OmniGen2: Towards Instruction-Aligned Multimodal Generation

Chenyuan Wu[1,2*], Jiahao Wang[1,3*], Pengfei Zheng[1,2*], Ruiran Yan[1,2*], Shitao Xiao[1*§], Xin Luo[1,2*],
Yueze Wang[1*], Wanli Li[1,4†], Xiyan Jiang[1,4†], Yexin Liu[1†], Junjie Zhou[1], Ziyi Xia[1],
Ze Liu[1,2], Chaofan Li[1], Haoge Deng[1,3], Kun Luo[1,3], Bo Zhang[4], Jiajun Zhang[3],
Dong Liu[2], Defu Lian[2], Xinlong Wang[1], Zhongyuan Wang[1], Tiejun Huang[1], Zheng Liu[1‡§]
[1] Beijing Academy of Artificial Intelligence, [2] University of Science and Technology of China,
[3] Institute of Automation, Chinese Academy of Sciences, [4] Zhejiang University
{stxiao, yzwang}@baai.ac.cn zhengliu1026@gmail.com

Abstract

*Multimodal generative models can process instructions in various modalities and demonstrate outstanding performance across a wide range of image generation tasks. However, their robustness in complex real-world scenarios remains limited due to insufficient generalized instruction alignment. We introduce **OmniGen2**, a unified multimodal generator designed to follow complex, fine-grained instructions. Our core contribution is a two-stage design that first builds a strong, world-knowledge-grounded foundation model and then aligns it using a progressive, multitask instruction tuning strategy. The foundation model features a streamlined architecture with decoupled decoding for versatile multimodal generation and a novel positional encoding scheme to improve learning efficiency. We ground this model in real-world knowledge using large-scale data construction pipelines. Building on this foundation, we propose a progressive, reinforcement-based alignment process. This phase carefully schedules training tasks and reward signals to foster cross-task knowledge transfer, significantly improving the model's instruction-following capabilities. Our models demonstrate competitive performance on standard benchmarks and our dedicated in-context generation benchmark, **OmniContext**. We have released our models, code, benchmark, and training datasets at https://github.com/VectorSpaceLab/OmniGen2.*

1. Introduction

Multimodal image generation has witnessed rapid progress in the past year. Generative models such as GPT-Image-1 [29], Flux [34], Qwen-Image [81], Seedream [66] and NanoBanana [25] demonstrate increasingly broad and versatile capabilities such as stylization, text rendering, in-context generation, and knowledge-driven generation, marking a significant step toward general-purpose generative intelligence. Given these diverse capabilities, it is essential to perform multimodal instruction alignment to ensure the controllability, semantic consistency and overall generation quality. This involves two key challenges. The first is constructing a robust and versatile foundation model. The model must be endowed with nascent instruction following capabilities and broad world knowledge, while strictly avoiding over-training. The second is aligning the foundation model. This alignment requires explicit and comprehensive reward signals and must ensure consistency across all generation tasks.

Existing open-source generation models are somewhat deficient as initial base model. Some models are specialized and can not handle tasks beyond their training scope while some are over-optimized towards specific aesthetic preferences, resulting in a severe loss of plasticity. Meanwhile, instruction alignment requires the foundation model to possess a deep understanding of multimodal semantic and task intent. Therefore, we aim to first establish a base model which is simple, versatile and flexible.

The versatility of a generative model depends a lot on the scale and diversity of its training data. Existing datasets are typically generated either via inpainting models [80], which have limited task coverage, or by retrieving images from the Internet [84], which results in limited data volume and low image quality. To address this, we develop extensive data construction pipelines that leverage video sources, providing richer in-context and editing examples.

A strong architecture is equally crucial. OmniGen2 achieves unified multimodal generation by conditioning the diffusion transformer on the variable-length hidden states of a Vision Language Model (VLM), effectively leverag-

*Co-first Authors and Listed in Alphabetical Order
†Core Contributor
‡Corresponding Author
§Project Lead

Figure 1. Overview of versatile abilities of OmniGen2.

ing the VLM's deep semantic understanding and rich world knowledge. To support diverse tasks, we introduce Omni-RoPE, which enhances spatial consistency across images and improves cross-image localization. While conceptually similar to MetaQuery [57], OmniGen2 differs in execution: rather than using fixed-length query tokens, it conditions the diffusion decoder on the VLM's variable-length hidden states, avoiding information bottlenecks. During the majority of the training process, the VLM is frozen, and optimization focuses on image rendering, making OmniGen2 more efficient than models like Mogao [40] and BAGEL [17].

Once the foundation model is obtained via pre-training and fine-tuning, we apply progressive reinforcement learning to facilitate instruction alignment of OmniGen2. Specifically, we adopt Group Relative Policy Optimization (GRPO) [67] and divide the instruction alignment process into multiple sequential stages. At each stage, appropriate reward is selected to optimize the alignment for specific target tasks. The training sequence is carefully organized to promote inter-task transfer.

Our extensive evaluation of OmniGen2 reveals its competitive performance across various task domains, including text-to-image (T2I) generation, image editing (Edit), and in-context generation (IC). Instruction alignment consistently and significantly improves the performance of the base model across all these tasks. Notably, for the in-context generation task, there is currently a lack of well-established public benchmark to systematically assess and compare the key capabilities of different models. To mitigate this limitation, we introduce the **OmniContext** benchmark, comprising eight task categories specifically designed to evaluate consistency across individuals, objects, and scenes.

In summary, our main contributions are as follows:

- We introduce OmniGen2, a powerful multimodal generative model that is systematically instruction aligned. The model demonstrates superior instruction following ability, context consistency, and generation quality across diverse task scenarios.
- We establish an end-to-end pipeline to achieve comprehensive instruction alignment. This pipeline spans from strong foundation model construction to dedicated multi-task alignment.
- We present the OmniContext benchmark, a rigorous suite designed to evaluate in-context image generation, provid-

ing the community with a standardized tool to measure progress in this key area.

2. Dataset Construction

To build a versatile and robust base model, high-quality, large-scale, and diverse training data is essential. Our goal is to equip the base model with broad world knowledge and the ability to generate varied content. A key challenge is the lack of high-quality public data for complex tasks like detailed image editing and consistent in-context generation. Therefore, we not only gather existing datasets but also build our own scalable pipelines to create the high-quality data needed to fill these gaps.

