Query-Kontext: An Unified Multimodal Model for Image Generation and Editing

Yuxin Song[1⋆], Wenkai Dong[1*], Shizun Wang[2*], Qi Zhang[1], Song Xue[1], Tao Yuan[1], Hu Yang[1], Haocheng Feng[1], Hang Zhou[1], Xinyan Xiao[1], and Jingdong Wang[1⋆⋆]

[1] Baidu Inc, China
[2] National University of Singapore, Singapore

Abstract. Unified Multimodal Models (UMMs) excel in text-to-image generation and editing but often entangle multimodal generative reasoning with high-fidelity visual synthesis. We introduce Query-Kontext, a novel approach that bridges a Vision-Language Model (VLM) and a diffusion model via multimodal "kontext" tokens. This design cleanly decouples the complex reasoning delegated to the VLM, including instruction understanding, grounding, and identity preservation, from the high-quality visual rendering executed by the diffusion model. We propose a three-stage progressive training strategy: (1) connecting the VLM to a lightweight diffusion head to activate generative reasoning; (2) scaling to a large, pre-trained diffusion model to enhance visual realism; and (3) incorporating a low-level image encoder for fine-grained instruction tuning. Supported by a comprehensive multi-task dataset, extensive experiments demonstrate that Query-Kontext matches or outperforms state-of-the-art task-specific and unified methods across diverse reference-to-image scenarios.

Keywords: Text-to-image generation · Image editing · Unified Multimodal Models

1 Introduction

Unified Multimodal Models (UMMs) have recently achieved notable progress in both image generation (T2I) [5, 9, 22, 23, 26, 30, 42, 48–50, 70] and editing (TI2I) [6, 21, 35, 36, 41, 62, 67, 80, 82]. Two prominent design paradigms have emerged from this work. The first assembled unified framework leverages external diffusion transformers, such as MMDiT [23,69], which are paired with off-the-shelf vision–language models (VLMs) or large language models (LLMs) to provide semantic conditioning. The second paradigm, naive UMMs, integrate generation and understanding more tightly through mixed-modal early-fusion transformers [17, 22, 44, 57, 59, 65, 85], where autoregressive modules with strong reasoning ability are jointly trained with diffusion modules specialized in visual synthesis.

⋆ Equal contribution.
⋆⋆ Corresponding author.

Fig. 1: Showcase of Query-Kontext model on multimodal reference-to-image tasks.

While these paradigms expand task coverage and streamline deployment, they also entangle multimodal generative reasoning and high-fidelity rendering. Consequently, the unique strengths of VLMs (semantic understanding, grounding, structured reasoning [2,3,19,20,61,77,78,84]) and diffusion models (photorealistic synthesis and detail fidelity [10–12, 29, 30, 46, 69]) cannot be fully exploited. We identify two sources of this limitation. First, *assembled* unified frameworks typically use a frozen VLM or LLM as a static feature extractor, narrowing the conditioning signal to only high-level semantics for the diffusion generator. Second, *native* UMMs force generative reasoning and visual rendering to be optimized jointly, introducing capacity competition and hindering generalization, particularly when tasks demand both fine-grained edits and strong semantic control. While attempts to mitigate these issues through methods like mixture-of-experts (e.g., LlamaFusion [53]) or mixture-of-transformers (e.g., BAGEL [22]) have been made, they only partially alleviate the tension.

In this work, we propose *Query-Kontext*, an economic ensemble UMM that leverages the multimodal *"kontext"* composed of semantic and coarse image conditions to cleanly decouple the generative reasoning of VLM from the high-fidelity rendering of diffusion model. To realize this separation, we develop a three-stage progressive training strategy. *Stage 1:* Bridge the VLM to a lightweight diffusion head through *"kontext"* tokens. Using parameter-efficient fine-tuning (LoRA) [31], we **unleash** the potential of VLM and steer it toward multimodal generative reasoning skills such as instruction following, spatial grounding, and identity-preserving image referencing. *Stage 2:* **Scale** the lightweight head to a well-trained large diffusion model (roughly 10× more parameters). We re-align both the text and *"kontext"* tokens from the VLM to the scaled diffusion model by using text-to-image generation and image-reconstruction objectives. *Stage 3:* Introduce a low-level image encoder [1] that injects fine-grained structural and textural cues into the diffusion model while keeping the VLM frozen. This step strengthens identity preservation [54, 63, 72, 79] and reconstruction fidelity in [33, 41, 76, 80] challenging editing scenarios.

In summary, our contributions are:

- We propose *Query-Kontext*, an economic ensemble UMM that decouples multimodal generative reasoning in VLMs from the high-fidelity visual rendering performed by diffusion models.
- We present a three-stage progressive training strategy that progressively aligns the VLM with increasingly capable diffusion generators while amplifying their respective strengths in generative reasoning and visual synthesis.
- We present a deliberate dataset curation scheme to collect real, synthetic, and carefully filtered open-source datasets to cover diverse multimodal reference-to-image scenarios.

2 Data Curation

We constructed a multimodal reference-to-image dataset (as summarized in Table 1) comprising a mixture of real, synthetic, and carefully curated open-source

datasets. This dataset spans five categories of tasks: text-to-image generation, image transformation, instruction editing, customized generation, and multi-subject composition.

Text-to-Image Generation and Image Reconstruction. We collected 30M open-source English image-text pairs (including ShareGPT-4o-Image [13], BLIP-3o [8], among others) as well as 170M in-house Chinese image-text pairs for text-to-image generation and image reconstruction tasks. The in-house data underwent extensive quality filtering based on image resolution, clarity, aesthetic score, watermark detection, and safety compliance. Among these Chinese data, 150M belong to general categories (balanced across diverse domains), and 20M come from specific vertical domains (e.g., artistic styles, logos, automobiles, text-containing images, celebrities, posters, etc.).

Image Transformation. Following the MetaQuery [49], we constructed naturally occurring image pairs from web corpora [14,39] and generated corresponding open-ended transformation instructions by leveraging multi-modal large language models (MLLMs). Specifically, we clustered images that share the same accompanying caption from sources like MMC4-core [14], OmniCorpus-CC [39] and OmniCorpus-CW [39] by using SigLIP [60] image features, then filtered these clusters by a similarity threshold to obtain 0.8M image transformation triplets (*see Appendix for example triplets*).

