

Unified Multimodal Autoregressive Modeling with Shared Context—Visual Tokenizer is Key to Unification

Wujian Peng^{1,2,*,#}, Lingchen Meng^{3,*,\ddagger}, Yuxuan Cai³, Xianwei Zhuang³,
Yuhuan Yang³, Rongyao Fang³, Chenfei Wu³, Junyang Lin³,
Zuxuan Wu^{1,2,\ddagger}, Shuai Bai^{3,\ddagger}

¹Institute of Trustworthy Embodied AI, Fudan University

²Shanghai Innovation Institute ³Qwen Team, Alibaba Inc.

*Equal Contributions, #Work done during internship at Qwen Team

\ddaggerProject Lead, \ddaggerCorresponding authors

Abstract

Unified Multimodal Modeling aims to integrate visual understanding and generation within a single system. However, existing approaches typically rely on two disparate visual tokenizers, which splits the representation space and hinders truly unified modeling. We propose UNiAR, a unified autoregressive framework where a single discrete visual tokenizer serves as the key bridge between understanding and generation, enabling a **shared context** in which the model can directly interpret its own generated visual tokens without additional re-encoding. UNiAR adapts a pretrained vision encoder with multi-level feature fusion and a lookup-free bitwise quantization scheme, preserving both high-level semantics and low-level details while scaling the effective visual vocabulary at minimal cost. Building on this, the unified autoregressive model adopts **parallel-bitwise-prediction** to jointly predict spatially grouped, multi-level visual codes, substantially reducing visual sequence length and accelerating generation. Finally, a diffusion-based visual decoder operates on discrete visual tokens to decode high-fidelity images. Through large-scale pre-training, followed by supervised fine-tuning and reinforcement learning, UNiAR achieves state-of-the-art performance on image generation and image editing while remaining competitive on multimodal understanding benchmarks.

Website: <https://sharelab-sii.github.io/uniar-web>

1 Introduction

Recent years have witnessed significant progress in visual understanding [1, 22, 43] and generation [11, 34, 60], largely driven by two distinct paradigms: autoregressive modeling and denoising diffusion. Increasingly, Unified Multimodal Models (UMMs) aim to integrate both capabilities into a single system [5, 10, 51, 58], enabling a model to understand and generate visual content in shared context—*i.e.*, to directly interpret its own visual generations without re-encoding. However, this integration is non-trivial due to an inherent tension in visual representation: understanding relies on high-level semantics, whereas generation demands low-level, high-frequency details. As a result, most existing works rely on two separate visual tokenizers [5, 10], which places understanding and generation in different representation spaces. Consequently, generated images must be re-encoded by the understanding tokenizer before the model can interpret them, breaking the desired “shared-context” goal and hindering true unification.

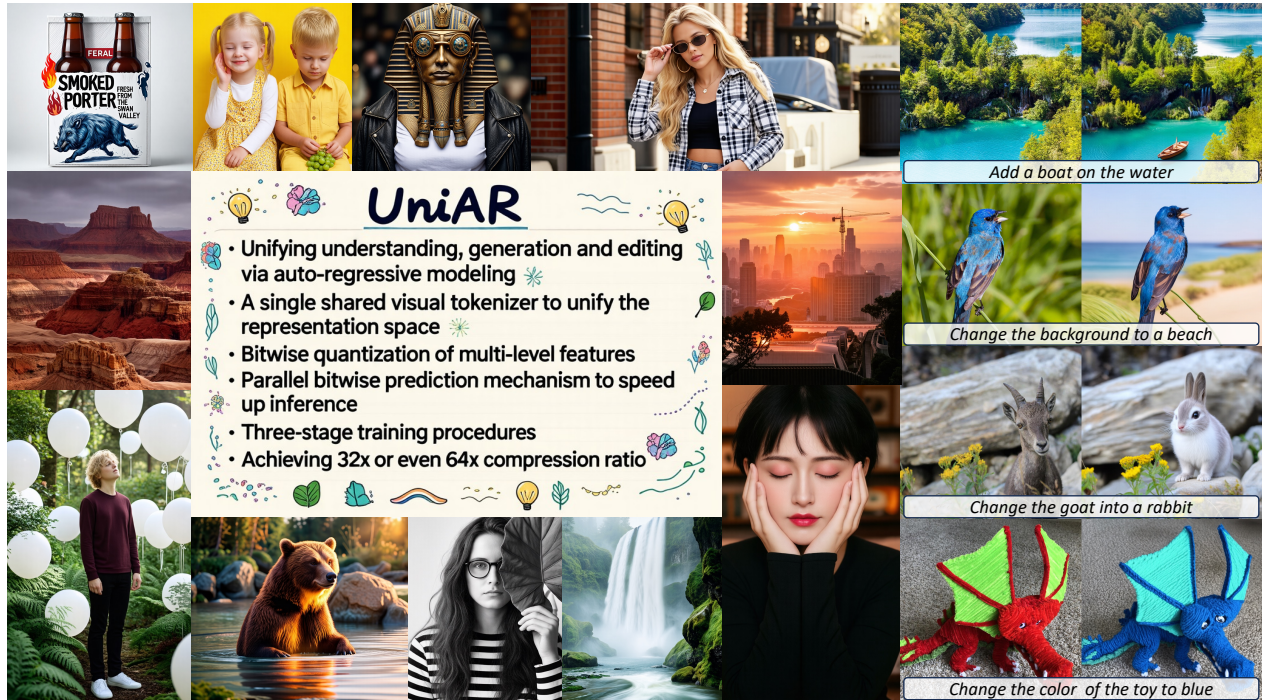


Figure 1 Visual content generated by UniAR. UniAR produces high-fidelity visual content with demonstrated efficacy in instruction-following and text rendering, alongside versatile image editing capabilities.

Using a single tokenizer to unify visual understanding and generation is a promising direction [14, 51, 58]. However, this approach faces non-trivial challenges: (1) how to design visual tokens that simultaneously meet the needs of both generation and understanding, and (2) how to scale up the tokenizer vocabulary at minimal cost. Motivated by these challenges, we propose UniAR, a UNified Auto-REgressive framework that leverages a single visual tokenizer to effectively unify generation and understanding in a shared context. To the best of our knowledge, **UniAR is the first unified model to leverage multi-level bitwise visual tokens, achieving effective and efficient visual understanding and generation.**

Specifically, we fuse multi-level visual features to improve representational capacity, preserving low-level details from shallow layers and high-level semantics from deeper layers, so that the resulting representation supports both discriminative and generative tasks. Moreover, we adopt a lookup-free binary quantization scheme that maps visual features to discrete binary vectors [18, 70]. In contrast to conventional vector quantization [56], this scheme removes the need for an explicit codebook and exponentially scales the theoretical vocabulary size (*e.g.*, a 64-bit vector yields 2^{64} unique codes), substantially expanding the vocabulary with only a small overhead. Building on this discrete visual tokenizer, we use a unified autoregressive model under a next-token prediction paradigm to jointly model understanding, generation, and editing. We further introduce a parallel bitwise prediction mechanism that predicts multi-level visual binary vectors within each 2×2 spatial grid simultaneously, effectively reducing the visual sequence length and achieving $32 \times$ visual compression ratio in autoregressive prediction. Finally, we employ a DiT-based visual decoder with resolution upsampling, which conditions solely on visual tokens (without text prompts) to reconstruct high-fidelity images; when employing upsampling, a 1024×1024 image requires predicting only 256 visual tokens.

