HIERAQUERY: BRIDGING MULTIMODAL UNDERSTANDING AND HIGH-QUALITY GENERATION THROUGH MULTI-SCALE QUERY LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Unified multi-modal LLMs enable the integration of visual understanding and generation in a single framework. Recent study shows that a set of learnable queries can serve as an effective interface between autoregressive multimodal LLMs and diffusion models, though the visual quality of generated images still lag behind dedicated generation models. Actually, it is hard for a single set of learnable queries to generate accurate visual representations in a single round of inference. Hence, we introduce `HieraQuery`, which leverages a hierarchy of learnable visual queries to generate high-quality visual contents in a coarse-to-fine manner. Specifically, several sets of learnable queries are provided to the language model, where preceding ones are used to generate images of lower resolution, focusing on the global structures of the generated content, while the subsequent ones serve as the condition for generating higher resolution images, concentrating on the fine-grained details. In addition, a multi-scale representation alignment strategy is proposed to enforce cross-scale consistency and accelerate convergence. Ablation analyses demonstrate that using the hierarchical visual queries can effectively improve the visual generation capability of unified multi-modal LLMs, and scaling up the number of scales proves an effective way for further improving the generation quality.

1 INTRODUCTION

The release of GPT-4o (OpenAI, 2025) in March 2025, which introduced native image generation, leads to prompt attention to unified models for understanding and generation (Tong et al., 2024; Team, 2024b; Pan et al., 2025). Users can now perform complex visual tasks such as image editing (Gong et al., 2024a; Feng et al., 2024), multi-view synthesis (Shi et al., 2023), style transfer (Li et al., 2024b), and even 3D rendering (Mildenhall et al., 2021) purely through natural language conversations. These capabilities are readily accomplished by specialized models (Gong et al., 2024b; Wang et al., 2024a; Huang et al., 2024a; Tan et al., 2024a), marking a major advance in vision-language intelligence. A major challenge in unifying multimodal understanding and generation lies in the inconsistency of visual feature spaces. Recent models such as TokenFlow (Qu et al., 2024) and Janus (Wu et al., 2024a) integrate diffusion-based decoders with token-based understanding models, achieving strong image generation but often at the cost of less precise understanding.

One crucial problem solved recently for image generation in unified autoregressive models is error accumulation. To tackle this issue, a set of learnable queries is employed to generate all continuous conditions simultaneously (Zhang et al., 2025; Pan et al., 2025). However, achieving high-quality generation and editing of visual content demands accurate generation of these continuous condition tokens, which remains difficult for models that produce all tokens at a single inference step.

Inspired by Chain-of-Sight (Huang et al., 2024b) and VAR (Tian et al., 2024), which exploit multi-scale visual structures to capture details at varying spatial scales for understanding and generation tasks, we propose `HieraQuery`, a unified autoregressive model capable of both visual understanding and generation at high precision. We present two key designs to effectively leverage multi-scale hierarchy for visual generation. The first is the multi-scale query learning. Several sets of learnable query tokens are sequentially connected and fed to the language model. The corresponding outputs

Figure 1: **The output results and multimodal interactive demos of `HieraQuery`**. Our model supports text-to-image generation, image editing, and image style transfer following textual instructions.

act as the condition for generating images at different resolutions. The query tokens are hierarchically structured from coarse to fine, with preceding tokens guiding the generation of lower-resolution images and the subsequent tokens refining details for higher-resolution outputs. The second component is the multi-scale representation alignment, where we align the intermediate representations of the diffusion transformer to the latent space of the same visual encoder to encourage cross-scale consistency and accelerate training.

Building on top of the conceptual idea of MetaQuery (Pan et al., 2025), we find `HieraQuery` resolves the previous limitation where an increase in the number of tokens did not notably enhance the visual quality in generation. A consistent improvement in the generation quality is observed when we gradually increase the number of queries with increased number of scales. Ablation analyses shows that the incoporation of `HieraQuery` demonstrates a 10pt improvement in the overall GenEval score when compared to the single scale baseline, while maintaining state-of-the-art understanding capabilites of the base MLLM. Additionally, it is demonstrated that the multi-scale representation alignment strategy notably improves the consistency between the generated images between different scales. Finally, we show `HieraQuery` is also capable of image editing with strong control fluency and contextual understanding. Some of the results are shown in Fig. 1.

2 Related Work

Multimodal Understanding Models Large language models (LLMs) leveraging pre-trained transformer architectures (Vaswani et al., 2017) have profoundly transformed natural language processing. The multi-modal systems for visual recognition are further reshaped by integrating with LLMs, where three core architectural components are typically identified: (1) a visual encoder for modality-specific feature extraction (Radford et al., 2021; Zhai et al., 2023; Dehghani et al., 2024), (2) a pre-trained LLM backbone for language understanding (Dubey et al., 2024; Bai et al., 2023; Brown et al., 2020), and (3) adaptive connector modules facilitating cross-modal alignment. Current implementations primarily diverge in the connector designs – early approaches like BLIP employ learnable query tokens for cross-modal interaction (Dai et al., 2023), while later frameworks such as LLaVA (Liu et al., 2024) adopt streamlined linear projection layers. Beyond static image comprehension, contemporary multimodal systems incorporate dynamic visual (video) and auditory modalities through unified sequence modeling approaches, evolving into omni-MLLM architectures (Bai et al., 2025; Guo et al., 2025) capable of processing heterogeneous input combinations.