Instruction Editing. For the image editing instruction task, we first aggregated approximately 3M image-instruction-image triplets from open-source datasets, including NHR-Edit [35](358k samples), GPT-Image-Edit [47](1.5M samples), MagicBrush [34](10k samples), and OmniEdit [67](1.2M samples). We further filtered the MagicBrush subset using CLIP-based image and text similarity scores, and translated all datasets' instructions into Chinese using a large language model. Building upon the methodologies of [35, 41, 67], we then constructed a synthetic data pipeline tailored for native Chinese instruction editing, producing an additional 300k high-quality triplets (*see Appendix for details on the data pipeline*). Finally, inspired by UniReal [16], we extended our dataset with video-based clusters derived from raw videos to cover more non-rigid editing tasks (e.g., motion changes, viewpoint shifts, view transitions such as zoom-in and zoom-out (*see Appendix for the video data examples*).

Customized Generation. We leveraged the open-source Subject-200K [56] and UNO-1M [72] datasets for customized (subject-driven) image generation. In addition, we augmented our data with portrait reference triplets synthesized using a dedicated model [4], which generates reference images of specific individuals. Through this approach, we accumulated approximately 0.3M portrait reference samples that maintain high facial similarity to the source subjects while exhibiting substantial diversity in poses, attire, and other attributes.

Multi-Subject Composition. Finally, we addressed multi-subject image composition using the open-source MUSAR-Gen [28] dataset and a new synthetic data pipeline (*see details in Appendix*). This yielded 40k multi-subject reference examples, each featuring compositions of multiple humans, objects, and

complex scenes, thereby enriching the dataset's coverage of realistic multi-entity interactions.

Table 1: The data outline about each training stage. † denotes only the Chinese prompt.

Stage	Task	Data source	Size
1, 2, 3	Image generation, Image reconstruction	ShareGPT-4o-Image, BLIP3o	30M
		in-house real data†	170M
1	Image transformation	mmc4, OmniCorpus	800K
3	Instruction editing	NHR-Edit, GPT-Edit, OmniEdit	3M
		In-house video data	2M
		In-house real data	300K
	Customized generation	subject200k	200K
		in-house real data	1.8M
	Multi-subject composition	MUSAR-Gen	29K
		GPT-4o synthesized data	40K

3 Query-Kontext

In this work, we propose Query-Kontext, a unified multimodal model for image generation and editing that delegates multimodal generative reasoning to the VLM while reserving the diffusion model's capability for high-quality visual synthesis. In Sec 3.1, we present the architectural design of the Query-Kontext model (Figure 2). In Sec 3.2, we design a three-stage progressive learning strategy and introduce the details of training recipe (Figure 3). In Sec 3.3, we introduce the implementation details of model hyper-parameters and infrastructures.

3.1 Architecture

As shown in Figure 2, Query-Kontext comprises four main components: a Multimodal Large Language Model (MLLM), a connector module, a Multimodal Diffusion Transformer (MMDiT), and a low-level image encoder (VAE). The MLLM is initialized with the Qwen2.5-VL model [3], which encodes and fuses multimodal inputs including the text prompt, input image(s), and a set of learnable query tokens. The output is a fixed-length sequence of *kontext* tokens $Q = \{q_1, \ldots, q_K\}$ which serves as coarse image-level conditioning for the diffusion decoder while providing high-level semantic cues. Intuitively, the *kontext* tokens Q encode what content should appear in the output image (the semantic information from the text prompt) and how the output should incorporate visual cues from the provided input images, as enforced by the training supervision in Sec. 3.2. The *kontext* Q and text tokens T are passed through a lightweight connector module to align them with the diffusion model's latent space. In

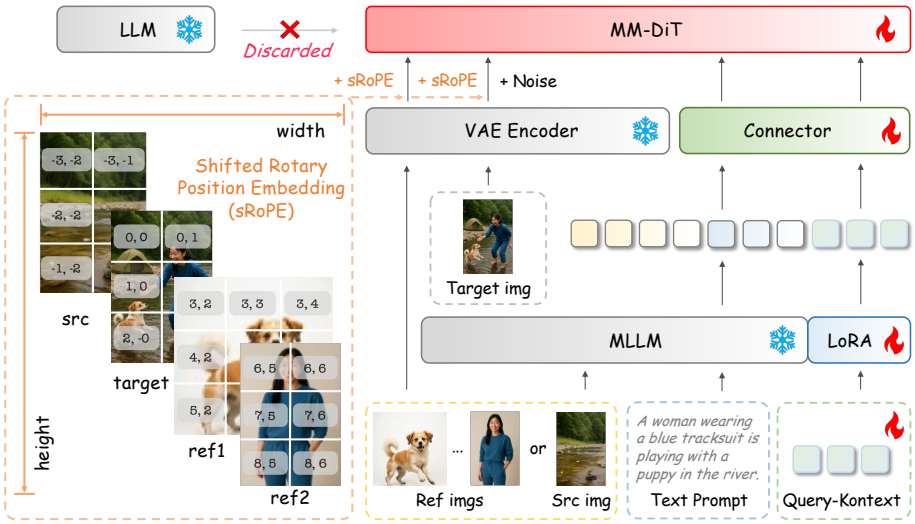

Fig. 2: The overall framework of the unified multi-modal to image generation and editing model, Query-Kontext.

practice, we concatenate the noisy latent, the low-level image latent with sRoPE, the text latent T and the "kontext" token Q from the connector, and subsequently feed them into MMDiT, thereby enriching the semantic context and image details, which are made available for the diffusion model. We initialize the diffusion model using our in-house MMDiT model and replace its original text encoder with the MLLM (*training details for this alignment are discussed in Sec. 3.2*). We concatenate the sequence of text T and *kontext* tokens Q from the MLLM with: (i) the noisy image latent at the current diffusion step t, and (ii) the low-level visual feature tokens extracted from the input image(s) by the VAE. The concatenated sequence is then fed into the MMDiT model in an in-context manner [36,83], allowing the diffusion model to attend to both the textual prompt and the visual cues from the input images.