We train the autoregressive model via three stages: large-scale pretraining on multimodal corpora, supervised fine-tuning on curated high-quality data, and reinforcement learning to further improve generation performance. Throughout the training process of the autoregressive model, the visual tokenizer and the visual decoder are kept frozen, and the decoder is introduced only during reinforcement learning to decode images

for reward computation. Extensive experiments show that UNiAR delivers state-of-the-art text rendering and instruction-following performance for image generation, while remaining competitive on multimodal understanding. Our contributions are three-fold:

- We propose UNiAR, a unified autoregressive framework with a single discrete visual tokenizer for unified modeling. In particular, multi-level feature fusion bridges the needs of generation and understanding, while lookup-free bitwise quantization scales up the tokenizer vocabulary efficiently.
- We introduce a unified autoregressive modeling paradigm with parallel bitwise prediction, together with a DiT-based decoder equipped with resolution upsampling, enabling efficient high-resolution generation with compact visual sequences.
- Extensive experiments validate the effectiveness of UNiAR, achieving state-of-the-art image generation performance while remaining competitive on standard multimodal understanding benchmarks.

2 Related Works

2.1 Visual Tokenizers

Visual understanding and generation impose distinct requirements on visual tokenization. Understanding tasks typically prioritize high-level semantic representations [6, 45, 55], whereas generative tasks require high-frequency details such as texture and color [7, 56, 57]. Previous works have attempted to accommodate both objectives by integrating dual-tokenizers [5, 53, 54]; however, this inherently partitions the visual context. Several studies have sought to bridge this gap by developing unified tokenizers supporting both objectives simultaneously [17, 35, 50, 71] or dual-codebook design [36, 44]. In this work, we employ a multi-level feature visual tokenizer that integrates features from various ViT layers to satisfy the requirements of both generation and comprehension. We quantize the visual features into discrete binary tokens, enabling autoregressive visual modeling for both tasks. In this respect, UNiAR is closely related to Infinity [18], which also adopts binary visual quantization [70]. However, there are several differences. Infinity is designed specifically for generation, while UNiAR targets unified multimodal modeling. Moreover, Infinity is optimized for reconstruction, whereas UNiAR learns semantic multi-level features that support both generation and understanding. The two methods also differ in modeling paradigm: Infinity adopts next-scale prediction, while UNiAR uses standard next-token prediction for both vision and language.

2.2 Unified Multimodal Models

Existing research explores various strategies for unified modeling. One paradigm [27, 42, 60] decouples tasks by interfacing a cascaded architecture composed of a frozen LMM and a diffusion model. However, these frameworks often lack architectural synergy between the understanding and generation components. Alternatively, another paradigm [10, 33, 37, 63, 72] employs hybrid Transformers that generate text autoregressively and images via flow-matching. However, these approaches typically rely on divergent training objectives or disrupt the inherent causal mechanisms of standard LLMs and LMMs, which increase the training cost. Distinct from these approaches, our UNiAR achieves deep integration by employing a unified visual tokenizer to align the representation space, and modeling both generation and understanding within a unified autoregressive framework.

Among existing works, X-Omni [14] is most closely related to ours. However, **UNiAR distinguishes itself in several dimensions:** (1) Quantization: instead of an explicit codebook, we employ lookup-free bitwise quantization across multi-level features, which expands theoretical codebook size with much lower overhead; (2) Efficiency: we introduce a parallel-bitwise-prediction paradigm, quadrupling the inference speed via reduced autoregressive steps; (3) Lightweight: UNiAR achieves competitive results with significantly fewer parameters, *i.e.* a more compact visual tokenizer (400M vs. 1B) and diffusion decoder (2.5B vs. 12B); and (4) AR-centric: X-Omni’s visual decoder feeds both text prompts and visual features, while our visual decoder is independent of text inputs; it functions purely for image decoding, leaving all semantic and structural generation to the autoregressive model.

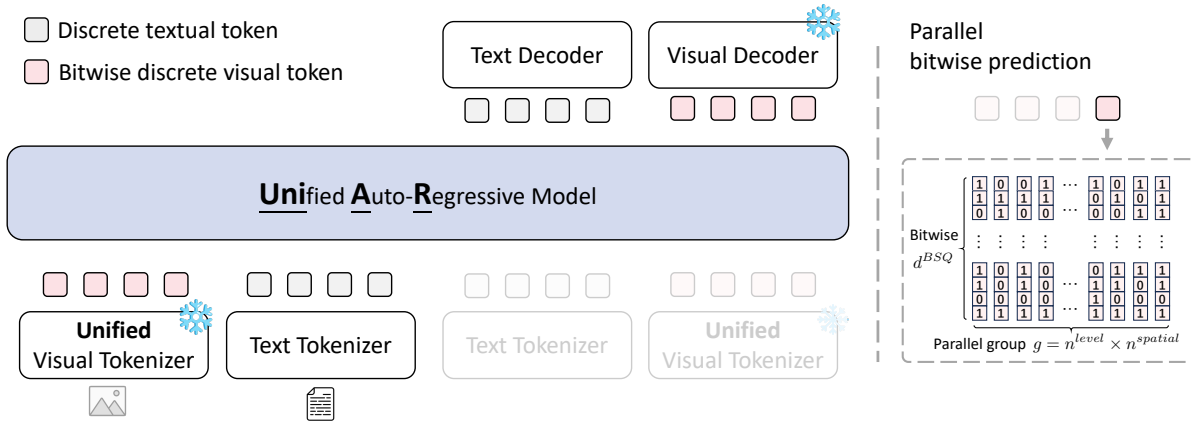


Figure 2 UNiAR Framework Overview. Our architecture integrates a **unified visual tokenizer** for bitwise quantization of image features into discrete semantic tokens, a **unified autoregressive backbone** that unifies generation and understanding via next-token prediction, and a **DiT-based decoder** for high-fidelity image decoding from predicted tokens. For visual generation, the auto-regressive model performs **parallel bitwise prediction** that predicts the next group of bit indices in parallel as shown in the right part. Notably, both the text and visual decoders are only required during inference and are not utilized during the pretraining stage.

3 UNiAR

The overall architecture of UNiAR comprises three core components: (1) a unified visual tokenizer that encodes images into discrete semantic tokens shared across understanding, generation and editing tasks (Sec. 3.1); (2) a unified autoregressive model for the joint prediction of visual and text tokens (Sec. 3.2); and (3) a DiT-based visual decoder that translates visual tokens back into pixel space (Sec. 3.3).

