Skywork UniPic: Unified Autoregressive Modeling for Visual Understanding and Generation

Skywork Multimodality Team
multimodal@skywork.ai

Abstract

We introduce Skywork UniPic, a unified autoregressive model that natively integrates image understanding, text-to-image generation, and image editing within a single 1.5B-parameter architecture—eliminating task fragmentation and connector modules. Our core insight is captured by a critical question for practical multimodal AI: *Can a compact unified model achieve state-of-the-art performance across understanding, generation, and editing tasks while maintaining accessibility for real-world applications?* Skywork UniPic answers this affirmatively through unprecedented parameter efficiency without capability compromise. Despite the inherent tension between pixel fidelity and semantic richness that plagues existing unified models, our approach establishes new benchmarks: surpassing most unified models on GenEval (0.86) for instruction following, setting a record on DPG-Bench for complex generation (85.5), and leading in editing performance (GEditBench-EN: 5.83; ImgEdit-Bench: 3.49). Remarkably, our approach can generate 1024×1024 resolution images on consumer hardware (RTX 4090). This achievement stems from three key innovations: a decoupled visual encoding strategy that employs Masked Autoencoder (MAR) for generation-focused representation while leveraging SigLIP2 for understanding-focused tasks within a shared autoregressive framework; a progressive training methodology scaling resolution from 256^2 to 1024^2 with dynamic parameter unfreezing; and meticulously curated hundred-million-scale datasets enhanced with task-specific reward models. By demonstrating that high-fidelity multimodal integration need not demand excessive resources, Skywork UniPic establishes a practical paradigm for deployable vision-language intelligence that bridges theoretical capability and real-world utility. Code and models are publicly released at https://huggingface.co/Skywork/Skywork-UniPic-1.5B.

1 Introduction

The rapid evolution of multimodal artificial intelligence has ushered in a paradigm shift toward unified models capable of seamlessly integrating visual perception, generation, and manipulation within a single architectural framework. Recent demonstrations like GPT-4o's[28] viral "Ghiblification" capability—transforming ordinary photographs into Studio Ghibli-style artworks through natural language interaction—highlight the transformative potential of such systems. These applications reveal a critical limitation in some conventional approaches[30, 39, 43]: fragmented pipelines where separate models handle understanding, generation, and editing. Such isolation impedes cross-modal synergy, inflates deployment costs through redundant model stacks, and disrupts natural multi-turn creative workflows. Consequently, the development of natively unified architectures that intrinsically support end-to-end visual comprehension, text-to-image synthesis, and instruction-driven editing has emerged as a pivotal challenge in multimodal artificial intelligence.

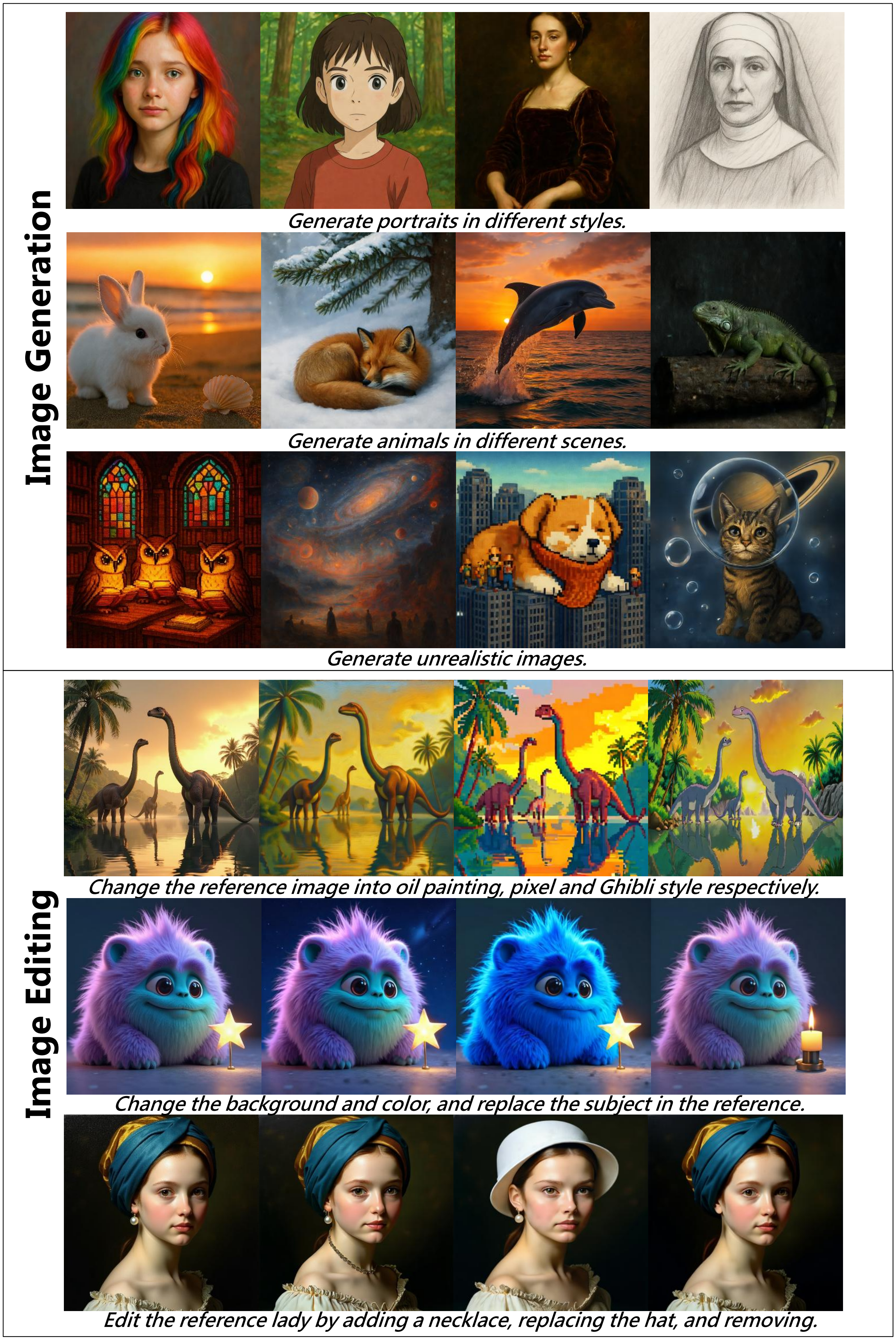


Figure 1: Showcases of our model's performance on editing and generation tasks.