Moreover, we distinguish between naive UMMs and assembled UMMs in Table 2. The comparison highlights whether each model trains from scratch, freezes parameters, or adapts pretrained components, and specifies the flow of information through text embeddings (TE), low-level image embeddings (LE), and query embeddings (QE). In particular, query embeddings naturally unleash the in-context learning capabilities of the VLM, enabling the model to reason over multimodal inputs and generate coherent images. Unlike prior methods, **our Query-Kontext integrates query embeddings alongside text and low-level image embeddings, while effectively decoupling understanding and generation modules for improved efficiency and flexibility.**

Furthermore, we design a *shifted 2D Rotary Position Embedding* (RoPE) scheme [55, 72] to incorporate multi-image positional conditioning and avoid

Table 2: Comparison of mainstream unified multimodal models on the modeling paradigms and the information flow. 🔥 denotes training from scratch, ❄ indicates freezing the parameters during training and ❄ → 🔥 represents training from a pretrained model. For the input modalities, "TE" refers to text embeddings, "LE" to low-level image embeddings, and "QE" to query embeddings.

Method	Module			Information		
	Understanding	Connector	Generation	TE	LE	QE
Native UMMs						
Janus-Pro [17]	🔥	-	🔥	✓	✗	✗
OmniGen2 [71]	❄ → 🔥	-	🔥	✓	✓	✗
BAGEL [22]	🔥	-	🔥	✓	✗	✗
Assembled UMMs						
Metaquery [49]	❄	🔥	❄ → 🔥	✓	✗	✓
Step1X-Edit [41]	❄	🔥	❄ → 🔥	✓	✓	✗
Uniworld-v1 [40]	❄	🔥	❄ → 🔥	✓	✓	✗
FLUX.1 Kontext [36]	❄	-	🔥	✓	✓	✗
Qwen-Image [70]	❄	-	🔥	✓	✓	✗
Query-Kontext (Ours)	❄ → 🔥	🔥	❄ → 🔥	✓	✓	✓

confusion among multiple reference images (as illustrated in Figure 2). In the standard diffusion architecture, each spatial position of a latent feature map (with size of $h \times w$) is identified by a 2D index (i, j), where $i \in [0, w-1]$ and $j \in [0, h-1]$. We introduce a task-specific prior to adjust these coordinates based on the fidelity requirements of the input images. For tasks requiring pixel-level fidelity to an input image (e.g., instruction-based editing), we treat the input image as a *source image*, denoted img_{src}. For tasks requiring identity preservation (e.g., personalized generation or multi-image composition), we treat the input image as a *reference image*, denoted img_{ref}. We then shift the coordinate indices of the VAE latent for each image type accordingly: for reference image latents, we shift indices into the positive quadrant, whereas for the source image latent, we shift into the negative quadrant.

As shown in the left subfigure of Figure 2, we adopt two concatenation orders depending on task requirements: 1) for the tasks requiring strict pixel fidelity (e.g., instruction-based editing), we concat(img_{src}, $noise$) with a shifted RoPE in the negative spatial direction. 2) for the asks emphasizing identity preservation with flexibility (e.g., customized generation), we concat($noise$, img_{ref}) along the positive direction. we define the coordinates for the n-th reference latent as:

$$\left(i^n_{\text{ref}},\, j^n_{\text{ref}}\right) = \left(i + w * n, j + h * n\right) \tag{1}$$

where $i \in [0, w-1]$, $j \in [0, h-1]$ and $n \in [1, N]$. Meanwhile, for the source image latent we shift the coordinates in the negative direction:

$$(i'_{\text{src}},\, j'_{\text{src}}) \;=\; (-i,\;\; -j)\,, \tag{2}$$

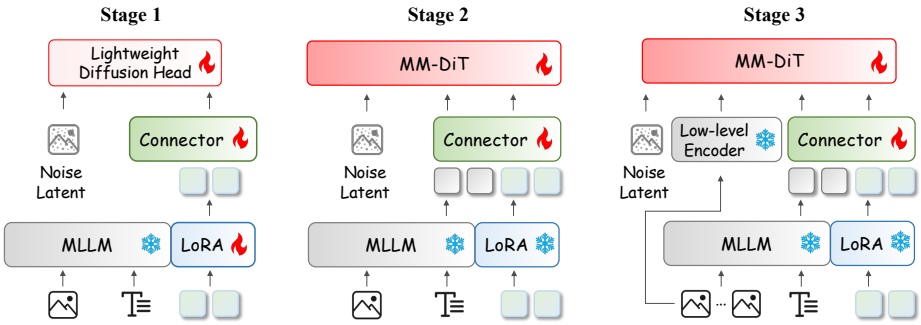

Fig. 3: Three training stages of Query-Kontext. Note that the Diffusion Head is only used in the Stage 1. In the Stage 2 and 3, we scale up Diffusion model to 10× parameters and keep MLLM frozen to provide coarse-grained image conditions.

where $i \in [0, w-1]$, $j \in [0, h-1]$ and $n \in [1, N]$. Finally, we add the shifted RoPE on the feature maps of the input image latent(s) and the noisy latent at their respective shifted coordinates (i.e., added element-wise to each spatial location).

3.2 Individualized-Teaching Curriculum

As shown in Figure 3, we propose a three-stage progressive training strategy that both unlocks the generative reasoning capabilities of the VLM and progressively aligns it with increasingly powerful diffusion generators. As a result, Query-Kontext, guided by multimodal *kontext* tokens, effectively decouples the multimodal generative reasoning of the VLM from the high-fidelity visual rendering carried out by diffusion models.

Stage 1: We unleash the generative reasoning potential of the MLLM through two key architectural designs: we first use learnable query tokens ("kontext") to represent a mixture of semantic cues and coarse-grained image conditions, and then align the output *kontext* tokens with a lightweight diffusion head that performs noisy prediction at a coarse level. We train all parameters of the connector, the diffusion head, and the MLLM's LoRA modules on a trio of tasks: text-to-image generation, image reconstruction, and image transformation (see Section 2). This training methodology preserves the MLLM's inherent language-vision understanding while cultivating its emergent ability for multimodal generative reasoning.

Stage 2: Next, we replace the lightweight diffusion head with our in-house diffusion model based on the MMDiT architecture for high-fidelity generation. In Stage 2, the full MLLM parameters (the LoRA parameters are merged into the MLLM) remain frozen, and we optimize the *kontext* tokens, the connector, and all parameters of the large diffusion model. In preliminary experiments, we observed that completely freezing the diffusion model was feasible for smaller head but failed for a larger diffusion model (*the experiments details and discussion*

Table 3: The data outline and training details about each training stage. Where, $Q.$ denotes the Query-kontext tokens, *Con.* is Connector module.