Foundational Knowledge and General Capabilities. To build a strong foundation, we first curate a massive dataset covering both multimodal understanding and text-to-image (T2I) generation. For the former, we adopt LLaVA-OneVision [35]. For T2I, we collect approximately 140M open-source image-text pairs from diverse datasets [9–12, 37, 38, 55, 65, 72], supplemented with 10M proprietary images annotated by Qwen2.5-VL-72B [3].

Advanced Capabilities for Editing and In-Context Tasks. To address the data scarcity in more complex domains, we develop dedicated construction pipelines. For image editing, we integrate public datasets such as SEED-Data-Edit [22] and OmniEdit [80], and further construct high-quality editing data using inpainting and video-based pipelines. For in-context generation and editing, we build our datasets from video sources to model consistent subjects across varying scenarios. We employ vision-language models for subject detection, segmentation, and semantic filtering, resulting in diverse and semantically consistent triplets for training.

Fostering Higher-Level Reasoning. Finally, to push the base model's capabilities beyond simple generation, we construct interleaved and reflection datasets to enhance temporal reasoning and self-correction capabilities in multimodal models. Detailed pipeline steps, examples and the capability of reflection are provided in Appendix 9.2, 9.3, 9.4, 9.5, 9.6.

3. Method

OmniGen2 is built on three key components: (1) a decoupled architecture for unified generation, (2) Omni-RoPE for efficient contextual learning, and (3) a multi-stage training and alignment curriculum that progresses from broad knowledge to fine-grained instruction following.

3.1. Overall Architecture

We aim to design a simple, efficient, and versatile architecture for multimodal generation. Following this principle, OmniGen2 utilizes decoupled pathways for text and image generation. It employs two distinct transformer modules to efficiently facilitate the concurrent support of both understanding and generation capabilities, as illustrated in Figure 2. The autoregressive transformer model is initialized from a VLM (Qwen2.5-VL-3B [3]). This VLM provides extensive world knowledge and deep understanding of multimodal instructions. The diffusion transformer is randomly initialized and dedicated solely to high-fidelity image synthesis.

The two modules operate in sequence. First, the VLM processes the input multimodal context. A special token, `<|img|>`, is learned to distinguish the understanding and generation tasks. The generation of this token triggers the image generation. The corresponding hidden states from the VLM are extracted and fed to the diffusion decoder as a condition. It encodes high-level semantic instruction. Besides the high-level semantic encoding from the VLM, we incorporate low-level image features to ensure consistency of fine-grained visual details for tasks like image editing. We utilize Flux-VAE [33] for this purpose. This approach avoids complex architectural modifications to the pre-trained VLM, thereby preserving its powerful instruction understanding capabilities.

3.2. Diffusion Decoder

OmniGen2 employs computational efficient conditioning mechanism for the diffusion decoder. Existing methods such as MetaQuery [57] compressing instructions into a fixed set of learnable query tokens can create an information bottleneck. In contrast, OmniGen2 directly leverages the rich hidden states from the VLM's final layer. Furthermore, we utilize only the hidden states corresponding to text tokens, as the VAE features already provide sufficient visual detail.

Within the diffusion decoder, we adopt a unified transformer backbone, following the architecture of Lumina-Image 2.0 [60], where the parameters are shared across modalities. This design choice is motivated by the motivation that language and vision share substantial semantic representations. Consequently, parameter sharing provides a more natural and efficient means of cross-modal alignment than maintaining separate pathways [17, 33, 34]. Meanwhile, this design facilitates more consistent information exchange between modalities. Before processed by the core transformer blocks, input conditioning signals (VLM hidden states, VAE features, and noisy latents) are aligned by a lightweight two-layer transformer refiner. This refiner shares the same architecture as the transformer block employed in Lumina-Image 2.0.

3.3. Omni-RoPE: Unified Positional Encoding

We introduce **Omni-RoPE**, a positional encoding scheme tailored for multimodal contexts with complex structural

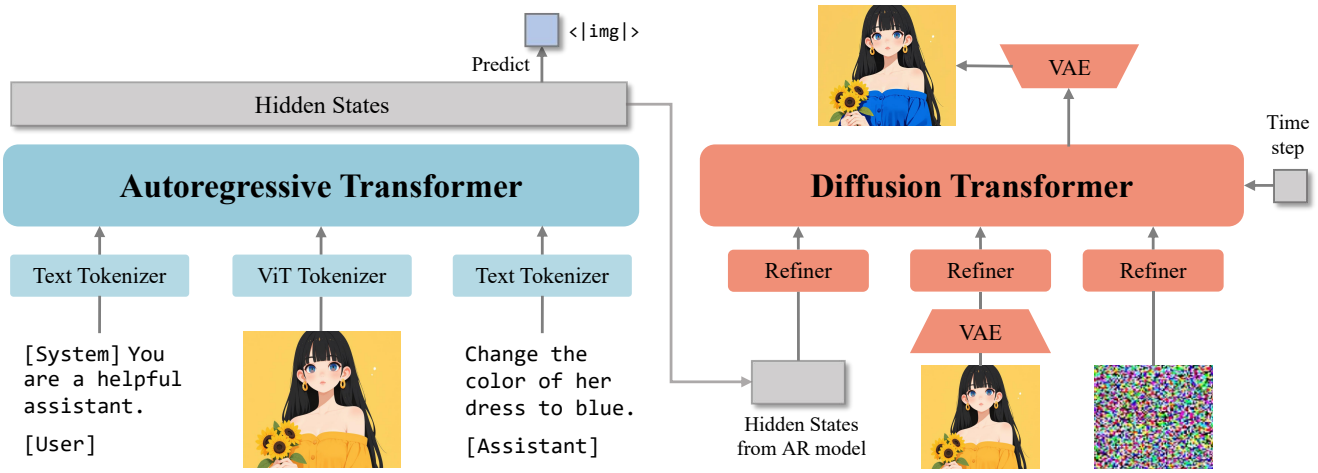

Figure 2. Architecture of OmniGen2. OmniGen2 employs separate transformers for autoregressive and diffusion. Two distinct image encoders are utilized: ViT encodes images for input into the text transformer, while VAE encodes images for the diffusion transformer.