Existing solutions face fundamental constraints. Methods using VQGAN/VAE representations[42, 51, 53, 61] prioritize pixel-level reconstruction at the expense of semantic richness, inherently weakening visual understanding capabilities. Alternative approaches[30, 39, 43] concatenate pre-trained vision-language and text-to-image models through ad-hoc connectors followed by joint fine-tuning. This piecemeal design fails to achieve deep integration, resulting in performance trade-offs between generation fidelity, editing precision, and reasoning depth. Moreover, prevailing efforts often resort to extreme scaling—deploying multi-billion-parameter models trained on trillion-scale datasets—raising serious concerns about computational efficiency and practical deployability. A crucial question thus remains unanswered: *Can a compact unified model achieve state-of-the-art performance across understanding, generation, and editing tasks while maintaining accessibility for real-world applications?*

We address this challenge through Skywork UniPic, a unified autoregressive model that redefines the efficiency frontier for multimodal integration. The model is built upon a single large language model (LLM), primarily consisting of a MAR encoder, a SigLIP2 encoder, a LLM backbone, and a MAR decoder. Our architecture fundamentally departs from quantization-based or connector-dependent paradigms by embedding image understanding, text-to-image generation, and image editing within a single end-to-end trainable framework. The core innovation lies in a decoupled visual encoding strategy: we employ the Masked Autoencoder (MAR[22]) as the backbone for generation-focused representation, optimized for high-fidelity synthesis, while integrating SigLIP2[41] for understanding-focused tasks. Critically, both encoders operate within a shared autoregressive objective, enabling bidirectional knowledge transfer where generation enhances visual detail modeling for understanding, and semantic understanding guides coherent editing. This design preserves architectural simplicity while resolving the longstanding tension between pixel-level fidelity and semantic comprehension.

Skywork UniPic achieves unprecedented parameter efficiency without sacrificing capability. With a compact 1.5B language backbone, it establishes new state-of-the-art results across critical benchmarks: surpassing contemporary models on GenEval[12] (0.86) for instruction following, achieving 85.5 on DPG-Bench[16] for complex generation, and leading among unified models on editing tasks (5.83 on GEditBench-EN[26], 3.49 on ImgEdit-Bench[56]), the visualization results as show in Figure1. Remarkably, it accomplishes this with approximately one-tenth the parameters of comparable systems like BAGEL[9] (14B) or UniWorld-V1[24] (19B), while generating 1024×1024 images on consumer-grade hardware (RTX 4090). This efficiency stems from three synergistic innovations: meticulous curation of a hundred-million-scale high-quality dataset emphasizing task balance and semantic diversity; novel text-to-image reward model trained via Group Relative Policy Optimization (GRPO[36]) and editing reward model to align with human preferences; and a progressive training curriculum that incrementally introduces task complexity while scaling resolution from 256^2 to 1024^2.

Our work makes three key contributions to unified multimodal modeling. First, we introduce the natively unified autoregressive architecture that intrinsically supports joint visual understanding, generation, and editing without requiring separate models or connectors, maintaining accessibility for real-world applications. Second, we resolve the semantic-fidelity dichotomy through a decoupled visual encoding strategy that optimizes representation pathways for distinct task requirements while maintaining cross-task synergy. Third, we demonstrate that rigorous data curation, targeted reward modeling, and progressive training enable state-of-the-art performance at unprecedented scale efficiency—proving that high-quality multimodal integration need not demand excessive computational resources. Through extensive validation across eight benchmarks and comprehensive ablation studies, we establish Skywork UniPic as a practical foundation for deployable multimodal systems. By open-sourcing the model weights, training code, and technical documentation, we aim to accelerate the adoption of efficient unified vision-language models in resource-constrained environments, bridging the gap between theoretical capability and real-world applicability.

2 Related Work

2.1 Semantic Encoders

Vision-language models (VLMs) have emerged as the cornerstone of multimodal understanding by introducing semantic encoders that effectively inject visual signals into language models, thereby endowing them with robust image comprehension capabilities. Among these, CLIP[34] established a foundational paradigm through its contrastive learning framework that aligns image and text embeddings in a shared space, enabling remarkable zero-shot classification and retrieval performance. Building on this foundation, SigLIP[57] refined the training methodology with a sigmoid-based loss function that eliminated temperature parameter dependencies, enabling more stable scaling. SigLIP2[41] integrates multiple advanced techniques—including captioning-based pretraining, self-supervised losses, and online data curation—to produce even richer semantic representations while preserving input aspect ratios across multiple resolutions. These progressive advancements in visual semantic encoding have significantly enhanced zero-shot classification, image-text retrieval, and transfer learning capabilities, establishing crucial foundations for unified models that must balance deep semantic understanding with high-fidelity generation—a balance that remains challenging for existing approaches due to the inherent tension between pixel-level detail preservation and conceptual representation.

2.2 Image Generation

Image generation has evolved through a succession of architectural paradigms, transitioning from the adversarial dynamics of Generative Adversarial Networks (GANs[13]) to the probabilistic foundations of diffusion models[15, 35, 37] and the sequential modeling capabilities of autoregressive approaches[38, 40]. Among these, Latent Diffusion Models (LDMs)[35] have gained widespread adoption by significantly improving computational efficiency—achieved through processing in a compressed latent space—while preserving fine-grained image details. Complementary to this, Vector Quantized models such as VQGAN[10] integrate discrete latent representations with adversarial training to enhance perceptual quality and reconstruction fidelity. However, the inherent quantization process often leads to a loss of semantic expressiveness due to discretization-induced information bottlenecks. Recently, the Masked Autoencoder (MAR) has emerged as a compelling framework for unified image modeling. By operating directly in pixel space and leveraging autoregressive generation through masked prediction, MAR circumvents the need for learned latent codes while maintaining strong generative performance, offering a promising alternative that balances architectural simplicity with effective representation learning.

2.3 Image Editing

Image editing has witnessed rapid advancements in recent years, especially in the realm of natural language-guided manipulation, where user instructions drive precise and semantically meaningful modifications to visual content. A pivotal contribution in this direction is Step1X-Edit [26], which established a comprehensive framework encompassing a scalable data generation pipeline across diverse editing tasks, alongside the introduction of GEdit-Bench—a standardized benchmark for evaluating instruction-following performance in image editing. Building on this progress, IC-Edit [59] introduced a context-aware generation mechanism leveraging diffusion Transformers, enabling zero-shot instruction following without architectural modifications, thereby demonstrating strong generalization across unseen editing commands. Concurrently, UltraEdit [60] addressed data scarcity and diversity limitations by constructing a large-scale, automatically curated dataset, significantly improving the quality and fine-grained controllability of language-driven edits. Despite these notable advances, a critical limitation persists: most current systems operate in isolation from broader vision-language understanding and generative modeling pipelines. They typically rely on specialized, standalone architectures that are decoupled from models responsible for image description, reasoning, or synthesis. This architectural fragmentation impedes the realization of seamless, multi-turn interactive workflows, in which users naturally alternate between describing scenes, issuing edit commands, and iteratively refining visual outputs through continuous natural language dialogue.

2.4 Unified Models

Unified multimodal models seek to combine visual understanding and generation within a single architecture, enabling seamless interaction between vision and language. These models can be grouped into four main paradigms: harmonization, decoupling, hybrid, and connector approaches.

The harmonization approach, exemplified by Harmon[49], uses a shared MAR[22] encoder for both tasks. It builds on findings that MAR representations achieve strong performance in linear probing and respond precisely to visual concepts, suggesting their potential for understanding beyond generation. In contrast, the decoupling strategy, as seen in Janus[46] and Janus-Pro[8], separates visual encoding into distinct pathways. This design addresses conflicting granularity demands while maintaining a unified Transformer backbone, improving flexibility and task specialization.