Stage	Stage 1	Stage 2	Stage 3
Task	Image Generation Image Reconstruction Image Transformation	Image Generation Image Reconstruction	Instruction Editing Customized Generation Multi-subject
Type	T2I, I2I, TI2I	T2I, I2I	T2I, TI2I
Training Param.	MLLM's LoRA, *Con.*, Diffusion head, $Q.$	*Con.*, MMDiT, $Q.$	MMDiT's LoRA, *Con.*, $Q.$
Global Batch Size	512	1024	512
Steps (K)	72	420	30
Learning Rate	1e-4	1e-4	2e-5

are available in Section 5). Therefore, we allow the diffusion model to be full-parameters fine-tuning in this stage. To keep training efficient, Query-Kontext is trained only on text-to-image generation and image reconstruction tasks in this stage, which accelerates convergence and reduces training cost for fast alignment from MLLM to the diffusion model.

Stage 3: Finally, we introduce a dedicated low-level image encoder for source or reference images to further refine the diffusion model for high-fidelity image referring. In Stage 3, the MLLM remains fully frozen, and we optimize only the Query-Kontext tokens and the connector. Additionally, we apply the LoRA-based fine-tuning to the diffusion model itself to preserve its high-quality image synthesis ability while extending it to all our tasks. This includes not only standard text-to-image generation but also instruction-guided image editing, user-customized image generation, and multi-subject composition tasks.

3.3 Implementation

Architecture. We initialize the MLLM from Qwen2.5-VL-7B and implement the connector as a two-layer MLP. (*details of architecture configuration are provided in Appendix*) Moreover, we implement the diffusion head in the stage a with a lightweight MMDiT architecture ($\sim 870M$ parameters). We set the max reference images $N = 2$ and $K = 128$ in Query-kontext $Q = \{q_1, \ldots, q_K\}$. We set rank $r_d = 256, \alpha_d = 256$ in the diffusion model's and rank $r_m = 128, \alpha_m = 256$ in the MLLM's LoRA.

Training recipe. The default configuration on the resolution with 512×512 is provided in Table 3. After Stage 3, we introduce a resolution upscaling stage using the same mixed multi-task dataset at a higher resolution. In this stage, the training resolution is increased to 1024×1024, the learning rate is further reduced to 1×10^{-5}, and training continues for an additional 3,000 steps with a global batch size of 256.

Table 4: Quantitative Evaluation results on GenEval [27]. † refer to the methods using the LLM rewriter.

Model	Single Object	Two Object	Counting	Colors	Position	Attribute Binding	Overall↑
Show-o [75]	0.95	0.52	0.49	0.82	0.11	0.28	0.53
Emu3-Gen [66]	0.98	0.71	0.34	0.81	0.17	0.21	0.54
PixArt-α [11]	0.98	0.50	0.44	0.80	0.08	0.07	0.48
SD3 Medium [24]	0.98	0.74	0.63	0.67	0.34	0.36	0.62
FLUX.1 [Dev] [5]	0.98	0.81	0.74	0.79	0.22	0.45	0.66
SD3.5 Large [24]	0.98	0.89	0.73	0.83	0.34	0.47	0.71
JanusFlow [45]	0.97	0.59	0.45	0.83	0.53	0.42	0.63
Lumina-Image 2.0 [51]	-	0.87	0.67	-	-	0.62	0.73
Janus-Pro-7B† [18]	0.99	0.89	0.59	0.90	0.79	0.66	0.80
HiDream-I1-Full† [7]	**1.00**	0.98	0.79	0.91	0.60	0.72	0.83
GPT-Image† [47]	0.99	0.92	0.85	0.92	0.75	0.61	0.84
Seedream 3.0† [26]	0.99	**0.96**	**0.91**	0.93	0.47	0.80	0.84
Qwen-Image† [70]	0.99	0.92	0.89	0.88	0.76	0.77	0.87
BAGEL† [22]	0.98	0.95	0.84	**0.95**	0.78	0.77	**0.88**
Query-Kontext†	0.98	0.94	0.81	0.91	**0.85**	**0.79**	**0.88**

4 Experiments

4.1 Quantitative results

Table 5: Quantitative Evaluation results on GEdit-Bench. G_SC is Semantic Consistency, G_PQ is Perceptual Quality, and G_O is Overall Score which is computed as the geometric mean of G_SC and G_PQ, averaged over all samples. All metrics are evaluated by GPT-4. We highlight the **best** and <u>second-best</u> values for each metric.

Model	GEdit-Bench-EN (Full set)↑			GEdit-Bench-CN (Full set)↑		
	G_SC	G_PQ	G_O	G_SC	G_PQ	G_O
Instruct-Pix2Pix [6]	3.58	5.49	3.68	-	-	-
AnyEdit [81]	3.18	5.82	3.21	-	-	-
MagicBrush [82]	4.68	5.66	4.52	-	-	-
UniWorld-v1 [40]	4.93	7.43	4.85	-	-	-
OmniGen [74]	5.96	5.89	5.06	-	-	-
OmniGen2 [71]	7.16	6.77	6.41	-	-	-
Gemini 2.0 [21]	6.73	6.61	6.32	5.43	6.78	5.36
BAGEL [22]	7.36	6.83	6.52	7.34	6.85	6.50
FLUX.1 Kontext [Pro] [36]	7.02	7.60	6.56	1.11	7.36	1.23
Step1X-Edit [42]	7.66	7.35	6.97	7.20	6.87	6.86
GPT Image 1 [High] [47]	7.85	<u>7.62</u>	7.53	7.67	<u>7.56</u>	7.30
Qwen-Image [70]	<u>8.00</u>	**7.86**	<u>7.56</u>	<u>7.82</u>	**7.79**	<u>7.52</u>
Query-Kontext	**8.36**	7.37	**7.66**	**8.39**	7.35	**7.65**

We evaluate Query-Kontext on a comprehensive suite of benchmarks, spanning text-to-image generation, instruction-guided editing, subject-driven customiza-

Table 6: Quantitative results for single-subject driven generation on Dreambooth. We highlight the **best** and second-best values.

