

[67] 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, 2024. 4

[68] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. arXiv preprint arXiv:2412.10302, 2024. 4

[69] Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. Sana: Efficient high-resolution image synthesis with linear diffusion transformers, 2024. 3

[70] Enze Xie, Junsong Chen, Yuyang Zhao, Jincheng Yu, Ligeng Zhu, Chengyue Wu, Yujun Lin, Zhekai Zhang, Muyang Li, Junyu Chen, et al. Sana 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer. arXiv preprint arXiv:2501.18427, 2025. 3

[71] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. arXiv preprint arXiv:2408.12528, 2024. 4

[72] 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 preprint arXiv:2412.15115, 2024. 4

[73] Qifan Yu, Wei Chow, Zhongqi Yue, Kaihang Pan, Yang Wu, Xiaoyang Wan, Juncheng Li, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Anyedit: Mastering unified high-quality image editing for any idea. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 26125–26135, 2025. 10

[74] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 11975–11986, 2023. 4

[75] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. Advances in Neural Information Processing Systems, 36:31428–31449, 2023. 10

[76] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 2023. 4

[77] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. 4

[78] Le Zhuo, Ruoyi Du, Han Xiao, Yangguang Li, Dongyang Liu, Rongjie Huang, Wenze Liu, Lirui Zhao, Fu-Yun Wang, Zhanyu Ma, et al. Lumina-next: Making lumina-t2x stronger and faster with next-dit. arXiv preprint arXiv:2406.18583, 2024. 3