Hybrid models like Show-o[53] integrate autoregressive and discrete diffusion mechanisms. This allows support for diverse tasks such as visual question answering, text-to-image generation, and mixed-modal synthesis. Connector-based methods, such as MetaQueries[30], use learnable queries to bridge autoregressive LLMs and diffusion models, enabling modular integration without architectural changes.

Recent advances include BAGEL[9], a large-scale decoder-only model trained on trillions of multimodal tokens. It demonstrates emergent capabilities in multimodal reasoning, including image manipulation, future frame prediction, and 3D navigation. OmniGen2[47] introduces separate decoding paths for text and images, along with a decoupled image tokenizer. This design preserves text generation quality while supporting in-context editing and achieving state-of-the-art performance on the OmniContext benchmark.

UniFluid[11] adopts a unified autoregressive framework with continuous visual tokens, showing that generation and understanding can mutually benefit under balanced training. Other notable models include BLIP3-o[6], which generates CLIP-space features via diffusion Transformers, and OpenUni[48], a lightweight open-source baseline. Despite significant progress, developing a compact unified model that achieves state-of-the-art performance across understanding, generation, and editing tasks while remaining practical for real-world deployment remains a critical challenge.

3 Method

We introduce Skywork UniPic, a unified multimodal generation and understanding model derived from the Harmon framework as shown in Figure 2.The architecture primarily consists of a MAR[22] encoder, a SigLIP2[41] encoder, a large language model (LLM), and a MAR decoder, each contributing to different stages of multimodal processing.

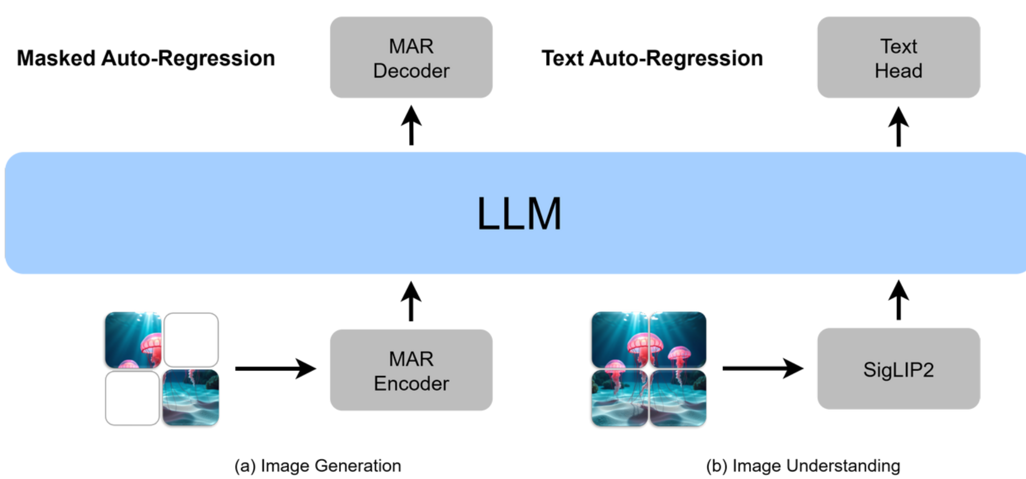


Figure 2: The overall framework of Skywork UniPic. (a) Image generation is achieved through a masked auto-regressive process using the MAR model. (b) Image understanding is performed using a SigLIP2 encoder to extract rich visual features, which are subsequently passed to an LLM for autoregressive text generation. They share a single LLM to promote consistent instruction-following and enable knowledge transfer between generation and understanding tasks

3.1 Model Architecture

In contrast to the original Harmon[49] framework, which employs a shared MAR encoder for both generation and understanding tasks, we found that this weight-sharing strategy may lead to task interference. To mitigate this issue, we decouple the encoders and adopt a SigLIP2 encoder for the understanding pathway. This design modification aims to enhance task-specific representation learning. We choose Qwen2.5-1.5B-Instruct[33] as LLM. To enhance the instruction-following capabilities of both the understanding and generation tasks, we employ a shared LLM across both branches.To bridge the dimensional gap between the encoder outputs and the LLM's hidden representations, we employ separate two-layer MLP on both the understanding and generation. These MLPs serve to align the feature dimensions, facilitating effective integration with the shared LLM.

Generation. We use the same VAE[35] from Harmon framework, to encode input images into a compact latent representation. The VAE remains frozen during the entire training process,this design choice prevents the degradation of low-level visual features and enables more stable convergence during multimodal training.

The Masked Auto-Regressive (MAR) model is specifically designed for image generation, leveraging masked prediction to enhance both generation quality and controllability. To retain these advantages in our framework, we adopt MAR-Huge as both the image encoder and decoder. This model contains approximately 1B parameters, with 20 layers on both the encoding and decoding sides, a hidden size of 1280, and 16 attention heads.We scale up the image resolution from 256×256 to 512×512 to enable finer-grained generation and enhance the model's capacity to capture high-frequency visual details. This modification also broadens the applicability of MAR to high-resolution image synthesis tasks.

Understanding. We adopt SigLIP2 as the image encoder on the understanding side, owing to its strong cross-modal alignment capabilities and superior performance on a wide range of vision-language benchmarks. Its lightweight architecture and efficient representation learning make it well-suited for integrating visual semantics into multimodal systems. To further enhance the model's visual understanding ability, we continue training based on the SigLIP2-so400m-patch16-512, which offers a solid foundation for cross-modal representation learning.

3.2 Training

The training objective combines task-specific losses:

- **Image Generation** (Diffusion loss):

$$\mathcal{L}_{\text{Gen}} = \mathbb{E}_{\varepsilon,t}\left[\left\|\varepsilon - \varepsilon_\theta\left(x_t \mid t, z\right)\right\|^2\right]$$

- **Image Understanding** (Cross-entropy loss):

$$\mathcal{L}_{\text{Und}} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{C} y_{n,i}\log(\hat{y}_{n,i})$$

These are integrated into the multi-task objective during joint training:

$$\mathcal{L}_{\text{Total}} = \lambda_{\text{Gen}}\mathcal{L}_{\text{Gen}} + \lambda_{\text{Und}}\mathcal{L}_{\text{Und}}$$

where λ coefficients evolve through training stages.

Training Pipeline. We employ a four-stage curriculum spanning hundred-million-scale pretraining and million-scale SFT samples, structured as a single, cohesive process. The pipeline begins with **Stage 1: MAR Pretraining (PT)**, which establishes foundational generation capabilities through dedicated training of the MAR encoder-decoder module, with particular emphasis on face reconstruction and complex object synthesis. This is followed by **Stage 2: MAR-LLM Alignment**, where MAR outputs are projected to the LLM embedding space while maintaining frozen LLM parameters, utilizing cosine annealing scheduling to accelerate the convergence of the projection layers. Subsequently, **Stage 3: Joint Optimization (CT)** unfreezes the LLM for cross-modal tuning under the multi-task objective $\mathcal{L}_{\text{Total}}$ with loss weights $\lambda_{\text{Gen}} = 1$ and $\lambda_{\text{Und}} = 0.01$, yielding 12-15% improvements in instruction adherence metrics. The process concludes with **Stage 4: Supervised Fine-tuning (SFT)**, which refines the unified model using reward-filtered samples (quality threshold > 0.9) and incorporates the full $\mathcal{L}_{\text{Total}}$ objective with editing loss components to polish the final task performance.