Method	DINO ↑	CLIP-I ↑	CLIP-T ↑
Tuning-free			
Textual Inversion [25]	0.569	0.780	0.255
DreamBooth [52]	0.668	0.803	0.305
BLIP-Diffusion [37]	0.670	0.805	0.302
Specialist Models			
ELITE [68]	0.647	0.772	0.296
Re-Imagen [15]	0.600	0.740	0.270
OminiControl [56]	0.684	0.799	0.312
FLUX.1 IP-Adapter [5]	0.582	0.820	0.288
UNO-FLUX [72]	0.760	0.835	0.304
Generalist Models			
OmniGen [73]	0.693	0.801	**0.315**
MIGE [58]	0.744	0.830	0.293
Metaquery [49]	0.737	0.851	0.301
BAGEL [22]	0.777	0.851	0.307
OmniGen2 [71]	0.749	0.830	0.310
Query-Kontext	**0.786**	**0.858**	0.307

Table 7: Quantitative results for multi-subject driven generation on Dreambench. We highlight the **best** and second-best values for each metric.

Method	DINO ↑	CLIP-I ↑	CLIP-T ↑
Tuning-free			
DreamBooth [52]	0.430	0.695	0.308
BLIP-Diffusion [37]	0.464	0.698	0.300
Specialist Models			
Subject Diffusion [43]	0.506	0.696	0.310
MIP-Adapter [32]	0.482	0.726	0.311
MS-Diffusion [64]	0.525	0.726	0.319
UNO-FLUX [72]	**0.542**	**0.733**	0.322
Generalist Models			
OmniGen [73]	0.511	0.722	0.331
BAGEL [22]	0.439	0.683	0.335
OmniGen2 [71]	0.488	0.716	0.332
Query-Kontext	0.532	0.731	**0.336**

tion, and multi-subject composition. Specifically, we report results on GenEval, GEdit-Bench, DreamBooth, and DreamBench.

On GenEval, Query-Kontext attains an overall score of 0.88, matching the SOTA result of unified UMM (BAGEL [22]), as illustrated in Table 4. Our results are reported based on Chinese prompts rewritten by DeepSeek [3]. On GEdit-Bench, Query-Kontext achieves the highest overall performance in instruction-guided editing, with scores of 7.66 on the English split and 7.65 on the Chinese split. These results surpass Qwen-Image (7.56 / 7.52) and GPT-Image (7.53 / 7.30), as shown in Table 5. We note that the Perceptual Quality score exhibits some shortcomings, primarily due to the lack of a reinforcement learning or supervised fine-tuning stage designed to enhance generation quality or photorealism. We leave this exploration in future work. For subject-driven generation on DreamBooth, Query-Kontext establishes new state-of-the-art results with DINO 0.786 and CLIP-I 0.858, significantly outperforming Metaquery (0.737 / 0.851) and UNO-FLUX (0.760 / 0.835), though with a slightly lower CLIP-T (0.307 vs. OmniGen's 0.315), as shown in Table 6. In Table 7, Query-Kontext achieves the best CLIP-T score (0.336) alongside competitive DINO (0.532) and CLIP-I results (0.731), on the multi-subject composition benchmark DreamBench.

4.2 Qualitative Results

We also provide qualitative comparisons across all task categories, including text-to-image generation, instruction editing, and customized generation, under both Chinese and English prompts. Representative examples are shown in Figure 1.

4.3 Shifted RoPE

[3] `https://chat.deepseek.com/`

We further examine the effect of the proposed shifted 2D-RoPE mechanism for handling reference images. With *source* input images, the model tends to preserve the pixel-level fidelity of the input, producing faithful reconstructions. In contrast, with *reference* input images, the model emphasizes instruction following and generalization, maintaining subject identity while generating more diverse outputs. Comparative results on the DreamBooth benchmark using source versus reference images are reported in Table 8.

Table 8: The comparison between the shifted RoPE on *source* or $reference$.

Method	DINO ↑	CLIP-I ↑	CLIP-T ↑
w/ src_{img}	0.865	0.914	0.289
w/ ref_{img}	0.786	0.858	0.307

4.4 Query–Kontext Convergence

In Stage 2, we analyze the convergence behavior of the diffusion model when conditioned on two settings: (i) text-only embeddings from an LLM and (ii) the mixed conditioning from both text tokens and Query–Kontext tokens generated by our fine-tuned VLM. We observe that replacing the LLM with our VLM leads to faster alignment of the diffusion model and produces superior visual results compared to the LLM-conditioned baseline, as shown in Figure 4. This demonstrates that decoupling multimodal reasoning from visual generation via Query–Kontext not only accelerates convergence but also unleashes the full potential of both the VLM and the diffusion model.

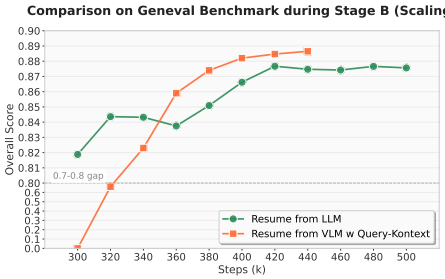

Fig. 4: Convergence validation of Query–Kontext. Comparison on in-house MMDiT between VLM re-alignment with Query–Kontext and LLM-based resumption.

4.5 Qualitative results on low-level encoder.

We provide comparative results of the low-level encoder in the editing task. As shown in Figure 5, incorporating the low-level encoder improves the ability to edit fine-grained image details.

4.6 Decoupling of the VLM and Diffusion model

We report the visualizations in Figure6 comparing Stage 1 and Stage 2 outputs. **Stage 1 Output:** Correct semantic layout and pose (proven reasoning) but blurry textures (limited rendering capacity). **Stage 2 Output:** The same semantic layout is maintained, but with photorealistic details (unlocked rendering capacity). This empirically validates that the VLM handles the reasoning (structure), and the Diffusion model handles the synthesis (texture), bridged effectively by the *Kontext* tokens.

Fig. 5: Qualitative comparative results for the low-level encoder.

Fig. 6: Qualitative comparison between the Stage 1 and Stage 2 outputs.