Unified Multimodal Large Language Models Both visual recognition and generation have benefited from adopting LLMs, thus it is very appealing to explore how to accomplish cross-modal understanding and generation in a unified framework. Pioneering works like Chameleon (Team, 2024a), Show-o (Xie et al., 2024b) and Emu3 (Wang et al., 2024b) have tried to directly adopt the unified VQ tokenizer to encode images for both multimodal understanding and generation. Since

VQ tokenizer usually cannot maintain high understanding performance of MLLMs due to the information loss during quantization, the Janus series (Wu et al., 2024b; Chen et al., 2025b) use separate encoders for understanding and generation. On the other hand, VILA-U (Wu et al., 2024c) and MUSE-VL (Xie et al., 2024c) try to align VQ-tokenizer with continuous VLM features, thereby improving the understanding performance of the model.

Another way to empower language models with image generation ability is to integrate MLLMs with diffusion models. For example, Emu (Sun et al., 2023) uses the LLM output as a condition for the pre-trained diffusion model, Transfusion (Zhou et al., 2024) train a single transformer by combining the cross-entropy loss with diffusion. Recent works further use learnable query tokens to extract semantic information from MLLM and use them as conditions for the diffusion model (Pan et al., 2025; Zhang et al., 2025). This approach enables the framework to have the ability to generate, edit, and communicate while maintaining the understanding performance of MLLM. However, such a framework requires the model to generate highly precise image representations using a single set of learnable queries, which is quite challenging. This proves to be the main bottleneck in further scaling up the generation quality (Pan et al., 2025), as the generation performance rapidly plateaus with the increasing number of learnable queries. Hence, in this work, we aim to make the generation of precise image representations easier, by introducing a hierarchy of visual query tokens that generates the image in a coarse-to-fine manner, thus bringing even more vitality to the community developing unified multi-modal LLMs.

3 Approach

`HieraQuery` compresses image representations into a sequence of continuous tokens, which are combined with discrete text tokens and further processed by a scaled auto-regressive Transformer (Bai et al., 2025; Lu et al., 2024) for end-to-end multimodal context learning. The general framework of `HieraQuery` follows MetaQuery (Pan et al., 2025), where the generation capability is provided by an external trainable diffusion model, conditioned on tokens produced by auto-regressive Transformers.

To address the difficulty of generating accurate visual representations with a single set of learnable queries in a single inference step, we present two key modifications in our `HieraQuery`. The first is the multi-scale query learning (Sec. 3.1), which enables the generation of visual content from a coarse global structures to fine-grained details. The second is the multi-scale representation alignment strategy (Sec. 3.2) for maintaining the cross-scale consistency between generated images of different resolutions. The overall framework of `HieraQuery` is presented in Fig. 2.

3.1 Multi-Scale Query Learning

To overcome the difficulty of directly generating accurate visual conditions for the image at the target resolution, we present multi-scale learnable query tokens in replacement of the learnable queries in (Pan et al., 2025). The multi-scale query tokens are structured from a coarse-to-fine order, where preceding queries are used as conditions for generating low-resolution images and the subsequent ones are used for generating high-resolution images.

Multi-Scale Learnable Tokens Construction Given an input image x, we define a set of scales $\mathcal{S} = \{s_1, s_2, \ldots, s_K\}$, where each s_k corresponds to a spatial resolution, e.g., $s_k \in \{4\times4, 8\times8, 16\times16\}$. Each scale s_k is associated with a dedicated set of learnable query tokens $Q_{s_k} \in \mathbb{R}^{N_{s_k}\times d}$, where N_{s_k} is the number of tokens for scale s_k, and d is the hidden dimension size. Formally, we initialize the multi-scale query tokens as:

$$Q = \{Q_{s_1}, Q_{s_2}, \ldots, Q_{s_K}\}, \quad Q_{s_k} = \text{Learnable Parameters}. \tag{1}$$

Each Q_{s_k} is designed to produce image contents at different granularities. The preceding ones focuses more on the global layout, color distributions and major objects, while the subsequent ones encodes fine-grained textures and detailed patterns. Given that modeling image content becomes progressively more challenging at higher granularities - it's generally easier to model lower resolution images compared to higher resolution ones – the quantity of learnable queries we consider rises as the resolution itself increases.

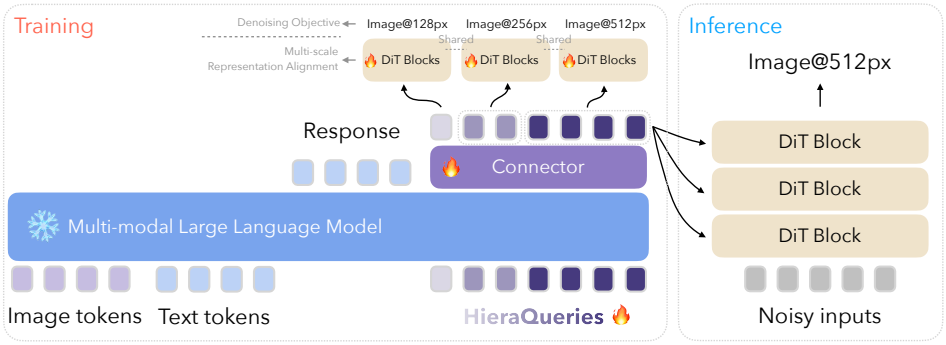

Figure 2: **The framework of `HieraQuery`.** Our model leverages multi-scale learnable query tokens for generating images at different resolutions, greatly reducing the difficulty at modeling high resolution images directly in a single inference step. Further, multi-scale representation alignment is introduced for cross-scale consistency between images of different resolutions.