Table 1: UniPic Training Configuration Across Learning Stages

Hyperparameter	**PT**	**Alignment**	**CT**	**SFT**
Learning rate	5.0×10^{-5}	1×10^{-5}	1.0×10^{-5}	5×10^{-6}
LR scheduler	Constant	Cosine decay	Cosine decay	Cosine decay
Weight decay	0.0	0.02	0.02	0.02
Gradient clipping	1.0	1.0	1.0	1.0
Optimizer	AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 10^{-15}$)			
Loss weights (U:G:E)*	0:1:0	–	0.01:1:1	0.01:1:1
Warmup ratio	0.05	0.05	0.01	0.01
Training epochs	800	3	3	2
EMA decay	0.9999	–	0.9999	0.995
Training samples	130M	130M	130M	3M
Image resolution (width × height)				
Generation	512×512	1024×1024	1024×1024	1024×1024
Understanding	256×256	512×512	512×512	512×512

Training Configuration. Hyperparameter shown in Table1. The training stack utilizes `bf16` mixed-precision and is optimized with DeepSpeed ZeRO-3[2]. We use a global batch size of 4096 for pre-training (PT) and 512 for supervised fine-tuning (SFT). The model architecture consists of an 800M parameter Multi-Agent Reasoning (MAR) module combined with a 1.5B parameter Large Language Model (LLM) backbone.

3.3 Data Quality Assurance

Reward Modeling. To ensure data quality, we develop two specialized reward models (RMs) based on the Qwen-VL architecture [4]: **Skywork-ImgReward**, designed to assess visual quality, and an **Skywork-EditReward**, aimed at evaluating the accuracy of image editing operations.

The image quality RM, Skywork-ImgReward, is trained using the Group Relative Policy Optimization (GRPO) algorithm [36], leveraging a custom-designed *paired ranking reward function*. This function combines a learned pairwise ranking score (r_θ) with a format-based score (r_{format}), formulated as:

$$r(x, y_i) = \underbrace{r_\theta(x, y_i)}_{\text{pairwise ranking}} + \underbrace{r_{\text{format}}(x, y_i)}_{\text{format reward}} \tag{1}$$

The training data for Skywork-ImgReward integrates several public datasets—Pick-a-Pic [19], ImageRewardDB [54], and HPSv2 [50]—augmented with our own curated samples focused on human figure quality.

The Skywork-EditReward is trained via supervised fine-tuning on a collection of high-quality editing datasets, including HumanEdit [3], UltraEdit [60], and SuperEdit-40K [21], enabling it to capture fine-grained instruction alignment and semantic correctness in image edits.

Data Curation. To ensure the quality of our training data, we implement a rigorous filtration pipeline. We first apply a threshold-based filter, discarding any samples with a reward score below 0.9. Subsequently, to address instances of weak instruction-following, we employ a multi-check mechanism that leverages vqascore[25] as an additional quality heuristic. An analysis of the rejected samples reveals four primary failure modes: instruction-alignment deviations, where the output content mismatches the prompt; visual artifacts, such as blurring, structural distortions, or detail loss; semantic inconsistencies, involving logical contradictions in the generated content; and edit non-compliance, characterized by inadequate responses to editing instructions or suboptimal quality. This curated dataset, demonstrating our pipeline's effectiveness across diverse categories like human figures, animals, and text rendering, ensures a high degree of data homogeneity and ultimately enhances the model's generalization capabilities.

*U:G:E = Understanding:Generation:Editing loss weights

4 Main Results

4.1 Evaluation Setup

To comprehensively assess the unified capabilities of Skywork UniPic, we adopt a multi-faceted evaluation strategy encompassing image understanding, text-to-image generation, and image editing across established benchmarks.

Benchmarks For text-to-image generation, we evaluate on GenEval[12] which measures compositional understanding and object-focused alignment, and DPG-Bench[16] which assesses complex instruction following and long prompt adherence capabilities. These benchmarks capture both fine-grained compositional reasoning and general-purpose generation quality.

Image editing capabilities are assessed using GEdit-Bench-EN[26] and ImgEdit-Bench[56] as primary evaluation suites. Built from authentic user requests covering diverse editing scenarios, these benchmarks closely mirror practical editing needs and provide comprehensive coverage of instruction-based image modification tasks including object addition/removal, style transfer, and attribute modification.

Evaluation Protocol All image generation tasks employ 64 sampling steps with 1024×1024 resolution outputs and classifier-free guidance scale of 7.5 for optimal quality-diversity trade-off. Performance assessment utilizes official benchmark scripts and automated evaluation metrics, with all scores reported from single evaluation runs without reranking or multi-sampling to ensure reproducible results.

Baselines We compare against several categories of state-of-the-art models. Unified models include OmniGen2[47], Janus-Pro[8], BAGEL[9], and UniWorld-V1[24]. Specialized generation models comprise FLUX.1-dev[55], SD3-medium[1], and DALL-E 3[27]. For editing capabilities, we compare against Step1X-Edit[26], IC-Edit[59], and AnyEdit[17].

Despite utilizing only 1.5B activated parameters, Skywork UniPic demonstrates competitive or superior performance compared to significantly larger unified models (typically 7B+ parameters), highlighting the effectiveness of our architectural design and training methodology.The corresponding performance metrics for each task are summarized in Figure 3.

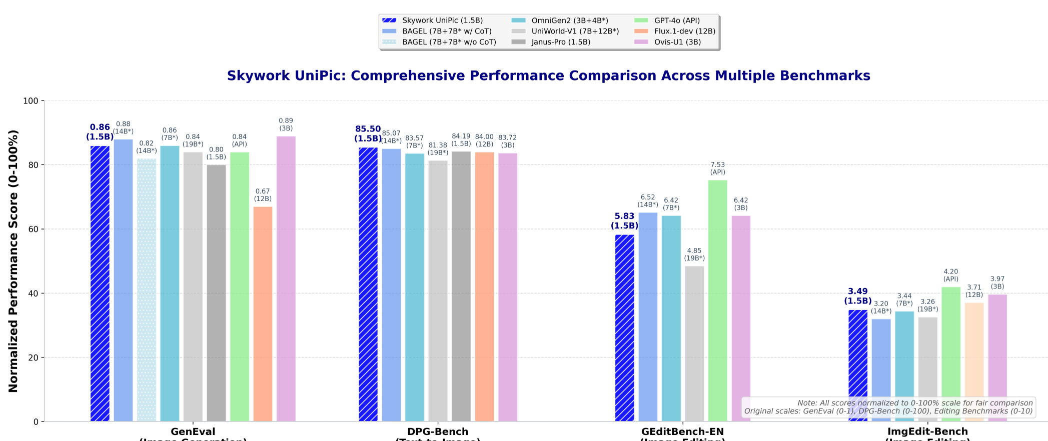


Figure 3: Performance comparison across multiple benchmarks. Skywork UniPic demonstrates competitive performance across understanding, generation, editing, and in-context tasks while maintaining exceptional parameter efficiency with only 1.5B activated parameters.