4.7 Qualitative results on unseen concept generalization and zero-shot tasks.

To verify the effectiveness of our reasoning-rendering decoupling in challenging scenarios, we evaluated the model on unseen concept generalization and zero-shot tasks. As illustrated in Figure 7(a) and (b), we conducted customized generation and instruction editing experiments using concepts absent from the instruction-tuning dataset, such as specific commercial brands and distinct artistic styles. Query-Kontext successfully performs complex spatial layout planning and style transfer for these unseen concepts. This confirms that the framework effectively leverages the "World Knowledg" inherent in the pre-trained VLM, transferring it to the diffusion head via Kontext tokens, thereby significantly outperforming baselines that rely solely on text conditioning. Furthermore, we assessed the model's generalized reasoning capabilities through a zero-shot outpainting task, as shown in Figure 7(c). Despite not being explicitly trained for outpainting, Query-Kontext correctly infers the global scene context from the reference image (e.g., harmoniously extending a portrait's body and background).

4.8 Comparison of various bridge modules.

We empirically compared three settings (as shown in Table 9): (a) our full Query-Kontext (text + queries), (b) MLP over text tokens only, and (c) Q-Former–style queries only. With the same diffusion backbone, setting (c) is significantly harder to optimize and converges to worse GenEval performance (e.g., 0.751 vs. 0.882 at

Fig. 7: Qualitative results on the unseen concept generalization and zero-shot task.

Table 9: Comparison of different conditioning designs during **Stage 1**. Results are reported using the overall score (↑) on the **Geneval** benchmark.

Setting	iters	20k	40k	60k	80k	100k
a. Query-Kontext	Text (512) & query (128)	**0.661**	**0.823**	**0.859**	**0.874**	**0.882**
b. MLP	only Text (512)	0.587	0.67176	0.757	0.770	0.814
c. Qformer	only query (640)	0.359	0.596	0.689	0.732	0.751

100k steps), while our decoupled design delivers both faster convergence and better final scores.Together, these aspects make Query-Kontext more than a direct Q-Former [38] adaptation. By explicitly activating the complementary capabilities of VLMs and diffusion models through this decoupled design, our approach provides a clear architectural distinction specifically tailored for controllable and stable diffusion conditioning, distinguishing it from both Q-Former–style bridges and existing UMMs.

5 Discussion and Conclusion

Discussion. Query-Kontext economically aligns a strong VLM with an MMDiT diffusion backbone: training on 192 NVIDIA H100 (80GB) GPUs uses roughly 10% of the compute typically required to train a large diffusion model or integrated multimodal transformer from scratch. The decoupled design also permits independent component scaling. With 0.9B/4B/10B diffusion backbones, we find that aligning a large frozen model through a lightweight connector can be difficult; unfreezing the diffusion model during Stage 2 provides a practical remedy. A systematic study of connector scaling remains future work.

Conclusion. Query-Kontext decouples VLM-based generative reasoning from diffusion-based rendering. Across generation, editing, customized synthesis, and multi-subject composition, the three-stage framework achieves competitive performance.

References

1. AI, S.: sd-vae-ft-ema (2024), https://huggingface.co/stabilityai/sd-vae-ft-ema
2. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv:2308.12966 (2023)
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
4. Baidu AI Cloud: Qianfan model center: Portrait generation model. https://console.bce.baidu.com/qianfan/modelcenter/model/buildIn/detail/am-t3uhhjzbys6w (2024), accessed: 2024
5. BlackForest: Flux. https://github.com/black-forest-labs/flux (2024)
6. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18392–18402 (2023)
7. Cai, Q., Chen, J., Chen, Y., Li, Y., Long, F., Pan, Y., Qiu, Z., Zhang, Y., Gao, F., Xu, P., et al.: Hidream-i1: A high-efficient image generative foundation model with sparse diffusion transformer. arXiv preprint arXiv:2505.22705 (2025)
8. Chen, J., Xu, Z., Pan, X., Hu, Y., Qin, C., Goldstein, T., Huang, L., Zhou, T., Xie, S., Savarese, S., et al.: Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. arXiv preprint arXiv:2505.09568 (2025)
9. Chen, J., Ge, C., Xie, E., Wu, Y., Yao, L., Ren, X., Wang, Z., Luo, P., Lu, H., Li, Z.: Pixart-sigma: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In: ECCV (2024)
10. Chen, J., Ge, C., Xie, E., Wu, Y., Yao, L., Ren, X., Wang, Z., Luo, P., Lu, H., Li, Z.: Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In: European Conference on Computer Vision. pp. 74–91. Springer (2024)
11. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wang, Z., Kwok, J.T., Luo, P., Lu, H., Li, Z.: Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In: ICLR (2024)
12. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., et al.: Pixart-*alpha*: Fast training of diffusion transformer for photorealistic text-to-image synthesis. arXiv preprint arXiv:2310.00426 (2023)
13. Chen, J., Cai, Z., Chen, P., Chen, S., Ji, K., Wang, X., Yang, Y., Wang, B.: Sharegpt-4o-image: Aligning multimodal models with gpt-4o-level image generation. arXiv preprint arXiv:2506.18095 (2025)
14. Chen, L., Bai, S., Chai, W., Xie, W., Zhao, H., Vinci, L., Lin, J., Chang, B.: Multimodal representation alignment for image generation: Text-image interleaved control is easier than you think. arXiv preprint arXiv:2502.20172 (2025)
15. Chen, W., Hu, H., Saharia, C., Cohen, W.W.: Re-imagen: Retrieval-augmented text-to-image generator. arXiv preprint arXiv:2209.14491 (2022)
16. Chen, X., Zhang, Z., Zhang, H., Zhou, Y., Kim, S.Y., Liu, Q., Li, Y., Zhang, J., Zhao, N., Wang, Y., et al.: Unireal: Universal image generation and editing via learning real-world dynamics. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12501–12511 (2025)