3.1 Unified Visual Tokenizing

To facilitate autoregressive image generation, visual inputs need to be encoded into discrete tokens. Unlike traditional reconstruction-based paradigms [48, 56, 67, 70], UNiAR adapts continuous features from **semantic** vision encoder into discrete representations. Specifically, we integrate Binary Spherical Quantization (BSQ) [70] into the visual encoder of a pre-trained Large Multimodal Model (LMM), and end-to-end finetune the model with visual understanding objectives. Unlike the conventional VQ-VAE [56], which maps visual feature to a specific index in an explicit codebook, BSQ quantizes features as a binary vector $\mathbf{u} \in \{0, 1\}^{d^{BSQ}}$. This approach enables the construction of vastly larger implicit vocabularies with minimal overhead, as the codebook size scales exponentially with the quantization dimension. In our approach, we set the BSQ dimension $d^{BSQ} = 64$, theoretically providing an expansive codebook of 2^{64} unique tokens. The quantization process is formally defined as:

$$\begin{aligned}
 \mathbf{v} &= \text{Encoder}x \\
 \mathbf{u} &= \text{BSQMLP}_{in}\mathbf{v} \\
 \mathbf{v}' &= \text{MLP}_{out}\mathbf{u} \\
 \mathbf{v} &= \text{Mergerv}'
 \end{aligned}
 \left. \vphantom{\begin{aligned} \mathbf{u} &= \text{BSQMLP}_{in}\mathbf{v} \\ \mathbf{v}' &= \text{MLP}_{out}\mathbf{u} \end{aligned}} \right\} \text{BSQ-added parameters}, \quad (1)$$

where x represents the input image and \mathbf{v} denotes the raw features extracted by the vision encoder. MLP_{in} and MLP_{out} refer to projection modules that map visual features to and from the BSQ quantization space, respectively. The binary vector $\mathbf{u} \in \{0, 1\}^{64}$ serves as the discrete bitwise representation, from which \mathbf{v}' is reconstructed. Following the BSQ quantize module, a spatial merger Merger aggregates 2×2 visual features into one token \mathbf{v} , projecting it to the hidden dimension of the LLM.

To further enrich the representation capacity, we adopt a multi-level fusing scheme that aggregates information across multiple layers of the visual encoder. Specifically, in addition to the final layer, we incorporate three

intermediate layers into the quantization process [1, 41]. This hierarchical design facilitates the extraction of multi-granular visual cues for both generation and understanding tasks.

Following the original BSQ formulation, we employ the soft entropy loss for quantization. In contrast, we replace the standard Mean Squared Error (MSE) loss of reconstruction-based decoder with the Cross-Entropy (CE) loss of LMM. Formally, the overall training objective is defined as follows:

$$\mathcal{L} = \mathcal{L}^{\text{CE}} \lambda^{\text{BSQ}} \cdot \mathcal{L}^{\text{BSQ}}, \quad (2)$$

where \mathcal{L}^{CE} is the CE loss of the original LMM and \mathcal{L}^{BSQ} is the original BSQ soft entropy loss. Once the discrete adaptation procedure is complete, the vision encoder is frozen to guarantee consistent visual indices for unified autoregressive modeling.

3.2 Unified Auto-Regressive Modeling

Our unified autoregressive model is built upon a pre-trained LLM backbone, integrated with the discrete visual tokenizer described in Sec. 3.1. We treat visual understanding and generation tasks equally, utilizing teacher-forcing paradigm for both modalities. To handle autoregressive prediction over an expansive discrete vocabulary, we implement bitwise prediction [18], applying a Cross-Entropy loss directly to the quantized bit vectors. Furthermore, to maximize the efficiency of autoregressive modeling while minimizing the visual context length, we introduce **parallel bitwise prediction**. This strategy enables the model to concurrently predict a group of BSQ indices, leveraging the compressed representations provided by the spatial merger and multi-level DeepStack. The visual prediction head comprises an RMSNorm layer followed by a linear projection. Given the LLM hidden states $\mathbf{h} \in \mathbb{R}^{d^{\text{LLM}}}$, the visual prediction logits \mathbf{y} are computed as:

$$\text{logits}^{\text{vis}} = W^{\text{vis}} \text{RMSNorm} \mathbf{h}, \quad (3)$$

where $W^{\text{vis}} \in \mathbb{R}^{d^{\text{LLM}} \times d^{\text{vis}}}$ is the linear layer mapping the hidden dimension to the visual prediction space. For each input visual token, UNIAAR predicts the next group BSQ indices. Therefore, the output dimension is defined as $d^{\text{vis}} = 2 \times d^{\text{BSQ}} \times g$, where the group size $g = n^{\text{level}} \times n^{\text{spatial}}$, representing the number of hierarchical DeepStack levels and spatial units in the spatial merger, respectively. For text generation, we retain the original output layer of the LLM.

Finally, the total training objective, \mathcal{L}^{AR} is as follows:

$$\mathcal{L}^{\text{AR}} = \mathcal{L}^{\text{text}} \lambda^{\text{vis}} \cdot \mathcal{L}^{\text{vis}}, \quad (4)$$

where $\mathcal{L}^{\text{text}}$ and \mathcal{L}^{vis} denote the auto-regressive loss of text tokens and visual tokens, respectively; λ^{vis} is the corresponding factor of visual part. This unified scheme enables multimodal large-scale pretraining while reconciling the discrepancies between visual understanding and generation.

Furthermore, to bridge the gap between teacher-forcing training and autoregressive inference, we implement **random visual index flipping** for generation tasks, following [18]. Specifically, given visual BSQ indices $\mathbf{u} \in \{0, 1\}^{\text{seq} \times d^{\text{BSQ}}}$, we randomly invert a subset of the bits. This perturbation simulates the cumulative errors typically encountered during autoregressive sampling, while the original, uncorrupted indices are retained as ground-truth labels. This technique significantly improves the stability of the visual generation. Consequently, the model produces high-quality images even under high-temperature sampling. This flipping strategy is critical for the subsequent reinforcement learning phase, as it facilitates stable exploration in high-temperature regimes.

3.3 Visual Decoding

We build our visual decoder upon a pretrained single-stream Diffusion Transformer (DiT), to reconstruct pixels from the visual tokens. To integrate visual guidance, we fuse the conditioning visual signal f_v with the

noisy hidden states z via element-wise addition [14, 69]. We optimize the decoder using Conditional Flow Matching (CFM) [28], with training objective formulated as:

$$\mathcal{L}^{\text{CFM}} = \mathbb{E}_{t, p_t, z | \epsilon, p_\epsilon} \|\mathcal{D}_\Theta z \oplus f_v, t - u_t z | \epsilon\|_2^2, \quad (5)$$

where \mathcal{D}_Θ denotes the visual decoder and $z \in \mathbb{R}^{h \times w \times d^{\text{dit}}}$ represents the DiT internal hidden state. To obtain the visual conditioning signal f_v , we process the predicted visual BSQ indices \mathbf{u} as follows. At each autoregressive step, the model predicts a 2×2 spatial region, where each grid cell contains n^{level} BSQ vectors from different encoder layers. For the spatial dimension, we flatten the 2×2 grid into the sequence dimension while preserving the original spatial order. For the multi-level features at each spatial position, we concatenate them along the feature dimension and project the result to d^{dit} , yielding $f_v \in \mathbb{R}^{h \times w \times d^{\text{dit}}}$. By fusing f_v with z via element-wise addition, the decoder reconstructs images with visual guidance.

To alleviate the computational burden of the autoregressive model, we introduce a resolution upsampling strategy. This enables UNiAR to synthesize high-resolution images while maintaining a low autoregressive budget. Specifically, we apply 2D bicubic interpolation to f_v to reach the target resolution. Notably, we place all semantic and layout generation within the AR model, while the visual decoder functions only as a token-to-image translator. Therefore, unlike X-Omni [14] and NextFlow [68], our decoder does **not** take any textual prompts as input, and is conditioned solely on f_v .