4.2 Text-to-Image Generation

We assess Skywork UniPic's T2I generation capabilities on two standard benchmarks: GenEval and DPG-Bench, which evaluate compositional understanding and long prompt following respectively. Our model demonstrates highly competitive performance, particularly when considering its resource efficiency.

Evaluation on GenEval. As shown in Table 2, Skywork UniPic achieves an overall score of 0.86 on GenEval, demonstrating strong compositional understanding across diverse generation tasks. The model performs particularly well on single object generation (98.44%) and two object composition (92.42%), while maintaining solid performance on color understanding (90.69%) and spatial positioning (89.00%). Counting tasks (74.06%) and color attribution (72.25%) present greater challenges, consistent with observations across unified models in the literature.

Table 2: Comprehensive comparison on GenEval benchmark. † denotes using rewritten prompts.

Model	Single	Two	Count	Color	Position	Attr	Overall
Diffusion Models							
SDv2.1[35]	0.98	0.51	0.44	0.85	0.07	0.17	0.50
SDXL[31]	0.98	0.74	0.39	0.85	0.15	0.23	0.55
IF-XL	0.97	0.74	0.66	0.81	0.13	0.35	0.61
LUMINA-Next[62]	0.92	0.46	0.48	0.70	0.09	0.13	0.46
SD3-medium[1]	0.99	0.94	0.72	0.89	0.33	0.60	0.74
FLUX.1-dev[55]	0.99	0.81	0.79	0.74	0.20	0.47	0.67
NOVA[18]	0.99	0.91	0.62	0.85	0.33	0.56	0.71
Autoregressive Models							
TokenFlow-XL[32]	0.95	0.60	0.41	0.81	0.16	0.24	0.55
Janus[46]	0.97	0.68	0.30	0.84	0.46	0.42	0.61
Janus Pro[8]	0.99	0.89	0.59	0.90	0.79	0.66	0.80
Emu3-Gen[45]	0.99	0.81	0.42	0.80	0.49	0.45	0.66
Show-o[53]	0.98	0.80	0.66	0.84	0.31	0.50	0.68
Unified Models							
OmniGen[52]	0.98	0.84	0.66	0.74	0.40	0.43	0.68
OmniGen2[47]	1.00	0.95	0.64	0.88	0.55	0.76	0.80
OmniGen2†	0.99	0.96	0.74	0.98	0.71	0.75	0.86
MetaQuery-XL†[30]	-	-	-	-	-	-	0.80
BLIP3-o† 4B[6]	-	-	-	-	-	-	0.81
BLIP3-o† 8B	-	-	-	-	-	-	0.84
BAGEL[9]	0.99	0.94	0.81	0.88	0.64	0.63	0.82
BAGEL†	0.98	0.95	0.84	0.95	0.78	0.77	0.88
UniWorld-V1[24]	0.99	0.93	0.79	0.89	0.49	0.70	0.80
UniWorld-V1†	0.98	0.93	0.81	0.89	0.74	0.71	0.84
Ovis-U1[44]	0.98	0.98	0.90	0.92	0.79	0.75	0.89
Proprietary Models							
GPT-4o[29]	0.99	0.92	0.85	0.92	0.75	0.61	0.84
Skywork UniPic	**0.98**	**0.92**	**0.74**	**0.91**	**0.89**	**0.72**	**0.86**

Evaluation on DPG-Bench. On DPG-Bench, Skywork UniPic achieves an overall score of 85.5, demonstrating competitive performance in long prompt following and complex scene understanding. Table 3 shows detailed comparisons across different evaluation categories, where our model maintains consistent performance across global coherence, entity recognition, attribute understanding, and relational reasoning.

4.3 Image Editing

Image editing represents a core strength of Skywork UniPic's unified architecture. We evaluate the model's editing capabilities on GEdit-Bench, which assesses instruction-based image modification across diverse scenarios.

Evaluation on GEdit-Bench. As demonstrated in Table 4, Skywork UniPic achieves strong performance with an overall score of 5.83, placing it among the top-tier unified models. The model demonstrates particular strength in semantic consistency (SC) with a score of 6.72, indicating robust instruction-following capabilities. While perceptual quality (PQ) scores show room for improvement

Table 3: Comprehensive comparison on DPG-Bench across different semantic categories.

Model	Global	Entity	Attribute	Relation	Other	Overall
Diffusion Models						
LUMINA-Next[62]	82.82	88.65	86.44	80.53	81.82	74.63
SDXL[31]	83.27	82.43	80.91	86.76	80.41	74.65
PlayGroundv2.5[20]	83.06	82.59	81.20	84.08	83.50	75.47
Hunyuan-DiT[23]	84.59	80.59	88.01	74.36	86.41	78.87
PixArt-$\sum$[7]	86.89	82.89	88.94	86.59	87.68	80.54
DALLE3[27]	90.97	89.61	88.39	90.58	89.83	83.50
SD3-medium[1]	87.90	91.01	88.83	80.70	88.68	84.08
FLUX.1-dev[55]	82.10	89.50	88.70	91.10	89.40	84.00
Autoregressive Models						
Show-o[53]	79.33	75.44	78.02	84.45	60.80	67.27
EMU3[45]	85.21	86.68	86.84	90.22	83.15	80.60
TokenFlow-XL[32]	78.72	79.22	81.29	85.22	71.20	73.38
Janus[46]	82.33	87.38	87.70	85.46	86.41	79.68
Janus Pro[8]	86.90	88.90	89.40	89.32	89.48	84.19
BLIP3-o 4B[6]	-	-	-	-	-	79.36
BLIP3-o 8B	-	-	-	-	-	81.60
Unified Models						
OmniGen[52]	87.90	88.97	88.47	87.95	83.56	81.16
OmniGen2[47]	88.81	88.83	90.18	89.37	90.27	83.57
BAGEL[9]	88.94	90.37	91.29	90.82	88.67	85.07
UniWorld-V1[24]	83.64	88.39	88.44	89.27	87.22	81.38
Ovis-U1[44]	82.37	90.08	88.68	93.35	85.20	83.72
Skywork UniPic	**89.65**	**87.78**	**90.84**	**91.89**	**91.95**	**85.50**

at 6.18, the model's ability to make precise, localized edits while preserving unmodified regions demonstrates the effectiveness of our unified architecture.

Evaluation on ImgEdit-Bench. To further validate our model's editing capabilities across diverse scenarios, we evaluate Skywork UniPic on ImgEdit-Bench, a comprehensive benchmark that covers nine distinct editing categories including Add, Adjust, Extract, Replace, Remove, Background, Style, Hybrid, and Action tasks. As demonstrated in Table 5, Skywork UniPic achieves competitive performance with an overall score of 3.49, establishing itself among the leading unified models in comprehensive image editing evaluation.

The results reveal several noteworthy patterns in our model's performance across different editing categories. Skywork UniPic demonstrates particularly strong capabilities in Action editing (4.04) and Style modification (4.76), benefiting from our progressive training methodology that incorporates diverse video-derived editing data. The model also shows solid performance in Background editing (3.77) and Replace operations (4.31), indicating robust understanding of spatial relationships and object substitution. While performance in Extract (2.06) and Adjust (3.51) tasks shows room for improvement, the overall balanced performance across all categories demonstrates the versatility of our unified architecture.