17. Chen, X., Wu, C., Wu, Z., Ma, Y., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C., Luo, P.: Janus-pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811 (2025)
18. Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C.: Janus-pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811 (2025)
19. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., Gu, L., Wang, X., Li, Q., Ren, Y., Chen, Z., Luo, J., Wang, J., Jiang, T., Wang, B., He, C., Shi, B., Zhang, X., Lv, H., Wang, Y., Shao, W., Chu, P., Tu, Z., He, T., Wu, Z., Deng, H., Ge, J., Chen, K., Zhang, K., Wang, L., Dou, M., Lu, L., Zhu, X., Lu, T., Lin, D., Qiao, Y., Dai, J., Wang, W.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024)
20. Chen, Z., Wang, W., Tian, H., Ye, S., Gao, Z., Cui, E., Tong, W., Hu, K., Luo, J., Ma, Z., et al.: Internvl2: Better than the best—expanding performance boundaries of open-source multimodal models with the progressive scaling strategy (2024), https://internvl.github.io/blog/2024-07-02-InternVL-2.0
21. DeepMind, G.: Gemini 2.0. https://gemini.google.com/ (2025)
22. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., et al.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025)
23. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: ICML (2024)
24. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: ICML (2024)
25. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)
26. Gao, Y., Gong, L., Guo, Q., Hou, X., Lai, Z., Li, F., Li, L., Lian, X., Liao, C., Liu, L., et al.: Seedream 3.0 technical report. arXiv preprint arXiv:2504.11346 (2025)
27. Ghosh, D., Hajishirzi, H., Schmidt, L.: Geneval: An object-focused framework for evaluating text-to-image alignment. Advances in Neural Information Processing Systems **36**, 52132–52152 (2023)
28. Guo, Z., Zhang, P., Wu, Y., Mou, C., Zhao, S., He, Q.: Musar: Exploring multi-subject customization from single-subject dataset via attention routing. arXiv preprint arXiv:2505.02823 (2025)
29. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020)
30. Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M., Salimans, T.: Cascaded diffusion models for high fidelity image generation. Journal of Machine Learning Research **23**(47), 1–33 (2022)
31. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv:2106.09685 (2021)
32. Huang, Q., Fu, S., Liu, J., Jiang, H., Yu, Y., Song, J.: Resolving multi-condition confusion for finetuning-free personalized image generation. arXiv preprint arXiv:2409.17920 (2024)
33. Huang, Y., Xie, L., Wang, X., Yuan, Z., Cun, X., Ge, Y., Zhou, J., Dong, C., Huang, R., Zhang, R., et al.: Smartedit: Exploring complex instruction-based image

editing with multimodal large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8362–8371 (2024)
34. Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6007–6017 (2023)
35. Kuprashevich, M., Alekseenko, G., Tolstykh, I., Fedorov, G., Suleimanov, B., Dokholyan, V., Gordeev, A.: Nohumansrequired: Autonomous high-quality image editing triplet mining. Available at SSRN 5381374 (2025)
36. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., et al.: Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. arXiv preprint arXiv:2506.15742 (2025)
37. Li, D., Li, J., Hoi, S.: Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. Advances in Neural Information Processing Systems **36**, 30146–30166 (2023)
38. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv:2301.12597 (2023)
39. Li, Q., Chen, Z., Wang, W., Wang, W., Ye, S., Jin, Z., Chen, G., He, Y., Gao, Z., Cui, E., et al.: Omnicorpus: A unified multimodal corpus of 10 billion-level images interleaved with text. arXiv preprint arXiv:2406.08418 (2024)
40. Lin, B., Li, Z., Cheng, X., Niu, Y., Ye, Y., He, X., Yuan, S., Yu, W., Wang, S., Ge, Y., et al.: Uniworld-v1: High-resolution semantic encoders for unified visual understanding and generation. arXiv preprint arXiv:2506.03147 (2025)
41. Liu, S., Han, Y., Xing, P., Yin, F., Wang, R., Cheng, W., Liao, J., Wang, Y., Fu, H., Han, C., Li, G., Peng, Y., Sun, Q., Wu, J., Cai, Y., Ge, Z., Ming, R., Xia, L., Zeng, X., Zhu, Y., Jiao, B., Zhang, X., Yu, G., Jiang, D.: Step1x-edit: A practical framework for general image editing. arXiv preprint arXiv:2504.17761 (2025)
42. Liu, S., Han, Y., Xing, P., Yin, F., Wang, R., Cheng, W., Liao, J., Wang, Y., Fu, H., Han, C., et al.: Step1x-edit: A practical framework for general image editing. arXiv preprint arXiv:2504.17761 (2025)
43. Ma, J., Liang, J., Chen, C., Lu, H.: Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–12 (2024)
44. Ma, Y., Liu, X., Chen, X., Liu, W., Wu, C., Wu, Z., Pan, Z., Xie, Z., Zhang, H., Yu, X., Zhao, L., Wang, Y., Liu, J., Ruan, C.: Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. arXiv preprint arXiv:2411.07975 (2024)
45. Ma, Y., Liu, X., Chen, X., Liu, W., Wu, C., Wu, Z., Pan, Z., Xie, Z., Zhang, H., Yu, X., et al.: Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7739–7751 (2025)
46. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: ICML (2021)
47. OpenAI: Gpt-image-1 (2025), `https://openai.com/index/introducing-4o-image-generation/`
48. OpenAI: Introducing 4o image generation (March 2025), `https://openai.com/index/introducing-4o-image-generation/`, accessed: 2025-05-09

49. Pan, X., Shukla, S.N., Singh, A., Zhao, Z., Mishra, S.K., Wang, J., Xu, Z., Chen, J., Li, K., Juefei-Xu, F., Hou, J., Xie, S.: Transfer between modalities with metaqueries. arXiv preprint arXiv:2504.06256 (2025)
50. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. In: ICLR (2024)
51. Qin, Q., Zhuo, L., Xin, Y., Du, R., Li, Z., Fu, B., Lu, Y., Yuan, J., Li, X., Liu, D., et al.: Lumina-image 2.0: A unified and efficient image generative framework. arXiv preprint arXiv:2503.21758 (2025)
52. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: CVPR. pp. 22500–22510 (2023)
53. Shi, W., Han, X., Zhou, C., Liang, W., Lin, X.V., Zettlemoyer, L., Yu, L.: Llamafusion: Adapting pretrained language models for multimodal generation. arXiv preprint arXiv:2412.15188 (2024)
54. Song, W., Jiang, H., Yang, Z., Quan, R., Yang, Y.: Insert anything: Image insertion via in-context editing in dit. arXiv preprint arXiv:2504.15009 (2025)
55. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024)
56. Tan, Z., Liu, S., Yang, X., Xue, Q., Wang, X.: Ominicontrol: Minimal and universal control for diffusion transformer. arXiv preprint arXiv:2411.15098 (2024)
57. Team, C.: Chameleon: Mixed-modal early-fusion foundation models. arXiv preprint arXiv:2405.09818 (2024)
58. Tian, X., Li, W., Xu, B., Yuan, Y., Wang, Y., Shen, H.: Mige: A unified framework for multimodal instruction-based image generation and editing. arXiv preprint arXiv:2502.21291 (2025)
59. Tong, S., Fan, D., Zhu, J., Xiong, Y., Chen, X., Sinha, K., Rabbat, M., LeCun, Y., Xie, S., Liu, Z.: Metamorph: Multimodal understanding and generation via instruction tuning. arXiv preprint arXiv:2412.14164 (2024)
60. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)
61. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., Lin, J.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv:2409.12191 (2024)
62. Wang, P., Shi, Y., Lian, X., Zhai, Z., Xia, X., Xiao, X., Huang, W., Yang, J.: Seededit 3.0: Fast and high-quality generative image editing. arXiv preprint arXiv:2506.05083 (2025)
63. Wang, Q., Bai, X., Wang, H., Qin, Z., Chen, A., Li, H., Tang, X., Hu, Y.: Instantid: Zero-shot identity-preserving generation in seconds. arXiv preprint arXiv:2401.07519 (2024)
64. Wang, X., Fu, S., Huang, Q., He, W., Jiang, H.: MS-diffusion: Multi-subject zero-shot image personalization with layout guidance. In: ICLR (2025), https://openreview.net/forum?id=PJqP0wyQek
65. Wang, X., Zhang, X., Luo, Z., Sun, Q., Cui, Y., Wang, J., Zhang, F., Wang, Y., Li, Z., Yu, Q., et al.: Emu3: Next-token prediction is all you need. arXiv preprint arxiv:2409.18869 (2024)