Multi-Scale Learnable Tokens Fusion and Processing To preserve scale-specific semantics, we introduce explicit scale boundary markers. For each scale s_k, we prepend and append special tokens:

$$\text{Input}_{s_k} = [\text{START}_{s_k}, Q_{s_k}, \text{END}_{s_k}], \tag{2}$$

where START_{s_k} and END_{s_k} are learnable tokens indicating the boundaries of s_k. Each query token also receives a dedicated positional encoding. Let $\mathbf{P}_{s_k} \in \mathbb{R}^{N_{s_k} \times d}$ denote the positional grid encoding for scale s_k, constructed according to the spatial grid associated with that resolution. The complete multi-scale token sequence fed into the language model is thus:

$$\mathbf{Z}_{\text{input}} = \text{Concat}\left(\text{Input}_{s_1}, \text{Input}_{s_2}, \ldots, \text{Input}_{s_K}\right) + \text{Concat}\left(\mathbf{P}_{s_1}, \mathbf{P}_{s_2}, \ldots, \mathbf{P}_{s_K}\right). \tag{3}$$

The language model $f_\theta(\cdot)$ processes $\mathbf{Z}_{\text{input}}$ to produce the hidden representations $\mathbf{H}$. The hidden representations are then provided to a trainable connector layer $g_\theta(\cdot)$ for fine-tuning the visual representations before feeding the output representations as the condition to diffusion transformers, which produces images of different resolutions $\hat{\mathbf{X}}_{\mathbf{s_1}}, \hat{\mathbf{X}}_{\mathbf{s_2}}, \hat{\mathbf{X}}_{\mathbf{s_3}}$:

$$\mathbf{H} = f_\theta(\mathbf{Z}_{\text{input}}), \tag{4}$$

$$\hat{\mathbf{X}}_{\mathbf{s_1}}, \hat{\mathbf{X}}_{\mathbf{s_2}}, \hat{\mathbf{X}}_{\mathbf{s_3}} = \text{DiT}(g_\theta(\mathbf{H})). \tag{5}$$

3.2 Multi-Scale Representation Alignment

To maintain the consistency between generate images of different resolutions, we introduce a simple yet effective Multi-Scale Representation Alignment strategy. This strategy leverages the intrinsic capabilities of unified multi-modal models, which include a visual encoder transforming image pixels into semantically rich representations. Specifically, our approach mirrors (Yu et al., 2024a), by aligning the intermediate hidden states extracted from the DiT backbone at various resolutions with the corresponding outputs of the visual encoder within the MLLM framework. However, due to the potential mismatches between the input image resolution and the DiT backbone resolution, we employ interpolation at the feature level to adjust the shape of the semantic representations to match that of the DiT hidden states prior to the alignment process.

4 Experimental Setup

4.1 Training Data

Basic Image-Text Pairs We use part of LAION-5B (Schuhmann et al., 2022) (372K samples), Wukong (Gu et al., 2022), Midjourney, Blip3o (Chen et al., 2025a), and other datasets that commonly used for diffusion model training. We apply aspect ratio (≤ 2.5), watermark detection (≤ 0.5), and CLIP alignment (≥ 0.45) thresholds for filtering. After preprocessing, the final sample counts are 4M (Wukong), 5M (Midjourney), 3M(Blip3o), and 2M (others).

Image Generation Datasets This part of training data includes InstructPix2Pix-clip-filtered (Brooks et al., 2023), where each pair of edited images is generated 100 times, and the best examples are chosem based on CLIP metrics (Sec.3.1.2 in InstructPix2Pix), SEED-Data-Edit-part2/3 (Ge et al., 2024), excluding part1 due to the poor visual quality, Ultra-edit (Zhao et al., 2024), SynCD (Kumari et al., 2025), Subjects200k (Tan et al., 2024b), HQ-edit (Hui et al., 2024), and MagicBrush (Zhang et al., 2023). It consists of 5,008,795 samples. In addition, our training data includes publicly available datasets commonly used for style transfer tasks, along with synthesized data generated using style prompts. The style data comprises a high-quality subset of WikiArt, covering 27 painting styles such as Impressionism, Realism, and Expressionism, and the StyleBooth, featuring 67 styles including cartoon and 3D, each with 717 image pairs. These two datasets contain 81,444 and 80,922 samples, respectively. We selected a total of 2 million relatively high-quality samples from these data sets for training

4.2 Implementation Details

Modules Based on the previous experience of MetaQuery (Pan et al., 2025), we continue to build generation capabilities on the powerful MLLM. Our experiments are based on Ming-Lite-Omni (InclusionAI, 2025), which is an open-source MoE-based Multimodal LLM, processing audio, video, image, and text inputs to generate multimodal outputs via balanced training strategies, maintaining robust text performance to advance omni-MLLM development. Following (Pan et al., 2025), we adopt Qwen2.5-0.5B (Qwen et al., 2025) as our connector, and initialize the trainable diffusion model with SD3-Medium (Esser et al., 2024) / SANA-1.6B (Xie et al., 2024a). Unless otherwise specified, the diffusion model is initialized using SD3-Medium. In order to make a fair comparison with the existing papers, we also re-trained with the SANA-1.6B initialization. Wherever additional transformations of the hidden-states dimension are required, we use a two-layer multilayer perceptron.