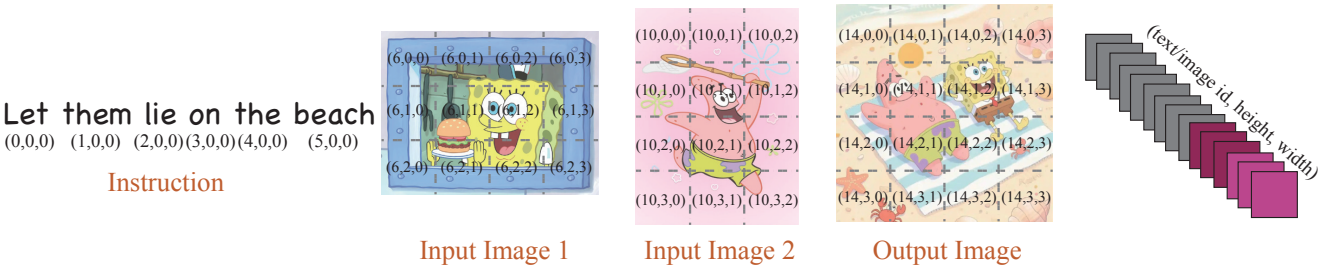

Figure 3. Illustration of **Omni-RoPE**. Each token in the k-th image is assigned a three-dimensional positional identifier $(\Delta_I^{(k)}, h, w)$, where $\Delta_I^{(k)}$ denotes the *instance identity* shared by all tokens within the same image, and (h, w) are the local 2D spatial coordinates computed from $(0, 0)$. This decomposition enables the model to distinguish different images while preserving local spatial consistency for tasks such as image editing.

Method	**PosID** $\mathrm{PosID}_k(h, w)$	**Steps to Target** ↓	**Final Loss** ↓
Lumina-Image-2.0's	$(0,\ h+\Delta_h,\ w+\Delta_w)$	∼2,500	0.017
Qwen2-VL's	$(\Delta_I,\ h+\Delta_I,\ w+\Delta_I)$	∼1,200	0.005
Omni-RoPE (Ours)	$(\Delta_I,\ h,\ w)$	∼**800**	0.003
+ Image Index Emb.		∼**800**	**0.002**

Table 1. Comparison of RoPE designs in the toy reconstruction task. Models are trained to reproduce the k-th image among randomly sampled inputs. We report the number of steps required to reach the target (*loss* < 0.014). Omni-RoPE achieves both faster convergence and lower final loss. **Note:** Δ_h and Δ_w denote the accumulated coordinate offsets in the height and width dimensions, respectively, while Δ_I represents the accumulated offset in the instance dimension.

correspondence. Conventional positional encodings cannot reliably distinguish multiple images or preserve spatial alignment across editing operations. Omni-RoPE addresses this limitation by extending Rotary Position Embedding (RoPE) [71] to a unified multimodal setting.

Unified formulation. As shown in Figure 3, each token at coordinates (h, w) in the k-th image is assigned a three-dimensional positional identifier:

$$\mathrm{PosID}_k(h, w) = (\Delta_I^{(k)},\ h,\ w), \tag{1}$$

where $\Delta_I^{(k)}$ denotes the *instance identity*, which distinguishes different images or modalities, and (h, w) are the *2D spatial coordinates*. All tokens from the same image share the same $\Delta_I^{(k)}$, while the spatial mapping remains unchanged, i.e., $\mathcal{P}_h^{(k)}(h) = h$ and $\mathcal{P}_w^{(k)}(w) = w$. For text tokens, this formulation naturally reduces to a standard 1D positional index.

This decomposition separates image identity from intra-image spatial layout. Spatial coordinates are computed locally from $(0, 0)$ within each image, ensuring that corresponding patches in input and output images receive identical embeddings, thereby preserving spatial alignment and edit consistency. Meanwhile, $\Delta_I^{(k)}$ provides an explicit channel for distinguishing visual instances, which is critical for in-context image generation and multi-image reasoning.

Toy experiment verification. To evaluate positional correspondence, we design a controlled toy task in which a randomly initialized model is trained to reconstruct the k-th image from multiple randomly sampled input images, thereby isolating the effect of positional encoding. We measure efficiency by the number of training steps required to reach a high-fidelity reconstruction target (*loss* < 0.014).

As reported in Table 1, the RoPE variants used in Lumina-Image-2.0 and Qwen2-VL [77] require substantially more training steps to converge, indicating weaker alignment across visual instances. In contrast, **Omni-RoPE** converges markedly faster and achieves the lowest reconstruction loss, demonstrating stronger spatial correspondence and instance discrimination. Incorporating an *image index embedding* [14] further improves the final reconstruction fidelity at no additional cost.

3.4. Foundation Model Training

We construct the foundation model using a two-stage training pipeline comprising from-scratch pre-training followed by supervised fine-tuning. To accommodate variable context lengths in unified multi-task training, we employ FlashAttention2 [16] for efficient sequence processing. The model is optimized using the Rectified Flow objective [2, 42, 47].

Pre-training. This stage focuses on learning general-purpose visual and semantic representations from large-scale datasets. The model is trained through a resolution-based curriculum ($256^2 \rightarrow 512^2 \rightarrow 1024^2$). For each resolution, we first conduct training on the text-to-image (T2I) task to establish strong text–image alignment. Then, we introduce a curated mixed-task dataset (covering image editing and in-context generation) to diversify the model's capabilities.

Supervised Fine-Tuning. After pre-training, the model undergoes SFT at 1024^2 resolution to refine high-level reasoning and compositional skills. We train on a mixture of curated datasets and distilled data from proprietary models, aiming to enhance instruction following and visual fidelity.

Through two-stage training, the model acquired initial instruction-following skills and versatile generation capability, setting a foundation for subsequent alignment. Detailed configurations for each stage are provided in Appendix 9.7.

3.5. Instruction Alignment

We perform online reinforcement learning for multi-task alignment via a progressive curriculum instead of a single joint training stage to ensure stability and avoid task interference. The key challenge is to achieve synergistic gains across tasks without degrading individual performance.