Compared to other unified models, Skywork UniPic outperforms OmniGen (2.96) and approaches the performance of leading specialized editing models like ICEdit (3.05) and Step1X-Edit (3.06), while maintaining the advantage of unified architecture that handles multiple modalities within a single framework. The superior performance of BAGEL (3.20) and UniWorld-V1 (3.26) on certain categories demonstrates the benefits of larger parameter scales and extensive training data, yet our model achieves comparable results with significantly fewer parameters, highlighting the efficiency of our architectural design and training strategy.

Table 4: Comprehensive comparison on GEdit-Bench-EN showing semantic consistency (SC) and perceptual quality (PQ) metrics. Higher scores are better for all metrics.

Model	**SC ↑**	**PQ ↑**	**Overall ↑**
Proprietary Models			
Gemini-2.0-flash[14]	6.73	6.61	6.32
GPT-4o[29]	7.85	7.62	7.53
Specialized Editing Models			
Instruct-Pix2Pix[5]	3.58	5.49	3.68
MagicBrush[58]	4.68	5.66	4.52
AnyEdit[17]	3.18	5.82	3.21
ICEdit[59]	5.11	6.85	4.84
Step1X-Edit[26]	7.09	6.76	6.70
Unified Models			
OmniGen[52]	5.96	5.89	5.06
OmniGen2[47]	7.16	6.77	6.41
BAGEL[9]	7.36	6.83	6.52
UniWorld-V1[24]	4.93	7.43	4.85
Ovis-U1[44]	-	-	6.42
Skywork UniPic	**6.72**	**6.18**	**5.83**

Table 5: Comprehensive comparison on ImgEdit-Bench showing performance across nine editing categories. Higher scores are better for all metrics.

Model	**Add**	**Adjust**	**Extract**	**Replace**	**Remove**	**Background**	**Style**	**Hybrid**	**Action**	**Overall**
Proprietary Models										
GPT-4o[29]	4.61	4.33	2.90	4.35	3.66	4.57	4.93	3.96	4.89	4.20
Specialized Editing Models										
MagicBrush[58]	2.84	1.58	1.51	1.97	1.58	1.75	2.38	1.62	1.22	1.90
Instruct-Pix2Pix[5]	2.45	1.83	1.44	2.01	1.50	1.44	3.55	1.20	1.46	1.88
AnyEdit[17]	3.18	2.95	1.88	2.47	2.23	2.24	2.85	1.56	2.65	2.45
UltraEdit[60]	3.44	2.81	2.13	2.96	1.45	2.83	3.76	1.91	2.98	2.70
Step1X-Edit[26]	3.88	3.14	1.76	3.40	2.41	3.16	4.63	2.64	2.52	3.06
ICEdit[59]	3.58	3.39	1.73	3.15	2.93	3.08	3.84	2.04	3.68	3.05
Unified Models										
OmniGen[52]	3.47	3.04	1.71	2.94	2.43	3.21	4.19	2.24	3.38	2.96
OmniGen2[47]	3.57	3.06	1.77	3.74	3.20	3.57	4.81	2.52	4.68	3.44
BAGEL[9]	3.56	3.31	1.70	3.30	2.62	3.24	4.49	2.38	4.17	3.20
UniWorld-V1[24]	3.82	3.64	2.27	3.47	3.24	2.99	4.21	2.96	2.74	3.26
Ovis-U1[44]	4.13	3.62	2.98	4.45	4.06	4.22	4.69	3.45	4.61	4.00
Skywork UniPic	**3.66**	**3.51**	**2.06**	**4.31**	**2.77**	**3.77**	**4.76**	**2.56**	**4.04**	**3.49**

4.4 Qualitative Results

Text-to-Image Generation Quality. Figure 4 presents qualitative comparisons between Skywork UniPic and both open-source and closed-source models on text-to-image generation tasks. Our model demonstrates competitive visual quality and strong adherence to textual prompts across diverse scenarios, from simple object generation to complex scene composition. The results show that despite its compact size, Skywork UniPic produces images with comparable fidelity and semantic accuracy to much larger specialized models.

Image Editing Capabilities. Figure 5 showcases Skywork UniPic's image editing performance compared to state-of-the-art editing models. The model demonstrates precise instruction following across various editing scenarios, including object addition/removal, style transfer, attribute modification, and complex compositional changes. Notably, the model maintains consistency in unedited regions while accurately implementing the requested modifications, highlighting the benefits of our unified architecture approach.

5 Limitation and Discussion

Limitations. While Skywork UniPic demonstrates strong performance across generation and editing tasks, certain limitations remain. As shown in Figure 6, the model occasionally struggles with complex or ambiguous instructions in text-to-image generation, leading to suboptimal instruction adherence. In the image editing setting, we observe cases where the model fails to respond to the editing prompt, resulting in incomplete or missing modifications. These limitations suggest that further refinement is needed in instruction grounding and editability robustness, particularly under challenging or compositional scenarios.

Emergence of Capabilities. Similar to observations in BAGEL[9], UniPic exhibits a clear, staged emergence of capabilities. Notably, text-to-image (T2I) generation appears in Stage 2 and is progressively refined, whereas more complex image editing capabilities emerge significantly later, only becoming evident in Stage 3 and Stage 4. This staggered manifestation reflects the inherent complexity of image editing, which demands a more sophisticated integration of visual-semantic alignment, conditional reasoning, and structural preservation compared to direct generation. In our work, we define an ability as emergent if it is absent in earlier training stages but materializes in later ones. This qualitative shift, often termed a phase transition, is consistent with our observation that UniPic's loss curves do not explicitly signal the onset of new capabilities, reinforcing the notion that training loss is an insufficient proxy for evaluating true model abilities.

To investigate this phenomenon, we evaluate model checkpoints from each stage by tracking average scores on standard VLM benchmarks (as a proxy for multimodal understanding), the GenEval[12] score (for generation), and the G-Edit[26]. Our experiments consistently show that **editing capabilities emerge later than generation capabilities**, a pattern that holds even when scaling image resolution from 256×256 up to 1024×1024. Interestingly, each resolution increase induces a temporary performance dip followed by a rapid recovery that surpasses the previous capability plateau, suggesting higher resolutions unlock higher performance ceilings. Furthermore, we find no clear evidence that simply scaling understanding-centric data (e.g., image-text matching) directly enhances these generative or editing capabilities. This observation underscores the necessity of generation-specific training regimes for mastering complex, instruction-following tasks.

6 Conclusion and Future Work

We present Skywork UniPic, a unified autoregressive model that achieves competitive performance across image understanding, text-to-image generation, and image editing tasks within a single 1.5B parameter architecture. Through decoupled visual encoding that employs MAR for generation and SigLIP2 for understanding, our model resolves the fundamental tension between pixel-level fidelity and semantic understanding that has constrained previous unified approaches.