4 Experiments

4.1 Training Recipe

We employ Qwen3-8B [65] as the LLM backbone, a BSQ-quantized SigLIP2-So400M ViT [55, 70] with DeepStack [1, 41] connections as the visual tokenizer, and a visual conditioned Stable Diffusion 3.5 Medium DiT [11] as the visual decoder. The training of UNiAR proceeds in a sequential, modular fashion. We first adapt the visual encoder into a discrete tokenizer via bitwise quantization (Sec. 3.1), after which it is frozen to ensure the visual codebook unchanged. We then train the DiT-based visual decoder (Sec. 3.3) to reconstruct images from the frozen encoder’s discrete indices, after which it is also frozen. With both components fixed, we proceed to train the autoregressive model.

We optimize the autoregressive model using a three-stage training procedure, during which the visual tokenizer and decoder are kept frozen. Since the next-token prediction involves only discrete visual tokens, the visual decoder is omitted during the initial two stages and only introduced during reinforcement learning to decode images for reward signals acquisition. We incorporate four task-specific Transformer layers following the LLM backbone dedicated to visual generation, which mitigates task competition between generation and understanding.

Pre-Training. The pretraining corpus comprises approximately 1T tokens, partitioned into 800B tokens for 8K context length pretraining and 200B tokens for 32K context length pretraining. We maintain a 1:1 ratio between visual understanding and generation data. For visual generation, the maximum resolution is constrained to 512×512 pixels during the 8K stage, and subsequently increased to 960×960 pixels for the 32K stage. During the pretraining, the tokens of visual generation are formulated as:

text prompt <image_gen> H W <vision_start> visual tokens <vision_end> ,

where the special token <image_gen> serves as a task-specific indicator for visual generation mode. To provide a spatial prior, we define H and W as the vertical and horizontal grid dimensions, respectively, which specify the token-count resolution for the generated output. Notably, while $\mathcal{L}^{\text{text}}$ is computed exclusively on the visual understanding datasets, \mathcal{L}^{vis} is optimized across both understanding and generation tasks, enforcing a unified representation space.

Supervised-Fine-Tuning. For SFT, we utilize a combination of public synthetic data and re-synthesized data, incorporating prompts from [3, 4] alongside prompts sampled from our pre-training corpus. We adopt the

Table 1 Performance comparison on GenEval. † denotes results with prompt rewriting, and ‡ denotes results from [64].

Method	Single	Two	Counting	Colors	Position	Color Attr.	Overall
<i>Gen. Only Models</i>							
SDXL	0.98	0.74	0.39	0.85	0.15	0.23	0.55
DALLE-3	0.96	0.87	0.47	0.83	0.43	0.45	0.67
SD3-medium	0.99	0.94	0.72	0.89	0.33	0.60	0.74
FLUX.1-dev†	0.98	0.93	0.75	0.93	0.68	0.65	0.82
<i>Unified Models</i>							
Emu3-Gen†	0.99	0.81	0.42	0.80	0.49	0.45	0.66
Show-o2	1.00	0.87	0.58	0.92	0.52	0.62	0.76
Janus-Pro	0.99	0.89	0.59	0.90	0.79	0.66	0.80
UniWorld-V1†	0.98	0.93	0.81	0.89	0.74	0.71	0.84
BAGEL	0.99	0.94	0.81	0.88	0.64	0.63	0.82
BAGEL†	0.98	0.95	0.84	0.95	0.78	0.77	0.88
OmniGen2†	0.99	0.96	0.74	0.98	0.72	0.75	0.86
X-Omni†	0.98	0.95	0.75	0.91	0.71	0.68	0.83
Emu3.5	-	-	-	-	-	-	0.86
NextFlow†	1.00	0.92	0.75	0.90	0.76	0.70	0.84
GPT-4o‡	0.99	0.92	0.85	0.92	0.75	0.61	0.84
UNIAR (ours)	1.00	0.95	0.75	0.94	0.77	0.67	0.85
UNIAR† (ours)	0.99	0.96	0.70	0.93	0.77	0.83	0.86

ChatML format for visual generation. In total, the SFT stage consists of approximately 50B tokens.

Reinforcement-Fine-Tuning. The next-token prediction scheme in UNIAR’s visual generation paradigm enables the seamless integration of reinforcement learning over discrete visual tokens. We introduce a reinforcement fine-tuning stage to further refine the model’s generation performance in image quality, text rendering and instruction following. The RL stage is applied exclusively to image generation task, while image editing and multimodal understanding tasks are not involved in this stage. More details about this stage can be found in [Sec. A.2](#).

4.2 Main Results

To comprehensively evaluate the performance of our unified model, we conduct a systematic assessment across a suite of benchmarks spanning visual content generation, image editing, and multimodal understanding.

Instruction Following. GenEval [15] is a widely-used benchmark for evaluating instruction-following in text-to-image synthesis. As shown in [Tab. 1](#), UNIAR achieves an overall score of 0.86, surpassing both the proprietary model GPT-4o [20] and the generation-only model Flux.1-dev. This result demonstrates UNIAR’s superior capability in precise instruction following during image generation, particularly in terms of accurate object counting, color-attribute binding, and spatial relationships.

Text Rendering. Text rendering is a pivotal capability for generative models, directly impacting their practical utility and user experience. However, it remains a significant challenge for existing text-to-image models. We evaluate UNIAR on the English subsets of OneIG-Bench [2] and LongText-Bench [14], with quantitative results summarized in [Tab. 2](#). UNIAR demonstrates state-of-the-art performance in this domain. Specifically, it achieves a score of 0.873 on OneIG-EN, surpassing GPT-4o [20]. Furthermore, on LongText-EN, which is a benchmark specifically tailored for long text rendering, UNIAR reaches 0.917, outperforming Gemini 2.5 Flash Image [8]. These results underscore that UNIAR is effective in text rendering.

Image Editing. We evaluate the image editing capabilities on ImgEdit Bench [66], with results shown in [Tab. 3](#). Our model achieves an overall score of 3.73, surpassing the Flux.1 Kontext [21] specifically designed for

Table 2 Comparison of text rendering performance. We conduct evaluation on the English splits of OneIG-Bench and LongText-Bench. UNiAR achieves strong text rendering performance, outperforming GPT-4o on OneIG-Bench and surpassing Gemini 2.5 Flash Image on LongText-Bench.

Method	Model Type	OneIG-EN	LongText-EN
FLUX.1-dev		0.523	0.607
HiDream-I1-Full		0.707	0.543
Kolors 2.0	Generation Only	0.427	0.258
Seedream 3.0		0.865	0.896
Qwen-Image		0.891	0.943
Janus-Pro		0.001	0.019
BAGEL		0.244	0.373
OmniGen2		0.680	0.561
X-Omni	Unified Model	0.901	0.900
GPT-4o		0.857	0.956
Gemini 2.5 Flash Image		0.894	0.869
UNiAR (ours)		0.873	0.917

Table 3 Performance comparison on ImgEdit-Bench.