We define a sequence of training tasks $\mathcal{S} = \langle \mathcal{T}_1, \ldots, \mathcal{T}_N \rangle$, where each task $\mathcal{T} = (\tau, \delta, \mathcal{R})$ consists of a **task type** $\tau \in \{\text{T2I, Edit, IC}\}$, a **task instance** δ, and a **reward signal** $\mathcal{R}$. Our goal is to cover all fundamental task types for comprehensive alignment.

For $\tau = \text{Edit}$ and $\tau = \text{IC}$, we adopt general-purpose tasks to enhance instruction-following and compositional abilities. As these tasks lack verifiable rewards, we employ learned reward models: EditScore [49] for Edit and Qwen2.5-VL-72B [3] for IC. For $\tau = \text{T2I}$, we select **GenEval**, which provides verifiable rewards and exhibits strong overlap with Edit and IC.

We exclude T2I tasks with limited generalization or high reward-hacking risk. In particular, aesthetic rewards such as HPSv3 [50] are omitted due to reward hacking, and specialized tasks (e.g., OCR) are excluded as they lack synergy with general instruction-following.

This yields a three-stage curriculum $\langle \mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3 \rangle$, trained with Flow-GRPO [45]:

- $\mathcal{T}_1 =$ (Edit, general editing, EditScore),
- $\mathcal{T}_2 =$ (T2I, GenEval, Verifiable Reward),
- $\mathcal{T}_3 =$ (IC, general in-context, Qwen2.5-VL-72B).

Our RL data includes 50k T2I prompts from Flow-GRPO [47], 110k editing samples from EditScore [49], and 180k in-context data from Echo-4o [87].

4. OmniContext Benchmark

Rigorous evaluation is essential for generalized instruction alignment, particularly for reference-based tasks testing core consistency. However, existing benchmarks fall short, lacking support for multiple input images and diverse tasks. For instance, DreamBench [64] only contains 30 objects and 25 prompt templates. And relying on simplistic metrics like CLIP-I fails on multi-subject evaluation and offers no explainability. To address these critical gaps, we introduce OmniContext, a comprehensive benchmark designed to assess a model's ability to generate content consistent with user-specified context images.

To bridge these gaps, we construct OmniContext using a large-scale, manually collected dataset of high-quality images including personal photos, open-source images, animation stills and AI-generated images. These images are grouped into three distinct categories — Character, Object, and Scene — and exhibit diverse coverage across various domains, as illustrated in Figure 4. We define three task categories(SINGLE, MULTIPLE, and SCENE), each with 50 examples per subtask. SINGLE uses one context image, MULTIPLE combines multiple subjects, and SCENE conditions on environmental context.

Image–prompt pairs are constructed through a hybrid process combining MLLMs and manual annotation. MLLMs first filter low-quality samples, after which experts select images based on clarity, aesthetics, and diversity. Prompts are generated with GPT-4o and refined for semantic and syntactic variety.

We use GPT-4.1 to assess outputs on three metrics: Prompt Following (PF), Subject Consistency (SC), and an Overall Score (geometric mean of PF and SC). Following VIEScore [32], GPT-4.1 provides both scores (0–10) and rationales to justify its evaluations. We believe the OmniContext will serve as a valuable resource for driving future research in controllable, reference-based image generation.

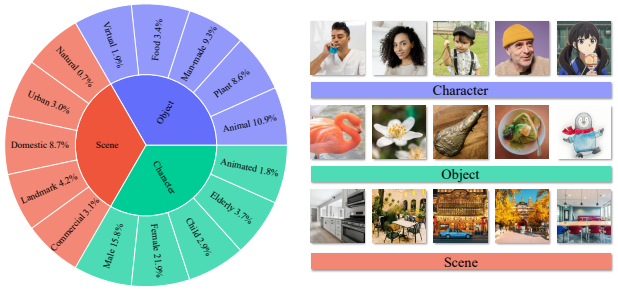

Figure 4. Overview of OmniContext benchmark. **Left:** Image genres included in OmniContext. **Right:** Example images for each genre in OmniContext.

5. Experiments

In this section, we conduct a comprehensive evaluation of OmniGen2 to demonstrate its unified capabilities across a wide spectrum of generation tasks. The overall comparison results are presented in Table 2.

5.1. Visual Understanding

OmniGen2 leverages Qwen2.5-VL-3B-Instruct [3] for visual understanding. As shown in Table 2, our model demonstrates robust multimodal comprehension, achieving solid scores of 79.1 on MMBench [48], 53.1 on MMMU [91], and 61.8 on MM-Vet [90]. These results confirm that OmniGen2 possesses a solid foundation for interpreting complex visual and textual instructions, which is essential for high-quality, instruction-aligned generation.

5.2. Text-to-Image Generation

We assess OmniGen2's T2I generation capabilities on two standard benchmarks: GenEval [23], which evaluates compositional generation, and OneIG-Bench [7] which evaluate models across multiple dimensions, including prompt-image alignment, text rendering precision, reasoning-generated content, stylization, and diversity.

As shown in Table 2, OmniGen2 delivers strong image generation performance on complex, compositional prompts, achieving an overall score of **0.95** on GenEval. This surpasses other powerful unified models such as UniWorld-V1 (0.84) and BAGEL (0.88), as well as Qwen-Image, a model specialized for T2I generation, highlighting the effectiveness of the RL alignment strategy in OmniGen2. On the more comprehensive OneIG-Bench, OmniGen2 continues to demonstrate competitive realism, achieving an overall score of 0.47, outperforming most existing models and trailing only behind large-scale models such as Gemini 2.5 Flash Image and Qwen-Image. More details are provided in Appendix 8.1 and 9.10.

5.3. Image Editing

Image editing is a cornerstone of OmniGen2's capabilities. We rigorously evaluate its performance across three diverse benchmarks: Emu-Edit [68], GEdit-Bench-EN [46] and ImgEdit-Bench [88]. The results collectively demonstrate that OmniGen2 achieve a strong performance in instruction-based image editing.