The model demonstrates strong empirical results: 0.86 on GenEval for compositional generation, 85.5 on DPG-Bench for complex instruction following, and 5.83 on GEdit-Bench for image editing, while maintaining efficient deployment on consumer hardware. Our comprehensive data construction pipelines from video sources address critical data scarcity in editing and in-context generation, and the introduced OmniContext benchmark provides standardized evaluation for in-context capabilities.

Key technical contributions include the decoupled encoding strategy that preserves both generation quality and understanding capabilities, systematic data construction methodologies for high-quality training corpus creation, and the integration of reflection mechanisms for iterative image generation improvement. The work demonstrates that unified multimodal models can achieve both strong performance and practical efficiency, challenging assumptions about the necessity of massive parameter scaling for capable multimodal systems.

Future work will address current limitations including multilingual performance disparities, fine-grained human body modification capabilities, and input quality sensitivity. The open-source release of model weights, training code, and datasets aims to facilitate further research in parameter-efficient unified multimodal architectures.

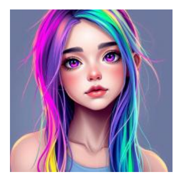
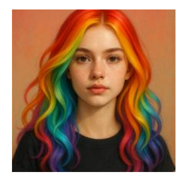
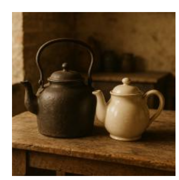
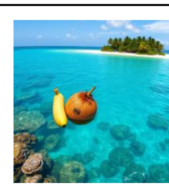
Prompt	Ours	Bagel	Kontext	GPT-4o
At sunset on the beach, a fluffy white rabbit pricks up its ears, curiously gazing at a scallop.				
A glossy-coated golden retriever stands on the park lawn beside a life-sized penguin statue.				
A textured green iguana sits still on a worn log against a shadowed wall.				
A giant pixel corgi sleeps on city skyscrapers. Tiny construction workers are knitting a huge scarf around its neck. The art looks like old Nintendo game sprites.				
Digital portrait of a girl with rainbow hair.				
A pencil sketch portrait of a nun.				
A vintage kitchen scene: a cast iron kettle and ceramic teapot resting on a rough-hewn wooden table.				
A banana and a hairy coconut float in crystal-clear turquoise waters above vibrant coral reefs near a palm-fringed island.				

Figure 4: Qualitative comparison of text-to-image generation results. Skywork UniPic produces high-quality images that accurately reflect textual prompts while maintaining competitive visual fidelity compared to both open-source and proprietary models.

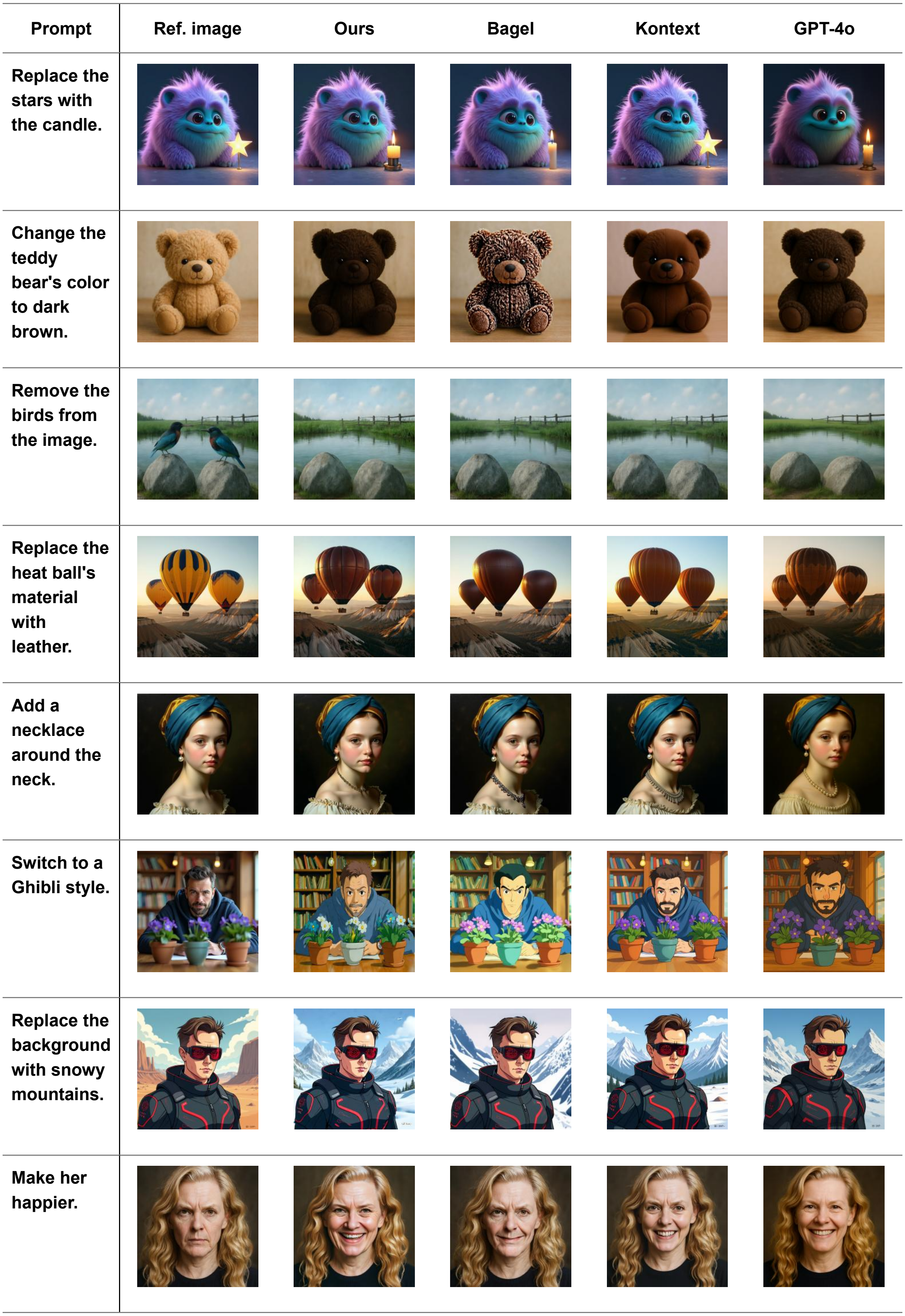


Figure 5: Qualitative comparison of image editing results. Skywork UniPic successfully handles diverse editing instructions while preserving image quality and maintaining consistency in unmodified regions, demonstrating the effectiveness of our unified approach.

Figure 6: Failure cases.