Training HieraQuery training process includes three stages: text-to-image pretraining, image reconstruction training, and image editing fine-tuning. Image reconstruction training is a preparation for editing training. The purpose is to make the output of the model consistent with the input, and use the reconstruction dataset converted from image-text pairs for training. Each stage is performed on the training dataset for two epochs, with the cosine learning rate strategy and warmup, and the learning rate increases from 1e-6 to 5e-5 and then decreases to 1e-6. The parameters of MLLM are frozen during the whole process, and all other parameters are trainable, and there is no special learning rate setting.

5 Experiments

5.1 Text-to-Image Generation

We conduct separate quantitative evaluations of `HieraQuery` on multimodal understanding and generation using public benchmarks. For multimodal generation, we evaluate text-to-image performance by FID score (Heusel et al., 2017) on MJHQ-30K (Li et al., 2024a) for visual aesthetic quality, and GenEval (Ghosh et al., 2024) and DPG-Bench (Hu et al., 2024) (both without prompt rewriting) for prompt alignment, respectively.

As shown in Tab. 1, our `HieraQuery` obtains a state-of-the-art 0.87 overall accuracy on GenEval without using any LLM rewriter. It can be seen that with a similar level of understanding performance, `HieraQuery` outperforms all other unified models in image generation on MJHQ-FID, GenEval and DPG-bench. Quantitative comparison is shown in Figure 3.

5.2 Understanding

We compare our model with existing approaches on understanding benchmarsk in Tab. 1, including MMB (Liu et al., 2025b), MMMU (Yue et al., 2024), AI2D (Kembhavi et al., 2016), and MM-Vet (Yu et al., 2024b). Our model demonstrates competitive performance among models of similar size, highlighting its robust capabilities in image-text understanding tasks. Furthermore, our model exhibits competitive performance among models of similar size, showcasing its robust capabilities in image-text understanding tasks.

Table 1: Quantitative results on generation benchmarks and parts of OpenCompass (Contributors, 2023) multimodal leaderboard. ‡ denotes closed-source models. "Und." and "Gen." denote "understanding" and "generation", respectively. † refers to the methods using LLM rewriter.

Type	Model	MMB ↑	MMMU↑	AI2D ↑	MM-Vet ↑	MJHQ FID ↓	GenEval ↑	DPG-Bench ↑
Und.	LLaVA-72B (Xie et al., 2024b)	84.5	56.6	86.2	60.6	×	×	×
	Qwen2-VL-72B (Bai et al., 2023)	85.9	64.3	88.3	73.9	×	×	×
	Qwen2.5-VL-7B (Bai et al., 2025)	87.8	67.9	88.2	76.7	×	×	×
	Emu3-Chat (Wang et al., 2024c)	58.5	31.6	-	37.2	×	×	×
	InternVL2.5-8B (Chen et al., 2024)	82	54.8	84.5	68.1	×	×	×
	DeepSeek-VL2 (Wu et al., 2024d)	81.2	50.7	84.5	60.0	×	×	×
	GPT-4o-20241120‡ (OpenAI, 2024)	84.3	70.7	84.9	74.5	×	×	×
	Step-1o‡ (StepFun, 2025)	87.3	69.9	89.1	82.8	×	×	×
Gen.	LlamaGen (Sun et al., 2024)	×	×	×	×	-	0.32	-
	LDM (Rombach et al., 2022)	×	×	×	×	-	0.37	-
	SDv1.5 (Rombach et al., 2022)	×	×	×	×	-	0.43	-
	PixArt-α (Chen et al., 2023)	×	×	×	×	6.14	0.48	71.6
	SDv2.1 (Rombach et al., 2022)	×	×	×	×	-	0.50	-
	DALL-E 2 (Ramesh et al., 2022)	×	×	×	×	-	0.52	-
	Emu3-Gen (Wang et al., 2024c)	×	×	×	×	-	0.54	-
	SDXL (Podell et al., 2023)	×	×	×	×	6.63	0.55	74.7
	DALL-E 3 (Betker et al., 2023)	×	×	×	×	-	0.67	83.5
	SD3-Medium (Esser et al., 2024)	×	×	×	×	11.92	0.74	84.1
Unified	DreamLLM (Dong et al., 2023)	-	-	-	36.6	-	-	-
	MetaMorph (Tong et al., 2024)	75.2	-	-	-	-	-	-
	Show-o (Xie et al., 2024b)	-	26.7	-	-	15.18	0.53	-
	TokenFlow-XL (Qu et al., 2024)	68.9	38.7	-	40.7	-	0.55	73.3
	Chameleon (Team, 2024b)	-	22.4	-	8.3	-	0.39	-
	Janus (Wu et al., 2024a)	69.4	30.5	-	34.3	10.10	0.61	-
	JanusFlow (Ma et al., 2024)	74.9	29.3	-	30.9	9.51	0.63	80.09
	Janus-Pro-1B (Chen et al., 2025b)	75.5	36.3	-	39.8	14.33	0.73	82.6
	Metaquery-XL (Pan et al., 2025)	83.5	58.6	-	66.6	6.02	0.80†	82.0
	OmniGen2 (Wu et al., 2025)	79.1	53.1	-	61.8	-	0.80	83.6
	Blip3-o (Chen et al., 2025a)	83.5	58.6	-	66.6	-	0.84	81.6
	BAGEL (Deng et al., 2025)	85.0	55.3	-	67.1	-	0.82	-
	UniWorld-V1 (Lin et al., 2025)	83.5	58.6	-	67.1	-	0.80	-
	Ours (`HieraQuery`)	80.7	54.3	84.9	74.0	**4.85**	**0.87**	**84.2**