As shown in Table 3, OmniGen2 exhibits an exceptional balance between edit accuracy and image preservation. On Emu-Edit, our model achieves the highest CLIP-Out score (0.311), indicating it most effectively applies the requested edits among all compared models. Concurrently, it secures the second-best scores for CLIP-I (0.896) and best scores for DINO (0.876), which measure the preservation of unedited regions. This combination highlights OmniGen2's proficiency in making precise, localized changes without disturbing the rest of the image. This strong instruction-following capability is further confirmed on GEdit-Bench, where OmniGen2 achieves the second-highest Semantic Consistency (SC) score of 7.58 and the highest Perceptual Quality (PQ) score of 7.94. This leads to a strong overall score of 7.21, placing it among the top-tier models. This score **outperforms** Gemini-2.5-Flash-Image and is **second only** to Qwen-Image-Edit among open-source models. As detailed in Table 2, OmniGen2 demonstrates compelling performance on the comprehensive ImgEdit-Bench, notably surpassing some strong open-source models like BAGEL. More details are provided in Appendix 8.2.

5.4. In-context Generation

A distinguishing feature of OmniGen2 is its capacity to perform in-context generation. We introduce the **OmniContext** benchmark to provide a comprehensive evaluation of the performance of the existing model in this domain. OmniContext comprises eight subtasks, with overall scores for each subtask presented in Table 4. As the inaugural model evaluated on this benchmark, OmniGen2 establishes a strong baseline, achieving an overall score of 7.95, which **surpass** the powerful open-sourced model Qwen-Image-Edit-2509. These results show OmniGen2's proficiency in disentangling the subject's identity from its original background and re-rendering it accurately according to new textual instructions. OmniGen2 exhibits significant improvements over competing models in all types of tasks, demonstrating superior prompt-following ability and subject consistency. Among closed-source models, GPT-4o [56] achieves the highest scores in the Overall metrics. More details are provided in Appendix 8.3, 9.8.

5.5. Ablation Study

Our ablation study, detailed in Table 5, validates our principled curriculum by demonstrating the critical importance of two key factors: the **selection** of tasks and reward signals,

Model	# Params	Understanding			Image Generation		Image Editing		In-context Generation		
		MMB↑	MMMU↑	MM-Vet↑	GenEval↑	OneIG-Bench-EN↑	ImgEdit-Bench↑	GEdit-Bench-EN↑	Single↑	Multiple↑	Scene↑
LLaVA-1.5 [44]	-	36.4	67.8	36.3	-	-	-	-	-	-	-
LLaVA-NeXT [43]	-	79.3	51.1	57.4	-	-	-	-	-	-	-
SDXL [58]	-	-	-	-	0.55	0.32	-	-	-	-	-
SD3-medium [1]	-	-	-	-	0.62	-	-	-	-	-	-
FLUX.1-dev [33]	-	-	-	-	0.66	0.43	-	-	-	-	-
Qwen-Image [81]	-	-	-	-	0.91	0.54	-	-	-	-	-
Instruct-P2P [5]	-	-	-	-	-	-	1.88	3.68	-	-	-
MagicBrush [92]	-	-	-	-	-	-	1.90	1.86	-	-	-
AnyEdit [89]	-	-	-	-	-	-	2.45	3.21	-	-	-
Step1X-Edit [46]	-	-	-	-	-	-	3.06	6.70	-	-	-
IC-Edit [94]	-	-	-	-	-	-	3.05	4.84	-	-	-
UNO [83]	-	-	-	-	-	-	-	-	6.72	4.48	3.59
InfiniteYou [31]	-	-	-	-	-	-	-	-	6.05	-	-
DreamO [54]	-	-	-	-	-	-	-	-	7.65	7.05	4.52
UMO [15]	-	-	-	-	-	-	-	-	7.78	7.14	6.78
Show-o [85]	-	-	27.4	-	0.68	-	-	-	-	-	-
Janus-Pro [13]	-	75.5	36.3	39.8	0.80	0.26	-	-	-	-	-
Emu3 [78]	-	58.5	31.6	37.2	0.54 / 0.66†	-	-	-	-	-	-
MetaQuery-XL [57]	7B + 1.6B*	83.5	**58.6**	66.6	0.80†	-	-	-	-	-	-
BLIP3-o 8B [11]	7B + 1.4B*	83.5	**58.6**	66.6	0.84†	0.31	-	-	-	-	-
BAGEL [17]	7B + 7B*	**85.0**	55.3	67.2	0.82 / 0.88†	0.36	3.20	6.52	6.25	6.02	5.08
UniWorld-V1 [41]	7B + 12B*	83.5	**58.6**	67.1	0.84†	-	3.26	4.85	-	-	-
Qwen-Image-Edit-2509 [81]	7B + 20B	-	-	-	-	-	**4.41**	**7.54**	8.74	**8.13**	6.55
Gemini 2.5 Flash Image [25]	-	-	-	-	-	**0.55**	4.28	7.10	**8.77**	8.06	7.01
OmniGen [84]	3.8B	-	-	-	0.68	-	2.96	5.06	6.46	5.26	4.34
OmniGen2	3B + 4B*	79.1	53.1	61.8	**0.95**	0.47	3.69	7.21	8.41	7.73	**7.86**

Table 2. Comparison of different models across Understanding, Generation, Editing, and In-context Generation tasks. *: The first term represents the number of parameters for text generation, while the second term corresponds to the number of parameters allocated for image generation. † refers to the methods using LLM rewriter.

Method	Emu-Edit			GEdit-Bench-EN		
	CLIP-I↑	CLIP-Out↑	DINO↑	SC↑	PQ↑	O↑
Gemini-2.0-Flash-Image [24]	-	-	-	6.73	6.61	6.32
Gemini-2.5-Flash-Image [25]	-	-	-	7.41	7.96	7.10
GPT-4o [56]	-	-	-	7.85	7.62	7.53
Instruct-Pix2Pix [5]	0.856	0.292	0.773	3.58	5.49	3.68
MagicBrush [92]	0.877	0.298	0.807	4.68	5.66	4.52
AnyEdit [89]	-	-	-	3.18	5.82	3.21
OmniGen [84]	-	-	-	5.96	5.89	5.06
ICEdit [94]	**0.907**	0.305	0.866	5.11	6.85	4.84
Step1X-Edit [46]	0.860	0.304	0.782	7.09	6.76	6.70
BAGEL [17]	0.839	0.307	0.753	7.36	6.83	6.52
UniWorld-V1 [41]	-	-	-	4.93	7.43	4.85
Qwen-Image-Edit-2509 [81]	-	-	-	**8.15**	7.86	**7.54**
OmniGen2	0.896	**0.311**	**0.876**	7.58	**7.94**	7.21