Method	Add	Adjust	Extract	Replace	Remove	Background	Style	Hybrid	Action	Overall
<i>Gen. Only Models</i>										
MagicBrush	2.84	1.58	1.51	1.97	1.58	1.75	2.38	1.62	1.22	1.90
Instruct-Pix2Pix	2.45	1.83	1.44	2.01	1.50	1.44	3.55	1.20	1.46	1.88
AnyEdit	3.18	2.95	1.88	2.47	2.23	2.24	2.85	1.56	2.65	2.45
UltraEdit	3.44	2.81	2.13	2.96	1.45	2.83	3.76	1.91	2.98	2.70
ICEdit	3.58	3.39	1.73	3.15	2.93	3.08	3.84	2.04	3.68	3.05
Step1X-Edit	3.88	3.14	1.76	3.40	2.41	3.16	4.63	2.64	2.52	3.06
FLUX.1 Kontext [Dev]	4.12	3.80	2.04	4.22	3.09	3.97	4.51	3.35	4.25	3.71
Qwen-Image-Edit	4.38	4.16	3.43	4.66	4.14	4.38	4.81	3.82	4.69	4.27
<i>Unified Models</i>										
OmniGen	3.47	3.04	1.71	2.94	2.43	3.21	4.19	2.24	3.38	2.96
BAGEL	3.56	3.31	1.70	3.3	2.62	3.24	4.49	2.38	4.17	3.20
UniWorld-V1	3.82	3.64	2.27	3.47	3.24	2.99	4.21	2.96	2.74	3.26
OmniGen2	3.57	3.06	1.77	3.74	3.20	3.57	4.81	2.52	4.68	3.44
GPT-Image-1 [High]	4.61	4.33	2.90	4.35	3.66	4.57	4.93	3.96	4.89	4.20
UNiAR (ours)	3.91	3.48	2.75	3.94	3.64	3.74	4.27	3.06	4.70	3.73

editing tasks, as well as powerful unified models such as BAGEL [10] and OmniGen2 [61]. This demonstrates that our UNiAR possesses strong image editing capabilities.

Multimodal Understanding. We conduct a comprehensive comparison across several multimodal understanding benchmarks, as shown in Tab. 4. UNiAR significantly outperforms existing unified models on OCR-oriented tasks, such as OCRBench [32], DocVQA [39], and InfoVQA [40], as well as video understanding benchmark like MVBench [25]. Our model achieves results surpassing LLaVA-Onevision on the overall performance, and even obtains performance comparable to state-of-the-art open-source VLMs, *e.g.* Qwen3-VL [1]. However, we observe that performance on MMMU slightly lags behind. We attribute this gap to two primary factors: (1) Lack of pure-text data during pretraining, of which broad linguistic and factual knowledge is particularly beneficial; (2) We have not integrated reinforcement learning for understanding, which is proved to be effective for improving performance on reasoning-related tasks [24, 52].

Table 4 Comparison on multimodal understanding benchmarks.

Method	RLWDQA	MMMU	ChartQA	OCRBench	DocVQA	InfoVQA	MVBench
<i>Und. Only Models</i>							
LLaVA-1.6 [30]	-	35.1	54.8	-	74.4	37.1	-
LLaVA-OV [22]	66.3	48.8	80.0	-	87.5	68.8	56.7
Qwen3-VL [1]	71.5	69.6	89.6	896	96.1	83.1	68.7
<i>Unified Models</i>							
Emu3 [58]	-	31.6	-	687	76.3	43.8	-
Janus-Pro [5]	-	41.0	-	-	-	-	-
BLIP3-o [3]	-	50.6	-	-	-	-	-
Mogao [26]	-	44.2	-	-	-	-	-
BAGEL [10]	-	55.3	-	-	-	-	-
Show-o2 [63]	-	48.9	-	-	-	-	56.4
X-Omni [14]	-	-	-	704	88.6	-	-
UNIAR (Ours)	68.5	44.3	75.9	833	91.4	70.0	62.3

Table 5 Results on multimodal benchmarks. # Patches denotes the number of visual patches for the LLM.

Model	ViT	# Patches	OKVQA	TextVQA	DocVQA	InfoVQA	ChartQA	SEED	MME
DINOv2	L/14	576	54.1	13.4	7.3	21.3	10.8	57.0	1345
CLIP	L/14	576	60.0	47.5	25.6	21.8	19.2	70.1	1481
SigLIP	L/14	576	59.3	44.1	16.9	20.7	14.4	66.8	1416
SigLIP 2 (NaFlex)	So/16	576	60.6	59.9	28.9	25.0	18.4	73.1	1536
COMP-SigLIP	So/14	576	61.0	62.5	34.0	26.0	25.0	74.3	1543
AIMv2 (336px)	L/14	576	60.8	53.6	26.6	22.8	19.2	71.8	1472
UNIAR-SigLIP2	So/16	576	60.4	63.1	38.0	23.8	26.8	72.8	1537

4.3 Visual Tokenizer for Multimodal Understanding

Setup. We follow the setup of AIMv2 [12] for multimodal instruction tuning. Specifically, we use a two-layer MLP to connect the visual tokenizer and the LLM, Llama 3.0-8B [16]. The parameters of the visual tokenizer are kept frozen, while the MLP connector and the LLM are trainable. We adopt the same hyperparameter settings as in AIMv2 and train the model for one epoch on the LLaVA-SFT dataset [29].

Evaluation. We conduct evaluations on general knowledge tasks (OKVQA [46], SEED-Bench [23], MME [13]) and text-rich tasks (InfoVQA [40], TextVQA [49], DocVQA [39], and ChartQA [38]). As shown in Tab. 5, our visual tokenizer exhibits strong multimodal understanding capabilities, achieving the best performance on TextVQA, DocVQA and ChartQA and outperforming strong baselines such as AIMv2 [12], SigLIP2 [55] and CoMP-SigLIP [6].

4.4 Ablation Studies

Random visual index flip. To simulate the error accumulation faced by autoregressive generation, we apply random flipping to the bitwise visual index during pretraining. As shown in figure 4, we sample images under different temperatures and observe that the introduction of visual index flipping substantially stabilizes the generation process, enabling the model to produce coherent outputs even at higher temperatures. This improved robustness to accumulated errors is particularly important for reinforcement learning, where high-temperature is needed to encourage exploration.

Multi-level visual features. A key design of UNIAR lies in modeling the visual tokenizer with multi-level features, which aims to preserve both low-level details and high-level semantics. To validate this, we train the

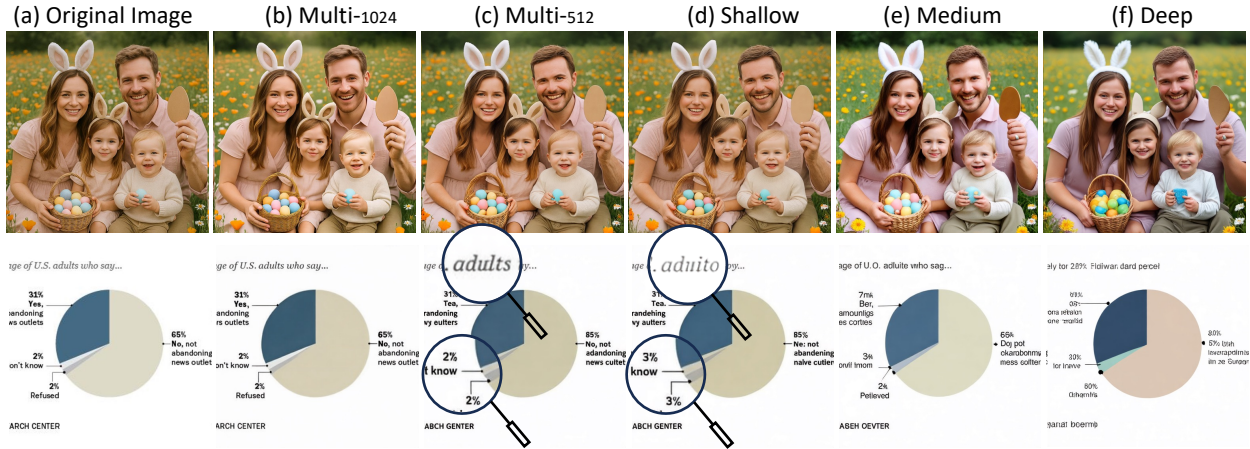


Figure 3 Image reconstruction conditioned on different levels of visual features. Shallow features preserve fine details while deep features capture semantics; multi-level conditioning yields the best results. Although our tokenizer is **not trained for reconstruction**, it still achieves strong reconstructions (b), highlighting the effectiveness of multi-level features.