Table 2: Quantitative evaluation on GEdit-Bench-EN. All metrics are reported as higher-is-better (↑). We report the results evaluated by GPT-4.1. The Intersection subset reflects the subset of prompts where all methods return valid responses with a total of 434 instances; the Full set includes all the 606 instances.

Model	GEdit-Bench-EN (Intersection subset) ↑			GEdit-Bench-EN (Full set) ↑		
	G_SC	G_PQ	G_O	G_SC	G_PQ	G_O
Instruct-Pix2Pix (Brooks et al., 2023)	3.473	5.601	3.631	3.575	5.491	3.684
MagicBrush (Zhang et al., 2023)	4.646	5.800	4.578	4.677	5.656	4.518
AnyEdit (Yu et al., 2025)	3.177	5.856	3.231	3.178	5.820	3.212
OmniGen (Xiao et al., 2025)	6.070	5.885	5.162	5.963	5.888	5.061
Step1X-Edit (Liu et al., 2025a)	7.183	6.818	6.813	7.091	6.763	**6.701**
Ours(`HieraQuery`)	**7.633**	**7.097**	**6.849**	**7.357**	**7.102**	6.616
Gemini (Gemini2, 2025)	6.697	6.638	6.322	6.732	6.606	6.315
GPT-4o (OpenAI, 2025)	7.844	7.592	7.517	7.850	7.620	7.534

5.3 Instruction Based Image Editing

As shown in Tab. 2, we use GEdit-Bench to quantitatively test the model's response to various editing instructions. As shown in Tab. 2, HieraQuery excels a wider range of interactive image editing tasks, including style transfer and object addition, deletion, and modification, outperforms existing open-source models in both SQ (Semantic Consistency), PQ (Perceptual Quality), and O (Overall Score). When editing human figures, it demonstrates a clear advantage in maintaining scene and character ID (As shown in Fig. 4).

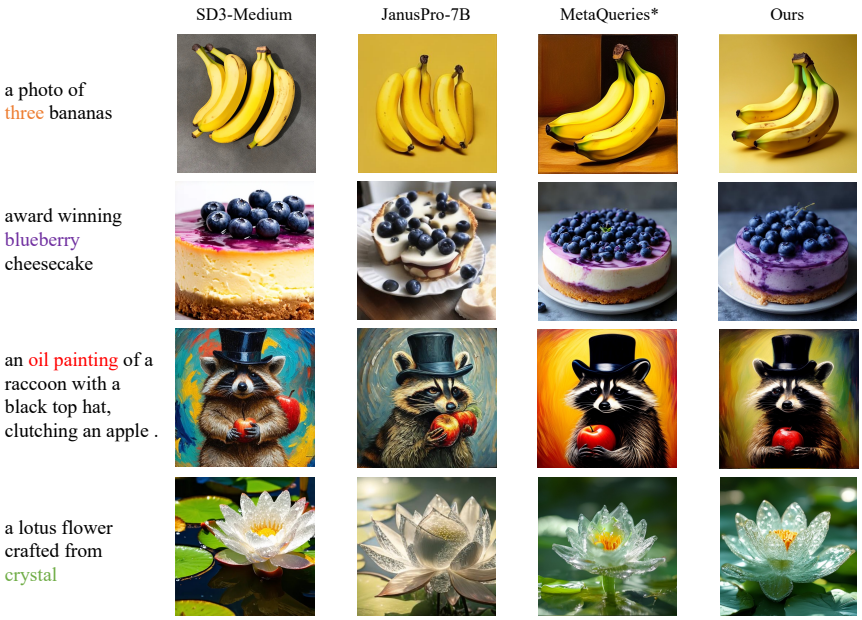

Figure 3: Quantitative comparison on text-to-image generation. Our method excels at prompt following, detail depiction, style accuracy, and subject integrity. * denotes models that are currently not publicly available and are therefore reproduced by us.