Table 3. Quantitative comparison on Emu-Edit [69] and GEdit-Bench-EN [46]. For Emu-Edit, CLIP-I/DINO measure consistency with the source image, while CLIP-Out measures alignment with the caption of target image, CLIP-B/32 [61] and DINO-S/16 [6] are leveraged for feature calculation. For GEdit-Bench, SC (Semantic Consistency) evaluates instruction following, and PQ (Perceptual Quality) assesses image naturalness and artifacts. Higher scores are better for all metrics.

and the **scheduling** of the training sequence. For task selection, we highlight four crucial findings: (1) Tasks with limited skill overlap can cause negative interference, as shown by `OCR only` training which degrades the GEdit Overall score from 6.28 to 6.13. (2) Conversely, well-chosen tasks with skill overlap such as instruction following exhibit strong synergy; the `Edit & GenEval` strategy surpasses both single-task baselines on their respective metrics (GenEval: 0.95 vs. 0.94; GEdit Overall: 7.19 vs. 7.01). (3) Reward signals about human preference pose significant risks, with `Edit & HPSv3` confirming reward hacking by inflating the PQ score to 8.22 at the severe cost of collapsing SC and IC scores. (4) accuracy reward signal is vital. As shown by `Edit only`, whose IC score is higher than `IC only` (7.71 vs. 7.38) because of excel performance of EditScore [49] to enhance instruction following. Beyond selection, the training sequence is equally vital. This is confirmed by comparing our final curriculum (`Edit & Geneval & IC`) against an alternative ordering (`Edit & IC & Geneval`), which results in a marked performance drop (GEdit Overall: 7.21 vs. 7.06). We also observe that prioritizing editing tasks leads to consistently better performance than T2I-first. We hypothesize this is because editing tasks with richer supervision build a robust foundation for subsequent learning. And Additional results on out-of-distribution (OOD) benchmarks are provided in Appendix 9.10, showing consistent improvements under our RL curriculum. These findings collectively prove that both careful selection and scheduling are essential to our principled alignment strategy.

6. Related Works

6.1. Multimodal Generation

Recent advances in multimodal generation have produced models capable of both understanding and generating content across text, images, and video. Diffusion-based models, including the Stable Diffusion series [19, 58, 63],

Method	SINGLE		MULTIPLE			SCENE			Average↑
	Character	Object	Character	Object	Char. + Obj.	Character	Object	Char. + Obj.	
Flux.1 Kontext max [34]	8.48	8.68	-	-	-	-	-	-	-
Gemini-2.0-Flash-Image [24]	5.06	5.17	2.91	2.16	3.80	3.02	3.89	2.92	3.62
Gemini-2.5-Flash-Image [25]	8.62	8.91	7.88	8.92	7.39	7.29	7.05	6.68	7.84
GPT-4o [56]	**8.90**	**9.01**	**9.07**	**8.95**	**8.54**	**8.90**	**8.44**	**8.60**	**8.80**
InfiniteYou [31]	6.05	-	-	-	-	-	-	-	-
UNO [83]	6.60	6.83	2.54	6.51	4.39	2.06	4.33	4.37	4.71
BAGEL [17]	5.48	7.03	5.17	6.64	6.24	4.07	5.71	5.47	5.73
Qwen-Image-Edit-2509 [81]	**8.35**	**9.13**	**7.65**	**8.85**	7.90	5.16	7.75	6.73	7.69
OmniGen [84]	7.21	5.71	5.65	5.44	4.68	3.59	4.32	5.12	4.34
OmniGen2	8.19	8.63	7.45	7.91	**7.93**	**7.75**	**7.91**	**7.93**	**7.95**

Table 4. Overall comparison of existing models on our proposed OmniContext benchmark. "Char. + Obj." indicates Character + Object.

Strategy	GenEval ↑	OmniContext ↑	GEdit ↑		
			SC	PQ	Overall
Base (w/o RL)	0.78	7.18	6.72	7.20	6.28
Single-Task					
Edit only	0.79	7.71	7.30	7.95	7.01
GenEval only	0.94	7.24	6.78	7.20	6.30
OCR only	0.78	7.33	6.65	7.15	6.13
IC only	0.78	7.38	6.97	6.98	6.39
Multi-Tasks					
Edit & GenEval	**0.95**	7.68	7.52	7.95	7.19
Edit & OCR	0.81	7.70	7.28	7.96	7.06
Edit & HPSv3	0.77	6.82	6.87	**8.22**	6.88
Edit & IC & GenEval	0.93	7.65	7.33	7.92	7.06
Geneval & Edit & IC	0.94	7.80	7.49	7.97	**7.21**
Edit & GenEval & IC (Ours)	**0.95**	**7.95**	**7.58**	7.94	**7.21**

Table 5. Ablation study of multi-task reinforcement learning strategies. T2I, Edit, and IC tasks are trained for 1500, 700, and 200 steps, respectively.

DALL·E [62], and Imagen [30], have achieved high-fidelity image synthesis, while methods like ControlNet [93] and T2I-Adapter [53] improve controllability, and StyleShot, InstructPix2Pix, and EMU-Edit [5, 21, 69] support fine-grained, instruction-guided editing. Unified image generation models such as OmniGen [84], UniReal [14], and related works [14, 51, 75, 84] extend this further, integrating multiple tasks into a single model. Building on this foundation, autoregressive multimodal models provide an alternative paradigm for unified generation [73, 74, 78]. There are also hybrid approaches such as Show-o and Transfusion [18, 26, 57, 70, 85, 95] combining autoregressive text generation with diffusion-based image modeling. Several works focus on adapting large language models for multimodal generation [11, 17, 40, 76, 82]: These works are trained with vast amount of data, obtaining powerful image understanding and image generation capabilities.