66. Wang, X., Zhang, X., Luo, Z., Sun, Q., Cui, Y., Wang, J., Zhang, F., Wang, Y., Li, Z., Yu, Q., et al.: Emu3: Next-token prediction is all you need. arXiv preprint arXiv:2409.18869 (2024)
67. Wei, C., Xiong, Z., Ren, W., Du, X., Zhang, G., Chen, W.: Omniedit: Building image editing generalist models through specialist supervision. In: ICLR (2024)
68. Wei, Y., Zhang, Y., Ji, Z., Bai, J., Zhang, L., Zuo, W.: Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In: CVPR. pp. 15943–15953 (2023)
69. William, P., Xie, S.: Scalable diffusion models with transformers. In: ICCV (2023)
70. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., ming Yin, S., Bai, S., Xu, X., Chen, Y., Chen, Y., Tang, Z., Zhang, Z., Wang, Z., Yang, A., Yu, B., Cheng, C., Liu, D., Li, D., Zhang, H., Meng, H., Wei, H., Ni, J., Chen, K., Cao, K., Peng, L., Qu, L., Wu, M., Wang, P., Yu, S., Wen, T., Feng, W., Xu, X., Wang, Y., Zhang, Y., Zhu, Y., Wu, Y., Cai, Y., Liu, Z.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025), https://arxiv.org/abs/2508.02324
71. Wu, C., Zheng, P., Yan, R., Xiao, S., Luo, X., Wang, Y., Li, W., Jiang, X., Liu, Y., Zhou, J., et al.: Omnigen2: Exploration to advanced multimodal generation. arXiv preprint arXiv:2506.18871 (2025)
72. Wu, S., Huang, M., Wu, W., Cheng, Y., Ding, F., He, Q.: Less-to-more generalization: Unlocking more controllability by in-context generation. arXiv preprint arXiv:2504.02160 (2025)
73. Xiao, S., Wang, Y., Zhou, J., Yuan, H., Xing, X., Yan, R., Li, C., Wang, S., Huang, T., Liu, Z.: Omnigen: Unified image generation. arXiv preprint arXiv:2409.11340 (2024)
74. Xiao, S., Wang, Y., Zhou, J., Yuan, H., Xing, X., Yan, R., Li, C., Wang, S., Huang, T., Liu, Z.: Omnigen: Unified image generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 13294–13304 (2025)
75. Xie, J., Mao, W., Bai, Z., Zhang, D.J., Wang, W., Lin, K.Q., Gu, Y., Chen, Z., Yang, Z., Shou, M.Z.: Show-o: One single transformer to unify multimodal understanding and generation. arXiv preprint arXiv:2408.12528 (2024)
76. Xu, Y., Kong, J., Wang, J., Pan, X., Lin, B., Liu, Q.: Insightedit: Towards better instruction following for image editing. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2694–2703 (2025)
77. Yao, H., Huang, J., Wu, W., Zhang, J., Wang, Y., Liu, S., Wang, Y., Song, Y., Feng, H., Shen, L., et al.: Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search. arXiv preprint arXiv:2412.18319 (2024)
78. Yao, H., Wu, W., Yang, T., Song, Y., Zhang, M., Feng, H., Sun, Y., Li, Z., Ouyang, W., Wang, J.: Dense connector for mllms. Advances in Neural Information Processing Systems **37**, 33108–33140 (2024)
79. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)
80. Ye, Y., He, X., Li, Z., Lin, B., Yuan, S., Yan, Z., Hou, B., Yuan, L.: Imgedit: A unified image editing dataset and benchmark. arXiv preprint arXiv:2505.20275 (2025)
81. Yu, Q., Chow, W., Yue, Z., Pan, K., Wu, Y., Wan, X., Li, J., Tang, S., Zhang, H., Zhuang, Y.: Anyedit: Mastering unified high-quality image editing for any idea. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 26125–26135 (2025)

82. Zhang, K., Mo, L., Chen, W., Sun, H., Su, Y.: Magicbrush: A manually annotated dataset for instruction-guided image editing. Advances in Neural Information Processing Systems **36**, 31428–31449 (2023)
83. Zhang, Z., Xie, J., Lu, Y., Yang, Z., Yang, Y.: In-context edit: Enabling instructional image editing with in-context generation in large scale diffusion transformer. arXiv preprint arXiv:2504.20690 (2025)
84. Zhao, C., Song, Y., Chen, J., Rong, K., Feng, H., Zhang, G., Ji, S., Wang, J., Ding, E., Sun, Y.: Octopus: A multi-modal llm with parallel recognition and sequential understanding. Advances in Neural Information Processing Systems **37**, 90009–90029 (2024)
85. Zhou, C., Yu, L., Babu, A., Tirumala, K., Yasunaga, M., Shamis, L., Kahn, J., Ma, X., Zettlemoyer, L., Levy, O.: Transfusion: Predict the next token and diffuse images with one multi-modal model. arXiv preprint arxiv:2408.11039 (2024)