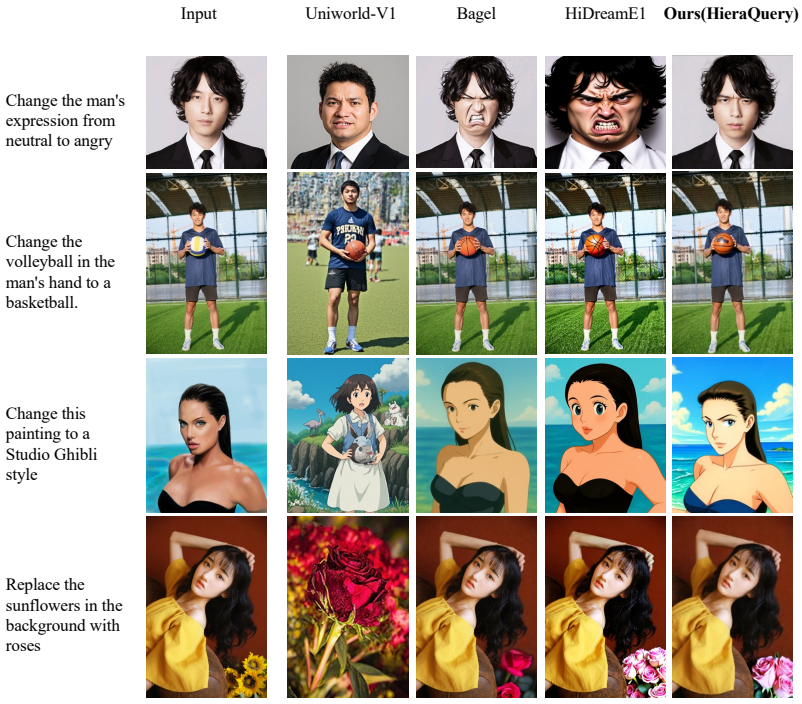

Figure 4: Illustration of image editing results compared to recent approaches such as Uniworld-V1 (Lin et al., 2025), Bagel (Deng et al., 2025), and HiDreamE1 (Cai et al., 2025).

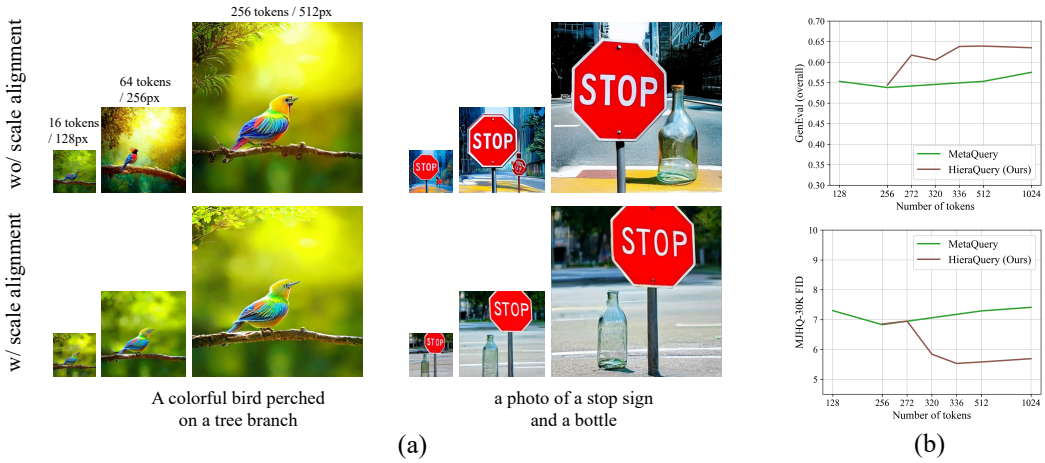

Figure 5: (a) Qualitative comparison of the ablation of multi-scale representation alignment method. The proposed representation alignment method can help tokens of different scales represent similar images, thereby coarse-to-fine improving the final high-resolution visual effect. (b) Study on the scaling of token numbers. As the number of tokens increases, our multi-scale learnable queries setting increases faster than single-scale in image generation performance.

Table 3: Ablation studies on the proposed motheds. "M" denotes *Multi-scale Query Learning*, "A" denotes *Multi-Scale Representation Alignment*. We report all aspects in GenEval (%).

M.	A.	Num tokens	Single Obj.↑	Two Obj.↑	Counting↑	Colors↑	Position↑	Color Attri.↑	Overall↑
×	×	256	96.8	55.8	40.6	82.7	18.5	32.0	54.4
×	×	336	96.4	56.4	42.2	82.6	18.5	32.3	54.7
✓	×	336	**99.0**	75.7	53.4	**86.9**	26.2	29.7	61.8
✓	✓	336	**99.0**	**76.7**	**58.7**	84.3	**27.0**	**33.2**	**63.2**

6 Ablations and Analysis

We perform comprehensive ablation studies to validate the effectiveness of multi-scale token learning and multi-scale representation alignment, as well as the designs of the multi-scale learnable tokens. To control the training budget, the ablation studies are conducted on a small-scale image generation dataset (with approximately 5 million image text pairs), and a small DiT structure (*i.e.,* SANA-1.6B (Xie et al., 2024a)).

6.1 Ablation on Multi-scale Leanable Tokens

We first conducted an ablation study on multi-scale learnable tokens. According to Table 3, Compared with Single-scale as baseline, we found that compared with Multi-scale, it improved by 8.4% on GenEval, 2.2% on Single Object and 12.8% on Counting, showing better generated image quality and prompt following ability.