Figure 4 Impact of random bitwise visual index flipping during pre-training. Top row: without flipping; Bottom row: with flipping.

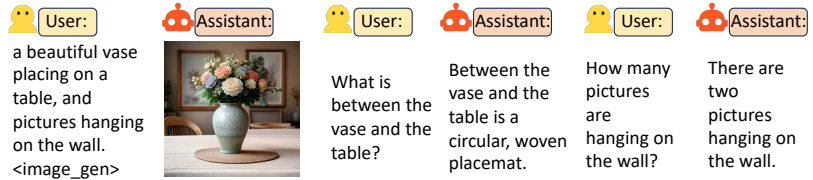


Figure 5 Emergent interleaved generation-understanding. Although not trained on interleaved multi-turn data, UNiAR can answer fine-grained questions about its own generated image within the same context, **without** an extra visual re-encoding step.

visual decoder conditioned on features from different ViT layers. As shown in [figure 3](#), deep-layer features mainly capture high-level semantics (*e.g.*, object categories) but lose high-frequency details, whereas shallow-layer features better support fine-grained reconstruction. Conditioning on multi-level features yields the best reconstruction quality. Interestingly, although our visual tokenizer is trained for multimodal understanding rather than image reconstruction, it can still recover most visual details under the 1024-resolution setting, including accurate rendering of textual content. This unexpectedly strong reconstruction performance further confirms the effectiveness of multi-level feature modeling in UNiAR.

4.5 Emerging Properties.

By using a single visual tokenizer for both understanding and generation, UNiAR unifies the representation space and enables an emergent capability for interleaved generation-understanding in a shared context. Notably, our training data does **not** contain multi-turn, interleaved generation-then-understanding dialogues. Nevertheless, as shown in [figure 5](#), when prompted with a vague instruction, UNiAR first generates an image and can then accurately answer follow-up questions about fine-grained details that were not specified in the original prompt. This suggests that UNiAR can directly interpret its own generated visual content within the same context, without an additional visual encoding pass. In contrast, approaches that rely on separate tokenizers (*e.g.* BAGEL and Janus-Pro) require re-encoding for understanding the generated image.

4.6 Effectiveness of Reinforcement Learning

UNIAR utilizes a discrete visual tokenizer and performs image prediction via auto-regression. This design is highly conducive to enhancing model capabilities through reinforcement learning. As illustrated in figure 6, the model’s text rendering performance improves significantly as the RL steps increase. When trained for 500 steps at a 512-pixel resolution, the score on OneIG-EN rose from 71.1 to 84.0. After an additional 100 steps of training at a higher resolution of 960, the metric further advanced to 87.3. This demonstrates that reinforcement learning can effectively improve the model performance.

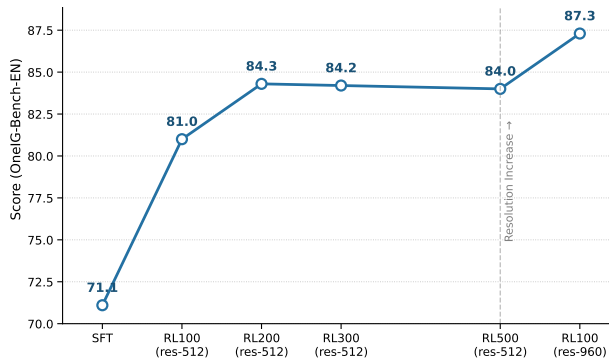


Figure 6 As the reinforcement learning steps increases, the text rendering capability improves significantly.

5 Conclusion

We proposed UNIAR, a unified autoregressive framework that unifies understanding, generation, and editing into a single model. UNIAR is built upon three key designs. First, it adopts a unified visual tokenizer that maps both generative and understanding tasks into the same semantic space, further strengthening model unification. Second, we successfully incorporate bitwise visual tokens for unified modeling, which substantially enlarges the representational capacity while incurring less computational overhead; meanwhile, we model multi-level visual features to improve expressiveness. Third, we introduce a parallel bitwise prediction mechanism together with a visual decoder with resolution upsampling, which significantly accelerates autoregressive prediction. Experiments shown that UNIAR achieves strong performance across generation, understanding, and editing.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No. 62472098) and the Science and Technology Commission of Shanghai Municipality (No. 25511106100).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Appendix

A Implementation Details

A.1 Pre-training and Supervised Fine-tuning

Our pre-training procedure is divided into two sequential stages with maximum sequence lengths of 8K and 32K tokens, respectively. We employ native resolution visual encoding during training, allowing for flexible data processing based on total token counts. For both understanding and generation tasks, we maintain a constant compression ratio of 32×32 . Consequently, a 512×512 image is represented by 256 tokens for autoregressive modeling. For the 8K stage, generation resolution is capped at a resolution of 512×512 pixels. While at the 32K stage, the generation resolution limit is increased to 960×960 pixels. Images exceeding these constraints are down-scaled while preserving their original aspect ratio. During PT, SFT and RL stages, the vision encoder remains frozen, while the Large Language Model and the spatial merger are fully trainable. Data is structured into three distinct formats to support unified modeling:

Table 6 Training hyperparameter configurations of UNiAR.

Hyperparameters	PT-8K	PT-32K	SFT
Learning rate (max:min)	5e-4:5e-5	5e-5:1e-5	1e-5:1e-6
LR scheduler	Cosine	Cosine	Cosine
Weight decay	0.05	0.05	0.05
Gradient norm clip	1.0	1.0	1.0
Optimizer	AdamW ($\beta_1 = 0.9, \beta_2 = 0.98$)		
Loss weight (Text: Vis)		1:10	
Warm-up steps	500	500	500
Global batch size	512	128	64
Training tokens	800B	200B	50B

- visual understanding: {prompt; **image_tokens**; **answers**} ;
- text-to-image generation: {prompt; **image_tokens**}
- image editing: {prompt; **reference_image_tokens**; **image_tokens**}.

The bold parts indicate the tokens that contribute to the autoregressive loss. For subsequent Supervised Fine-Tuning (SFT), we adapt these templates into the ChatML format to facilitate multi-turn dialogue and instruction following. We list the detailed training recipe in Tab. 6.