6.2. Reinforcement Learning in Diffusion Model

Reinforcement learning has increasingly been adopted to improve alignment in diffusion and flow-based models. For text-to-image (T2I) generation, early works such as DDPO and DPOK [4, 20] formulated diffusion sampling as a sequential decision process and optimized via KL-regularized policy updates. Follow-up approaches including ReFL, AlignProp [8, 59] refined this paradigm with more stable reward optimization, improved credit assignment across denoising steps, and scalable preference-learning from human or synthetic feedback. More recently GRPO [67] has become prominent due to its training stability and efficiency. GRPO-based alignment method like DanceGRPO [86], Flow-GRPO [45], and Mix-GRPO [36] further push the boundaries of alignment technology, outperforming traditional methods in both accuracy and scalability [27, 39, 79]. For image editing or in-context generation, RL has also been used to enforce text alignment and editing faithfulness to ensure consistency between input and output [15, 28, 49, 52]. Despite the rapid progress, most existing approaches optimize RL for a single task or a narrow alignment objective. In contrast, our work introduces a multi-task RL pipeline that jointly aligns the model's behavior across all three critical scenarios, achieving comprehensive, all-around alignment.

7. Conclusion

In this work, we present OmniGen2, a generative model that is systematically instruction aligned. OmniGen2 explores two directions to enhance alignment performance: constructing a robust and flexible base model, and developing a multi-task RL alignment scheme. OmniGen2 utilizes a simple, efficient and flexible architecture to support diverse multimodal generation tasks. Our experiments across standard benchmarks and our propose novel OmniContext benchmark demonstrate OmniGen2's semantic consistency, versatile capabilities, and superior generation quality. Instruction alignment has consistently and significantly enhanced the base model across various tasks. These results suggest that instruction alignment may represent a crucial step toward realizing general multimodal systems.

References

[1] Stability AI. Sd3-medium. https://stability.ai/news/stable-diffusion-3-medium, 2024. 7

[2] Michael S Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. *arXiv preprint arXiv:2209.15571*, 2022. 5

[3] 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, 5, 6

[4] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning, 2024. 8

[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. 7, 8

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 7

[7] Jingjing Chang, Yixiao Fang, Peng Xing, Shuhan Wu, Wei Cheng, Rui Wang, Xianfang Zeng, Gang Yu, and Hai-Bao Chen. Oneig-bench: Omni-dimensional nuanced evaluation for image generation. *arXiv preprint arXiv:2506.07977*, 2025. 6

[8] Chaofeng Chen, Annan Wang, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. Enhancing diffusion models with text-encoder reinforcement learning. In *European Conference on Computer Vision*, pages 182–198. Springer, 2024. 8

[9] Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model, 2024. 3

[10] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.

[11] Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025. 7, 8

[12] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023. 3

[13] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025. 7

[14] Xi Chen, Zhifei Zhang, He Zhang, Yuqian Zhou, Soo Ye Kim, Qing Liu, Yijun Li, Jianming Zhang, Nanxuan Zhao, Yilin Wang, et al. Unireal: Universal image generation and editing via learning real-world dynamics. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12501–12511, 2025. 5, 8

[15] Yufeng Cheng, Wenxu Wu, Shaojin Wu, Mengqi Huang, Fei Ding, and Qian He. Umo: Scaling multi-identity consistency for image customization via matching reward. *arXiv preprint arXiv:2509.06818*, 2025. 7, 8

[16] Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023. 5

[17] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. 2, 3, 7, 8

[18] Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. Dreamllm: Synergistic multimodal comprehension and creation. *arXiv preprint arXiv:2309.11499*, 2023. 8

[19] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 7

[20] Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36:79858–79885, 2023. 8

[21] Junyao Gao, Yanchen Liu, Yanan Sun, Yinhao Tang, Yanhong Zeng, Kai Chen, and Cairong Zhao. Styleshot: A snapshot on any style. *arXiv preprint arXiv:2407.01414*, 2024. 8

[22] Yuying Ge, Sijie Zhao, Chen Li, Yixiao Ge, and Ying Shan. Seed-data-edit technical report: A hybrid dataset for instructional image editing. *arXiv preprint arXiv:2405.04007*, 2024. 3

[23] 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. 6

[24] Google. Gemini 2.0 flash. https://developers.googleblog.com/en/experiment-with-gemini-20-flash-native-image-generation, 2025. 7, 8

[25] Google. Introducing gemini 2.5 flash image. https://developers.googleblog.com/en/introducing-gemini-2-5-flash-image/, 2025. Accessed: 2025-09-18. 1, 7, 8

[26] Agrim Gupta, Linxi Fan, Surya Ganguli, and Li Fei-Fei. Metamorph: Learning universal controllers with transformers. *arXiv preprint arXiv:2203.11931*, 2022. 8

[27] Xiaoxuan He, Siming Fu, Yuke Zhao, Wanli Li, Jian Yang, Dacheng Yin, Fengyun Rao, and Bo Zhang. Tempflow-grpo: When timing matters for grpo in flow models. *arXiv preprint arXiv:2508.04324*, 2025. 8

[28] Ziwei Huang, Ying Shu, Hao Fang, Quanyu Long, Wenya Wang, Qiushi Guo, Tiezheng Ge, and Leilei Gan. From competition to synergy: Unlocking reinforcement learning for subject-driven image generation. *arXiv preprint arXiv:2510.18263*, 2025. 8

[29] 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. 1

[30] Imagen-Team-Google. Imagen 3, 2024. 8

[31] Liming Jiang, Qing Yan, Yumin Jia, Zichuan Liu, Hao Kang, and Xin Lu. Infiniteyou: Flexible photo recrafting while preserving your identity. *arXiv preprint arXiv:2503.16418*, 2025. 7, 8

[32] Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. Viescore: Towards explainable metrics for conditional image synthesis evaluation, 2023. 5

[33] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 3, 7

[34] Black Forest Labs. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space. 2025. 1, 3, 8

[35] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 3

[36] Junzhe Li, Yutao Cui, Tao Huang, Yinping Ma, Chun Fan, Miles Yang, and Zhao Zhong. Mixgrpo: Unlocking flow-based grpo efficiency with mixed ode-sde, 2025. 8