6.2 Ablation on Multi-scale Representation Alignment

Based on multi-scale learnable tokens, we further conducted ablation experiments on multi-scale representation alignment. According to Table 3, we can find that representation alignment plays a role in further improving the generation effect. Multi-scale representation alignment improves the overall effect of the model by 1.5% under the GenEval measurement. A qualitative comparison is shown in Figure 5(a).

Table 4: Ablation studies on combinations of multiple scales and the number of overall tokens.

Scales Combination	Aspect Ratio	Number of Queries	**MJHQ FID)↓**	**GenEval (Overall)↑**
1.0x	1:1	256	6.87	0.54
1.0x	any	256	6.85	0.57
0.25x, 1.0x	any	272	6.95	0.61
0.5x, 1.0x	any	320	5.84	0.60
0.25x, 0.5x, 1.0x	any	336	**5.53**	**0.64**
0.25x, 0.5x, 1.0x	any	512	5.58	**0.64**
0.25x, 0.5x, 1.0x	any	1024	5.69	0.63

6.3 Ablation on designs of Multi-scale Leanable tokens

We then conduct an experimental study on the combination of different scales. We start with a single scale (1.0x) and gradually add more and smaller scales (0.5x, 0.25x). And to further verify that our method is valid for more aspect ratios, we first expand the fixed 512x512 generation size to the native aspect ratio generation sizes, which is achieved by smart resizing (Bai et al., 2025) and constructing ratio buckets.

As shown in Table 4, adding a smaller resolution at the front position can significantly improve the generated quality. Adding 0.25x and 0.5x scales alone can bring 3% to 4% improvement, while adding them together can further bring a slight improvement.

The proposed multi-scale method actually increases the number of learnable tokens. To study the relationship between the number of tokens and the generation effect, we compared the experimental results with those of the single-scale method (Pan et al., 2025), as shown in Figure 5(b). We can see that in the existing single-scale learnable tokens method, the generation quality does not continue to improve effectively as the number of tokens increases (after exceeding 256). In contrast, multi-scale learnable tokens can achieve better generation results by introducing only a small number of additional tokens. The above experiments show that the multi-scale method can make tokens be used more efficiently as a bridge between understanding and generation.

7 Discussion

Limitations Although the proposed method has greatly improved the quality of text-to-image generation, it is limited by the fact that the editing datasets are currently open source datasets and some datasets are not of high quality, so the editing ability may be limited in some cases. This may be further improved by increasing the number of datasets and filtering criteria.

Conclusions In this paper, we mainly explore the role of multi-scale query learning in unified MLLM. Based on the existing method of building unified MLLM with leanable tokens, we proposed grouped tokens to generate images progressively from low resolution to high resolution, and improved the consistency of generated multiscale images by aligning the middle layer features of the multi-scale diffusion model. After detailed experiments, we verified that this token arrangement can efficiently extract the information required for generation from MLLM. We verified the effectiveness of the method in text-to-image task, image style transfer, and fine-grained editing, while maintaining the understanding performance of the original MLLM unchanged.

REFERENCES

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

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

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*, pp. 18392–18402, 2023.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, 2020.

Qi Cai, Jingwen Chen, Yang Chen, Yehao Li, Fuchen Long, Yingwei Pan, Zhaofan Qiu, Yiheng Zhang, Fengbin Gao, Peihan Xu, et al. Hidream-i1: A high-efficient image generative foundation model with sparse diffusion transformer. *arXiv preprint arXiv:2505.22705*, 2025.

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

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.

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*, 2025b.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

OpenCompass Contributors. Opencompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass, 2023.

W Dai, J Li, D Li, AMH Tiong, J Zhao, W Wang, B Li, P Fung, and S Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. arxiv 2023. *arXiv preprint arXiv:2305.06500*, 2, 2023.

Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim M Alabdulmohsin, et al. Patch n'pack: Navit, a vision transformer for any aspect ratio and resolution. *Advances in Neural Information Processing Systems*, 36, 2024.

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.

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.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. URL https://arxiv.org/abs/2403.03206.

Yutong Feng, Biao Gong, Di Chen, Yujun Shen, Yu Liu, and Jingren Zhou. Ranni: Taming text-to-image diffusion for accurate instruction following. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4744–4753, 2024.

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.

Google Gemini2. Experiment with gemini 2.0 flash native image generation, 2025.

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, 2024.

Biao Gong, Siteng Huang, Yutong Feng, Shiwei Zhang, Yuyuan Li, and Yu Liu. Check locate rectify: A training-free layout calibration system for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6624–6634, 2024a.

Biao Gong, Shuai Tan, Yutong Feng, Xiaoying Xie, Yuyuan Li, Chaochao Chen, Kecheng Zheng, Yujun Shen, and Deli Zhao. Uknow: A unified knowledge protocol with multimodal knowledge graph datasets for reasoning and vision-language pre-training. *Advances in Neural Information Processing Systems*, 37:9612–9633, 2024b.

Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Niu Minzhe, Xiaodan Liang, Lewei Yao, Runhui Huang, Wei Zhang, Xin Jiang, et al. Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. *Advances in Neural Information Processing Systems*, 35:26418–26431, 2022.