7 Contributions

Core Contributors: Peiyu Wang, Yi Peng, Yimeng Gan, Liang Hu, Eric Li*, Xuchen Song*

Contributors: Tianyidan Xie, Xiaokun Wang, Yichen Wei, Chuanxin Tang, Bo Zhu, Changshi Li, Hongyang Wei, Yang Liu, Yahui Zhou

* Project Lead

References

[1] Stability AI. Stable diffusion 3 medium: Multimodal diffusion transformer for photorealistic text-to-image generation. https://stability.ai/news/stable-diffusion-3-medium, 2025. 8, 9, 10

[2] Reza Yazdani Aminabadi, Samyam Rajbhandari, Minjia Zhang, Ammar Ahmad Awan, Cheng Li, Du Li, Elton Zheng, Jeff Rasley, Shaden Smith, Olatunji Ruwase, and Yuxiong He. Deepspeed inference: Enabling efficient inference of transformer models at unprecedented scale, 2022. 7

[3] Jinbin Bai, Wei Chow, Ling Yang, Xiangtai Li, Juncheng Li, Hanwang Zhang, and Shuicheng Yan. Humanedit: A high-quality human-rewarded dataset for instruction-based image editing, 2025. 7

[4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. 7

[5] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions, 2023. 11

[6] 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. 5, 9, 10

[7] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis, 2023. 10

[8] 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. 5, 8, 9, 10

[9] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining, 2025. 3, 5, 8, 9, 10, 11, 12

[10] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis, 2021. 4

[11] 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. 5

[12] Dhruba Ghosh, Hanna Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment, 2023. 3, 8, 12

[13] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. 4

[14] Google. Gemini 2.0 flash. https://developers.googleblog.com/en/experiment-with-gemini-20-flash-native-image-generation, 2025. 11

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. 4

[16] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment, 2024. 3, 8

[17] Houcheng Jiang, Junfeng Fang, Ningyu Zhang, Guojun Ma, Mingyang Wan, Xiang Wang, Xiangnan He, and Tat seng Chua. Anyedit: Edit any knowledge encoded in language models, 2025. 8, 11

[18] Nan Jiang, Chengxiao Wang, Kevin Liu, Xiangzhe Xu, Lin Tan, Xiangyu Zhang, and Petr Babkin. Nova: Generative language models for assembly code with hierarchical attention and contrastive learning, 2025. 9

[19] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation, 2023. 7

[20] Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation, 2024. 10

[21] Ming Li, Xin Gu, Fan Chen, Xiaoying Xing, Longyin Wen, Chen Chen, and Sijie Zhu. Superedit: Rectifying and facilitating supervision for instruction-based image editing, 2025. 7

[22] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization, 2024. 3, 4, 5

[23] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, Dayou Chen, Jiajun He, Jiahao Li, Wenyue Li, Chen Zhang, Rongwei Quan, Jianxiang Lu, Jiabin Huang, Xiaoyan Yuan, Xiaoxiao Zheng, Yixuan Li, Jihong Zhang, Chao Zhang, Meng Chen, Jie Liu, Zheng Fang, Weiyan Wang, Jinbao Xue, Yangyu Tao, Jianchen Zhu, Kai Liu, Sihuan Lin, Yifu Sun, Yun Li, Dongdong Wang, Mingtao Chen, Zhichao Hu, Xiao Xiao, Yan Chen, Yuhong Liu, Wei Liu, Di Wang, Yong Yang, Jie Jiang, and Qinglin Lu. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding, 2024. 10

[24] Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, Yatian Pang, and Li Yuan. Uniworld-v1: High-resolution semantic encoders for unified visual understanding and generation, 2025. 3, 8, 9, 10, 11

[25] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation, 2024. 7

[26] Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, Guopeng Li, Yuang Peng, Quan Sun, Jingwei Wu, Yan Cai, Zheng Ge, Ranchen Ming, Lei Xia, Xianfang Zeng, Yibo Zhu, Binxing Jiao, Xiangyu Zhang, Gang Yu, and Daxin Jiang. Step1x-edit: A practical framework for general image editing, 2025. 3, 4, 8, 11, 12

[27] OpenAI. Improving image generation with better captions. `https://cdn.openai.com/papers/dall-e-3.pdf`. 8, 10

[28] OpenAI. Gpt-4o system card, 2024. 1

[29] OpenAI. Gpt-4o. https://openai.com/index/introducing-4o-image-generation, 2025. 9, 11

[30] 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. 1, 3, 5, 9

[31] 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, 2023. 9, 10

[32] 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, 2024. 9, 10

[33] Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. 6

[34] 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. 3

[35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 4, 6, 9

[36] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. 3, 7

[37] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations, 2021. 4

[38] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. arXiv preprint arXiv:2406.06525, 2024. 4

[39] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners, 2024. 1, 3

[40] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. 2024. 4

[41] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. 3, 4, 5

[42] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning, 2018. 3

[43] 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, 2024. 1, 3

[44] Guo-Hua Wang, Shanshan Zhao, Xinjie Zhang, Liangfu Cao, Pengxin Zhan, Lunhao Duan, Shiyin Lu, Minghao Fu, Xiaohao Chen, Jianshan Zhao, Yang Li, and Qing-Guo Chen. Ovis-u1 technical report, 2025. 9, 10, 11

[45] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Tao Li, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Tiejun Huang, and Zhongyuan Wang. Emu3: Next-token prediction is all you need, 2024. 9, 10

[46] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. Janus: Decoupling visual encoding for unified multimodal understanding and generation, 2024. 5, 9, 10

[47] Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, Ze Liu, Ziyi Xia, Chaofan Li, Haoge Deng, Jiahao Wang, Kun Luo, Bo Zhang, Defu Lian, Xinlong Wang, Zhongyuan Wang, Tiejun Huang, and Zheng Liu. Omnigen2: Exploration to advanced multimodal generation, 2025. 5, 8, 9, 10, 11

[48] Size Wu, Zhonghua Wu, Zerui Gong, Qingyi Tao, Sheng Jin, Qinyue Li, Wei Li, and Chen Change Loy. Openuni: A simple baseline for unified multimodal understanding and generation. arXiv preprint arXiv:2505.23661, 2025. 5

[49] 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, 2025. 4, 5

[50] 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, 2023. 7

[51] Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, Song Han, and Yao Lu. Vila-u: a unified foundation model integrating visual understanding and generation, 2025. 3

[52] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation, 2024. 9, 10, 11

[53] 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, 2024. 3, 5, 9, 10

[54] 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, 2023. 7

[55] Chenglin Yang, Celong Liu, Xueqing Deng, Dongwon Kim, Xing Mei, Xiaohui Shen, and Liang-Chieh Chen. 1.58-bit flux, 2024. 8, 9, 10

[56] 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, 2025. 3, 8

[57] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023. 4

[58] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing, 2024. 11

[59] Zechuan Zhang, Ji Xie, Yu Lu, Zongxin Yang, and Yi Yang. In-context edit: Enabling instructional image editing with in-context generation in large scale diffusion transformer, 2025. 4, 8, 11

[60] Haozhe Zhao, Xiaojian Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. Ultraedit: Instruction-based fine-grained image editing at scale, 2024. 4, 7, 11

[61] 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, 2024. 3

[62] Le Zhuo, Ruoyi Du, Han Xiao, Yangguang Li, Dongyang Liu, Rongjie Huang, Wenze Liu, Lirui Zhao, Fu-Yun Wang, Zhanyu Ma, Xu Luo, Zehan Wang, Kaipeng Zhang, Xiangyang Zhu, Si Liu, Xiangyu Yue, Dingning Liu, Wanli Ouyang, Ziwei Liu, Yu Qiao, Hongsheng Li, and Peng Gao. Lumina-next: Making lumina-t2x stronger and faster with next-dit, 2024. 9, 10