A.2 Reinforcement Learning

Reward system. We adopt multiple rewards to improve different aspects of the model. Specifically, for image quality, we use HPSv2 [62] and UnifiedReward [59] to enhance aesthetic quality and reduce artifacts. For text rendering, we apply PaddleOCR [9] to recognize texts in generated images and compute the reward based on the edit distance to the ground-truth text, encouraging accurate text rendering. For instruction following, we use the object-detector-based reward from FlowGRPO [31], which checks whether the generated image correctly reflects fine-grained concepts in the prompt, including object categories, counts, attributes, and relations. Finally, all rewards are normalized to the range 0, 1 and averaged as the final reward for each sample.

Data collection. To maximize coverage across diverse domains, we curate prompts from multiple sources, including (1) prompts sampled from our SFT data, (2) open-source datasets such as BLIP3o-60k [3] and ShareGPT-4o-Image [4], the GenEval training set and the OCR training set created by FlowGRPO [31], and (3) long-text prompts synthesized using a large language model.

Training details. We optimize the model with GRPO [47], using a constant learning rate of 5×10^{-6} . To prevent over-optimization, we add a KL-divergence term to the training loss with a coefficient of 0.01. Each batch contains 32 randomly sampled prompts, and we generate 16 images for each prompt. We train in two stages: (1) 500 steps at 512×512 resolution to quickly improve image quality and instruction following; (2) 100 additional steps at a higher 960×960 resolution to enhance long-text rendering.

A.3 Evaluation Configurations

Autoregressive sampling. For both text-to-image generation and image editing, we adopt classifier-free guidance (CFG) [19] during autoregressive sampling to improve instruction following. Specifically, we compute the final prediction by combining the conditional and unconditional predictions with a guidance scale $s = 2.5$ for GenEval and $s = 2.0$ for other benchmarks. During evaluation, we set the sampling temperature to 0.1.

Inference resolutions. On GenEval and ImgEdit, we autoregressively generate 512×512 visual tokens and apply $\times 2$ upsampling in the DiT decoder to obtain 1024×1024 output images. For text-rendering benchmarks

Table 7 Training efficiency comparison between continuous and discrete visual tokens.

	Continuous	Discrete
Time/Iter (s)	35.4	24.5

Table 8 GPU hours for each training stage. The total training cost is approximately 33k GPU hours.

	PT-8K	PT-32K	SFT	RL
GPU Hours	19k	10k	2k	1.9k

Table 9 Inference efficiency comparison.

	Downsample ratio	#Token	Time (s)
Janus-Pro (7B)	$\times 16$	4096	101.9
X-Omni (7B)	$\times 16$	4096	119.7
UniAR (8B) w/o decoder upsample	$\times 32$	1024	53.5
UniAR (8B) w/ decoder upsample	$\times 64$	256	13.0

(LongText Bench and OneIG-Bench), we autoregressively generate 1280×704 images and do not use decoder super-resolution.

A.4 Training and Inference Cost Analysis

Regarding training efficiency, discrete visual tokens improve pre-training throughput by approximately 30% over continuous tokens at a sequence length of 8K, as visual inputs can be pre-tokenized and stored as bit-packed representations offline. Detailed training elapsed times are provided in [Tab. 7](#). We also report the training time for each stage in [Tab. 8](#).

Regarding inference efficiency for image generation, we compare UniAR with Janus-Pro [5] and X-Omni [14]. We report the time required for the autoregressive generation stage, excluding the cost of the visual decoder. All models are evaluated on the same A100 GPU without CFG, and we report the time required to generate a 1024-resolution image. UniAR achieves significantly faster generation, mainly due to its higher $32\times$ downsampling ratio, which quadratically reduces the number of prediction steps. When the decoder uses upsampling, the number of predicted tokens is further reduced to 256, leading to an additional speedup.

B Limitations and Future Works

Due to resource constraints, we do not include pure-text data in joint pre-training, leaving considerable room for optimizing the data mixture for both understanding and generation. In addition, our RL training is currently limited to image generation. In the future, we aim to further explore large-scale pre-training by scaling both the dataset size and model parameters. Furthermore, we find that there is much room in our post-training, particularly through Reinforcement Learning. We will focus on developing reward models for specific domains, such as aesthetics, instruction-following, and text-rendering. Finally, we plan to extend the capabilities of each individual domain while exploring the synergy between them.

References

- [1] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibao Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. [arXiv preprint arXiv:2511.21631](#), 2025.
- [2] Jingjing Chang, Yixiao Fang, Peng Xing, Shuhan Wu, Wei Cheng, Rui Wang, Xianfang Zeng, Gang Yu, and Hai-Bao Chen. Oneig-bench: Omni-dimensional nuanced evaluation for image generation. In [NeurIPS](#), 2025.
- [3] Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. [arXiv preprint arXiv:2505.09568](#), 2025.
- [4] Junying Chen, Zhenyang Cai, Pengcheng Chen, Shunian Chen, Ke Ji, Xidong Wang, Yunjin Yang, and Benyou Wang. Sharegpt-4o-image: Aligning multimodal models with gpt-4o-level image generation. [arXiv preprint arXiv:2506.18095](#), 2025.
- [5] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. [arXiv preprint arXiv:2501.17811](#), 2025.
- [6] Yitong Chen, Lingchen Meng, Wujian Peng, Zuxuan Wu, and Yu-Gang Jiang. Comp: Continual multimodal pre-training for vision foundation models. [arXiv preprint arXiv:2503.18931](#), 2025.
- [7] Yitong Chen, Zuxuan Wu, Xipeng Qiu, and Yu-Gang Jiang. Catok: Taming mean flows for one-dimensional causal image tokenization. In [CVPR](#), 2026.
- [8] Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. [arXiv preprint arXiv:2507.06261](#), 2025.
- [9] Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiakuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, Yue Zhang, Wenyu Lv, Kui Huang, Yichao Zhang, Jing Zhang, Jun Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. Paddleocr 3.0 technical report, 2025. URL <https://arxiv.org/abs/2507.05595>.
- [10] 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.
- [11] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In [ICML](#), 2024.
- [12] Enrico Fini, Mustafa Shukor, Xiujun Li, Philipp Dufter, Michal Klein, David Haldimann, Sai Aitharaju, Victor G Turrisi da Costa, Louis Béthune, Zhe Gan, et al. Multimodal autoregressive pre-training of large vision encoders. In [CVPR](#), pages 9641–9654, 2025.
- [13] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. In [NeurIPS](#), 2025.
- [14] Zigang Geng, Yibing Wang, Yeyao Ma, Chen Li, Yongming Rao, Shuyang Gu, Zhao Zhong, Qinglin Lu, Han Hu, Xiaosong Zhang, et al. X-omni: Reinforcement learning makes discrete autoregressive image generative models great again. [arXiv preprint arXiv:2507.22058](#), 2025.
- [15] Dhruva Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In [NeurIPS](#), 2023.