[37] Xianhang Li, Haoqin Tu, Mude Hui, Zeyu Wang, Bingchen Zhao, Junfei Xiao, Sucheng Ren, Jieru Mei, Qing Liu, Huangjie Zheng, Yuyin Zhou, and Cihang Xie. What if we recaption billions of web images with llama-3? *arXiv preprint arXiv:2406.08478*, 2024. 3

[38] Xiaotong Li, Fan Zhang, Haiwen Diao, Yueze Wang, Xinlong Wang, and Ling-Yu Duan. Densefusion-1m: Merging vision experts for comprehensive multimodal perception. *2407.08303*, 2024. 3

[39] Yuming Li, Yikai Wang, Yuying Zhu, Zhongyu Zhao, Ming Lu, Qi She, and Shanghang Zhang. Branchgrpo: Stable and efficient grpo with structured branching in diffusion models. *arXiv preprint arXiv:2509.06040*, 2025. 8

[40] Chao Liao, Liyang Liu, Xun Wang, Zhengxiong Luo, Xinyu Zhang, Wenliang Zhao, Jie Wu, Liang Li, Zhi Tian, and Weilin Huang. Mogao: An omni foundation model for interleaved multi-modal generation. *arXiv preprint arXiv:2505.05472*, 2025. 2, 8

[41] 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. 7

[42] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 5

[43] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 7

[44] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 7

[45] 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. 5, 8

[46] 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, 7

[47] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 5

[48] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024. 6

[49] Xin Luo, Jiahao Wang, Chenyuan Wu, Shitao Xiao, Xiyan Jiang, Defu Lian, Jiajun Zhang, Dong Liu, and Zheng Liu. Editscore: Unlocking online rl for image editing via high-fidelity reward modeling. *arXiv preprint arXiv:2509.23909*, 2025. 5, 7, 8

[50] Yuhang Ma, Yunhao Shui, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. *CoRR*, abs/2508.03789, 2025. 5

[51] Chaojie Mao, Jingfeng Zhang, Yulin Pan, Zeyinzi Jiang, Zhen Han, Yu Liu, and Jingren Zhou. Ace++: Instruction-based image creation and editing via context-aware content filling. *arXiv preprint arXiv:2501.02487*, 2025. 8

[52] Yanting Miao, William Loh, Suraj Kothawade, Pascal Poupart, Abdullah Rashwan, and Yeqing Li. Subject-driven text-to-image generation via preference-based reinforcement learning. *Advances in Neural Information Processing Systems*, 37:123563–123591, 2024. 8

[53] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models, 2023. 8

[54] Chong Mou, Yanze Wu, Wenxu Wu, Zinan Guo, Pengze Zhang, Yufeng Cheng, Yiming Luo, Fei Ding, Shiwen Zhang, Xinghui Li, et al. Dreamo: A unified framework for image customization. *arXiv preprint arXiv:2504.16915*, 2025. 7

[55] Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, et al. Docci: Descriptions of connected and contrasting images. *arXiv preprint arXiv:2404.19753*, 2024. 3

[56] OpenAI. Gpt-4o. https://openai.com/index/introducing-4o-image-generation, 2025. 6, 7, 8

[57] Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, et al. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025. 2, 3, 7, 8

[58] 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. *arXiv preprint arXiv:2307.01952*, 2023. 7

[59] Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation. 2023. 8

[60] Qi Qin, Le Zhuo, Yi Xin, Ruoyi Du, Zhen Li, Bin Fu, Yiting Lu, Jiakang Yuan, Xinyue Li, Dongyang Liu, et al. Lumina-image 2.0: A unified and efficient image generative framework. *arXiv preprint arXiv:2503.21758*, 2025. 3

[61] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 7

[62] 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. 8

[63] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 7

[64] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 5

[65] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 3

[66] Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025. 1

[67] 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, 8

[68] Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8871–8879, 2024. 6

[69] Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8871–8879, 2024. 7, 8

[70] Weijia Shi, Xiaochuang Han, Chunting Zhou, Weixin Liang, Xi Victoria Lin, Luke Zettlemoyer, and Lili Yu. Llamafusion: Adapting pretrained language models for multimodal generation. *arXiv preprint arXiv:2412.15188*, 2024. 8

[71] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 4

[72] Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, et al. Journeydb: A benchmark for generative image understanding. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[73] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14398–14409, 2024. 8

[74] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 8

[75] Xueyun Tian, Wei Li, Bingbing Xu, Yige Yuan, Yuanzhuo Wang, and Huawei Shen. Mige: A unified framework for multimodal instruction-based image generation and editing. *arXiv preprint arXiv:2502.21291*, 2025. 8

[76] 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. 8

[77] 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. 5

[78] 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. 7, 8

[79] Yibin Wang, Zhimin Li, Yuhang Zang, Yujie Zhou, Jiazi Bu, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. Pref-grpo: Pairwise preference reward-based grpo for stable text-to-image reinforcement learning. *arXiv preprint arXiv:2508.20751*, 2025. 8

[80] Cong Wei, Zheyang Xiong, Weiming Ren, Xeron Du, Ge Zhang, and Wenhu Chen. Omniedit: Building image edit-

ing generalist models through specialist supervision. In *The Thirteenth International Conference on Learning Representations*, 2024. 1, 3

[81] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. 1, 7, 8

[82] 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. 8

[83] Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation. *arXiv preprint arXiv:2504.02160*, 2025. 7, 8

[84] 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. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13294–13304, 2025. 1, 7, 8

[85] 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. 7, 8

[86] Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025. 8

[87] Junyan Ye, Dongzhi Jiang, Zihao Wang, Leqi Zhu, Zhenghao Hu, Zilong Huang, Jun He, Zhiyuan Yan, Jinghua Yu, Hongsheng Li, et al. Echo-4o: Harnessing the power of gpt-4o synthetic images for improved image generation. *arXiv preprint arXiv:2508.09987*, 2025. 5

[88] Yang Ye, Xianyi He, Zongjian Li, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, and Li Yuan. Imgedit: A unified image editing dataset and benchmark. *arXiv preprint arXiv:2505.20275*, 2025. 6

[89] 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. 7

[90] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 6

[91] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. 6

[92] 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, 2024. 7

[93] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3836–3847, 2023. 8

[94] 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. *arXiv preprint arXiv:2504.20690*, 2025. 7

[95] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024. 8