Qingpei Guo, Kaiyou Song, Zipeng Feng, Ziping Ma, Qinglong Zhang, Sirui Gao, Xuzheng Yu, Yunxiao Sun, Tai-Wei Chang, Jingdong Chen, et al. M2-omni: Advancing omni-mllm for comprehensive modality support with competitive performance. *arXiv preprint arXiv:2502.18778*, 2025.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

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

Siteng Huang, Biao Gong, Yutong Feng, Xi Chen, Yuqian Fu, Yu Liu, and Donglin Wang. Learning disentangled identifiers for action-customized text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7797–7806, 2024a.

Ziyuan Huang, Kaixiang Ji, Biao Gong, Zhiwu Qing, Qinglong Zhang, Kecheng Zheng, Jian Wang, Jingdong Chen, and Ming Yang. Accelerating pre-training of multimodal llms via chain-of-sight. *Advances in Neural Information Processing Systems*, 37:75668–75691, 2024b.

Mude Hui, Siwei Yang, Bingchen Zhao, Yichun Shi, Heng Wang, Peng Wang, Yuyin Zhou, and Cihang Xie. Hq-edit: A high-quality dataset for instruction-based image editing. *arXiv preprint arXiv:2404.09990*, 2024.

InclusionAI. Ming-lite-omni. https://github.com/inclusionAI/Ming/, 2025.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images, 2016.

Nupur Kumari, Xi Yin, Jun-Yan Zhu, Ishan Misra, and Samaneh Azadi. Generating multi-image synthetic data for text-to-image customization. *arXiv preprint arXiv:2502.01720*, 2025.

Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint arXiv:2402.17245*, 2024a.

Wen Li, Muyuan Fang, Cheng Zou, Biao Gong, Ruobing Zheng, Meng Wang, Jingdong Chen, and Ming Yang. Styletokenizer: Defining image style by a single instance for controlling diffusion models. In *European Conference on Computer Vision*, pp. 110–126. Springer, 2024b.

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.

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

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*, 2025a.

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*, pp. 216–233. Springer, 2025b.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.

Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai yu, Liang Zhao, Yisong Wang, Jiaying Liu, and Chong Ruan. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation, 2024.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

OpenAI. Gpt-4o system card, 2024. URL `https://arxiv.org/abs/2410.21276`.

OpenAI. Introducing 4o image generation. `https://openai.com/index/introducing-4o-image-generation/`, 2025.

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.

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.

Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K Du, Zehuan Yuan, and Xinglong Wu. Tokenflow: Unified image tokenizer for multimodal understanding and generation. *arXiv preprint arXiv:2412.03069*, 2024.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan,

Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

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. *arXiv preprint arXiv:2103.00020*, 2021.

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.

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*, pp. 10684–10695, 2022.

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. *arXiv preprint arXiv:2210.08402*, 2022.

Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023.

StepFun. step-1o. https://www.stepfun.com/, 2025.

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

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*, 2023.

Shuai Tan, Biao Gong, Xiang Wang, Shiwei Zhang, Dandan Zheng, Ruobing Zheng, Kecheng Zheng, Jingdong Chen, and Ming Yang. Animate-x: Universal character image animation with enhanced motion representation. *arXiv preprint arXiv:2410.10306*, 2024a.

Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. *arXiv preprint arXiv:2411.15098*, 2024b.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models, 2024a. URL https://arxiv.org/abs/2405.09818.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024b.

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

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017.

Wen Wang, Qiuyu Wang, Kecheng Zheng, Hao Ouyang, Zhekai Chen, Biao Gong, Hao Chen, Yujun Shen, and Chunhua Shen. Framer: Interactive frame interpolation. *arXiv preprint arXiv:2410.18978*, 2024a.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Tao Li, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Tiejun Huang, and Zhongyuan Wang. Emu3: Next-token prediction is all you need, 2024b. URL https://arxiv.org/abs/2409.18869.

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*, 2024c.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. Janus: Decoupling visual encoding for unified multimodal understanding and generation, 2024a. URL https://arxiv.org/abs/2410.13848.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*, 2024b.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, Ze Liu, Ziyi Xia, Chaofan Li, Haoge Deng, Jiahao Wang, Kun Luo, Bo Zhang, Defu Lian, Xinlong Wang, Zhongyuan Wang, Tiejun Huang, and Zheng Liu. Omnigen2: Exploration to advanced multimodal generation, 2025. URL https://arxiv.org/abs/2506.18871.

Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024c.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, and et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding, 2024d. URL https://arxiv.org/abs/2412.10302.

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*, pp. 13294–13304, 2025.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, et al. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. *arXiv preprint arXiv:2410.10629*, 2024a.

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*, 2024b.

Rongchang Xie, Chen Du, Ping Song, and Chang Liu. Muse-vl: Modeling unified vlm through semantic discrete encoding, 2024c. URL https://arxiv.org/abs/2411.17762.

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*, pp. 26125–26135, 2025.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024a.

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. In *International conference on machine learning*. PMLR, 2024b.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*, 2024.

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

Hong Zhang, Zhongjie Duan, Xingjun Wang, Yingda Chen, Yuze Zhao, and Yu Zhang. Nexus-gen: A unified model for image understanding, generation, and editing. *arXiv preprint arXiv:2504.21356*, 2025.

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

Haozhe Zhao, Xiaojian Shawn Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. Ultraedit: Instruction-based fine-grained image editing at scale. *Advances in Neural Information Processing Systems*, 37:3058–3093, 2024.

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