- [16] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. [arXiv preprint arXiv:2407.21783](#), 2024.
- [17] Jiaming Han, Hao Chen, Yang Zhao, Hanyu Wang, Qi Zhao, Ziyang Yang, Hao He, Xiangyu Yue, and Lu Jiang. Vision as a dialect: Unifying visual understanding and generation via text-aligned representations. In *NeurIPS*, 2025.
- [18] Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *CVPR*, 2025.
- [19] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. [arXiv preprint arXiv:2207.12598](#), 2022.
- [20] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. [arXiv preprint arXiv:2410.21276](#), 2024.
- [21] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. [arXiv preprint arXiv:2506.15742](#), 2025.
- [22] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. [arXiv preprint arXiv:2408.03326](#), 2024.
- [23] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. In *CVPR*, 2024.
- [24] Jiase Li, Jingyang Chen, Yuxun Qu, Jianzhong Ju, Zhenbo Luo, Jian Luan, Shijie Xu, Zhenru Lin, Junyou Zhu, Boshen Xu, et al. Xiaomi mimo-vl-miloco technical report. [arXiv preprint arXiv:2512.17436](#), 2025.
- [25] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *CVPR*, pages 22195–22206, 2024.
- [26] Chao Liao, Liyang Liu, Xun Wang, Zhengxiong Luo, Xinyu Zhang, Wenliang Zhao, Jie Wu, Liang Li, Zhi Tian, and Weilin Huang. Mogao: An omni foundation model for interleaved multi-modal generation. [arXiv preprint arXiv:2505.05472](#), 2025.
- [27] 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.
- [28] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. [arXiv preprint arXiv:2210.02747](#), 2022.
- [29] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024.
- [30] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llanext: Improved reasoning, ocr, and world knowledge, 2024.
- [31] Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. In *NeurIPS*, volume 38, pages 40783–40818, 2025.
- [32] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024.
- [33] Zhiheng Liu, Weiming Ren, Haozhe Liu, Zijian Zhou, Shoufa Chen, Haonan Qiu, Xiaoke Huang, Zhaochong An, Fanny Yang, Aditya Patel, et al. Tuna: Taming unified visual representations for native unified multimodal models. [arXiv preprint arXiv:2512.02014](#), 2025.
- [34] Zhuohan Liu, Wujian Peng, Yitong Chen, and Zuxuan Wu. Compositional text-to-image generation via region-aware bimodal direct preference optimization. In *CVPR*, pages 36604–36614, 2026.
- [35] Jiasen Lu, Liangchen Song, Mingze Xu, Byeongjoo Ahn, Yanjun Wang, Chen Chen, Afshin Dehghan, and Yinfei Yang. Atoken: A unified tokenizer for vision. [arXiv preprint arXiv:2509.14476](#), 2025.

- [36] Chuofan Ma, Yi Jiang, Junfeng Wu, Jihan Yang, Xin Yu, Zehuan Yuan, Bingyue Peng, and Xiaojuan Qi. Unitok: A unified tokenizer for visual generation and understanding. In *NeurIPS*, volume 38, pages 129274–129297, 2025.
- [37] Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, et al. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In *CVPR*, 2025.
- [38] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pages 2263–2279, 2022.
- [39] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *WACV*, pages 2200–2209, 2021.
- [40] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *WACV*, pages 1697–1706, 2022.
- [41] Lingchen Meng, Jianwei Yang, Rui Tian, Xiyang Dai, Zuxuan Wu, Jianfeng Gao, and Yu-Gang Jiang. Deepstack: Deeply stacking visual tokens is surprisingly simple and effective for Imms. In *NeurIPS*, 2024.
- [42] 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.
- [43] Wujian Peng, Lingchen Meng, Yitong Chen, Yiweng Xie, Yang Liu, Tao Gui, Hang Xu, Xipeng Qiu, Zuxuan Wu, and Yu-Gang Jiang. Inst-it: Boosting instance understanding via explicit visual prompt instruction tuning. In *NeurIPS*, 2025.
- [44] 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. In *CVPR*, 2025.
- [45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*. PmLR, 2021.
- [46] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *ECCV*, pages 146–162. Springer, 2022.
- [47] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
- [48] Fengyuan Shi, Zhuoyan Luo, Yixiao Ge, Yujiu Yang, Ying Shan, and Limin Wang. Scalable image tokenization with index backpropagation quantization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16037–16046, 2025.
- [49] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, pages 8317–8326, 2019.
- [50] Hao Tang, Chenwei Xie, Xiaoyi Bao, Tingyu Weng, Pandeng Li, Yun Zheng, and Liwei Wang. Unilip: Adapting clip for unified multimodal understanding, generation and editing. *arXiv preprint arXiv:2507.23278*, 2025.
- [51] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.
- [52] Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.
- [53] Rui Tian, Mingfei Gao, Haiming Gang, Jiasen Lu, Zhe Gan, Yinfei Yang, Zuxuan Wu, and Afshin Dehghan. Unigen-1.5: Enhancing image generation and editing through reward unification in reinforcement learning. *arXiv preprint arXiv:2511.14760*, 2025.
- [54] Rui Tian, Mingfei Gao, Mingze Xu, Jiaming Hu, Jiasen Lu, Zuxuan Wu, Yinfei Yang, and Afshin Dehghan. Unigen: Enhanced training & test-time strategies for unified multimodal understanding and generation. In *NeurIPS*, 2025.

- [55] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. [arXiv preprint arXiv:2502.14786](#), 2025.
- [56] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *NeurIPS*, 2017.
- [57] Junke Wang, Yi Jiang, Zehuan Yuan, Bingyue Peng, Zuxuan Wu, and Yu-Gang Jiang. Omnitokenizer: A joint image-video tokenizer for visual generation. *NeurIPS*, 37:28281–28295, 2024.
- [58] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiyang Yu, et al. Emu3: Next-token prediction is all you need. [arXiv preprint arXiv:2409.18869](#), 2024.
- [59] Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. [arXiv preprint arXiv:2503.05236](#), 2025.
- [60] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. [arXiv preprint arXiv:2508.02324](#), 2025.
- [61] Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyang Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. [arXiv preprint arXiv:2506.18871](#), 2025.
- [62] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. [arXiv preprint arXiv:2306.09341](#), 2023.
- [63] Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. In *NeurIPS*, volume 38, pages 47490–47518, 2025.
- [64] Zhiyuan Yan, Junyan Ye, Weijia Li, Zilong Huang, Shenghai Yuan, Xiangyang He, Kaiqing Lin, Jun He, Conghui He, and Li Yuan. Gpt-imgeval: A comprehensive benchmark for diagnosing gpt4o in image generation. [arXiv preprint arXiv:2504.02782](#), 2025.
- [65] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. [arXiv preprint arXiv:2505.09388](#), 2025.
- [66] Yang Ye, Xianyi He, Zongjian Li, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, and Li Yuan. Imgedit: A unified image editing dataset and benchmark. In *NeurIPS*, 2025.
- [67] Lijun Yu, José Lezama, Nitesh Bharadwaj Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G Hauptmann, et al. Language model beats diffusion-tokenizer is key to visual generation. In *ICLR*, 2024.
- [68] Huichao Zhang, Liao Qu, Yiheng Liu, Hang Chen, Yangyang Song, Yongsheng Dong, Shikun Sun, Xian Li, Xu Wang, Yi Jiang, et al. Nextflow: Unified sequential modeling activates multimodal understanding and generation. [arXiv preprint arXiv:2601.02204](#), 2026.
- [69] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023.
- [70] Yue Zhao, Yuanjun Xiong, and Philipp Krähenbühl. Image and video tokenization with binary spherical quantization. In *ICLR*, 2025.
- [71] Yue Zhao, Fuzhao Xue, Scott Reed, Linxi Fan, Yuke Zhu, Jan Kautz, Zhiding Yu, Philipp Krähenbühl, and De-An Huang. Qlip: Text-aligned visual tokenization unifies auto-regressive multimodal understanding and generation. [arXiv preprint arXiv:2502.05178](#), 2025.
- [72] 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. In *ICLR*, 2025.