# EvoTok: A Unified Image Tokenizer via Residual Latent Evolution for Visual Understanding and Generation

Yan Li[1*], Ning Liao[1*†], Xiangyu Zhao[1], Shaofeng Zhang[2], Xiaoxing Wang[1], Yifan Yang[3], Junchi Yan[1], and Xue Yang[1✉]

[1]Shanghai Jiao Tong University
[2]University of Science and Technology of China
[3]Microsoft Corporation
[*]Equal Contributors, [†]Project Leader, [✉]Corresponding Author

https://github.com/VisionXLab/EvoTok

**Abstract.** The development of unified multimodal large language models (MLLMs) is fundamentally challenged by the granularity gap between visual understanding and generation: understanding requires high-level semantic abstractions, while image generation demands fine-grained pixel-level representations. Existing approaches usually enforce the two supervision on the same set of representation or decouple these two supervision on separate feature spaces, leading to interference and inconsistency, respectively. In this work, we propose **EvoTok**, a unified image tokenizer that reconciles these requirements through a residual evolution process within a shared latent space. Instead of maintaining separate token spaces for pixels and semantics, EvoTok encodes an image into a cascaded sequence of residual tokens via residual vector quantization. This residual sequence forms an evolution trajectory where earlier stages capture low-level details and deeper stages progressively transition toward high-level semantic representations. Despite being trained on a relatively modest dataset of 13M images, far smaller than the billion-scale datasets used by many previous unified tokenizers, EvoTok achieves a strong reconstruction quality of 0.43 / 0.25 rFID at 256×256 / 384×384 resolutions on ImageNet-1K, respectively. When integrated with a large language model, EvoTok shows promising performance across 8 out of 9 visual understanding benchmarks, and remarkable results on image generation benchmarks such as GenEval, GenAI-Bench, and DPG. These results demonstrate that modeling visual representations as an evolving trajectory provides an effective and principled solution for unifying visual understanding and generation.

**Keywords:** Unified Image Tokenizer · Visual Understanding and Generation · Residual Evolution

## 1 Introduction

Recent advances in Multimodal Large Language Models (MLLMs) [3,8,28,35,39] have revolutionized visual reasoning, while diffusion models [48, 51] and au-

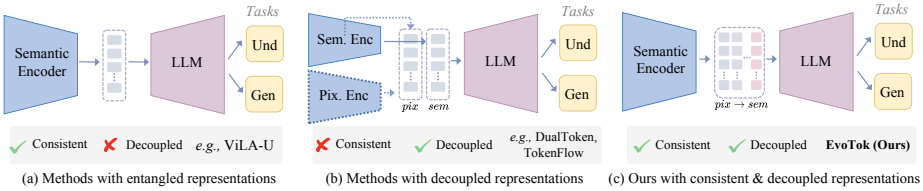



**Fig. 1: Different paradigms of current MLLMs and unified models.** (a) Entangled unified MLLMs share semantic and pixel features, enabling both understanding and generation but tightly coupling the two objectives. (b) Decoupled unified MLLMs separate semantic and pixel encoders or feature layers to disentangle tasks. (c) Our unified evolution MLLM evolves pixel features into semantic representations within a shared space while keeping decoupled latents, achieving aligned understanding and generation while preserving decoupled representations.

toregressive frameworks [54, 56] have enabled high-fidelity image synthesis. Although these paradigms have evolved separately, their complementary nature has sparked growing interest for unified visual understanding and generation, the core of which is designing a unified image tokenizer supporting both tasks.

Specifically, one line of work [45,63] unifies the text-image contrastive loss and VQ-based image reconstruction loss upon the same set of visual representations extracted by the designed tokenizers for consistency, as shown in Fig. 1 (a). This design forces the representation to serve as a bridge, capturing both the high-level semantics essential for understanding and the low-level pixel fidelity required for generation. However, it introduces an inherent optimization conflict: understanding tasks prioritize semantic alignment and discriminability, while generation tasks focus on fine-grained reconstruction and structural fidelity. These competing objectives interfere with each other within the shared feature space, undermining the representation's effectiveness for either task. Therefore, explicitly decoupling the features for these two tasks is essential.

To achieve feature decoupling, another line of approaches [12,17,34,53] learns semantic-wise features for understanding and pixel-wise features for generation independently, as shown in Fig. 1 (b). Nonetheless, this overly independent decoupling directly compromises the intrinsic consistency and correlation between the two types of features. The understanding and generation branches become isolated in the feature space, failing to share visual structures and semantic priors, which is detrimental to effective modeling within a unified architecture.

Building on these analysis, we advocate for a holistic design principle: *a unified tokenizer should not only decouple representations to mitigate task conflicts, but also preserve consistency between them to facilitate the sharing of visual appearance and semantic information.* In this work, we address both objectives by evolving pixel-level features into semantic-level features within a shared space.

We propose **EvoTok**, a unified image tokenizer that achieves task-decoupled and consistent visual representations through a *residual evolution process*, as shown in Fig. 1 (c). It leverages cascaded residual vector quantization to decom-

pose an image into a sequence of residual tokens within a *shared latent space*. Specifically, earlier residual stages capture spatial structures and perceptual details that are critical for image reconstruction, while deeper stages gradually steer the representation toward semantic features aligned with pretrained vision-language models. By organizing the visual representation along this residual trajectory, our tokenizer naturally reconciles the representational requirements of both tasks: pixel-level features are preserved for high-fidelity generation, while the accumulated representation encodes semantic abstractions suitable for visual understanding. Importantly, both forms of information naturally coexist within a single, unified latent space for task-consistency, rather than being maintained in separate ones.

Experimentally, despite being trained on a significantly smaller data scale with 13M images compared to baselines using billion-scale training data, the proposed EvoTok achieves strong rFID scores of 0.43 and 0.25 at 256×256 and 384×384 resolutions on ImageNet-1K [13], respectively. Moreover, EvoTok contributes to the best performance on 8 out of 9 visual understanding tasks (e.g., SEEDBench [32], MMMU [69], and MME [14]), and dominating complex image generation benchmarks, including GenEval [18], GenAI-Bench [26], and DPG [23].

Our contributions are summarized as follows:

(1) We propose EvoTok, a unified image tokenizer that represents images as a residual latent evolution trajectory within a shared latent space.

(2) We introduce a unified training objective that strikes a balance between task decoupling and cross-task consistency, learning visual representations that are simultaneously effective for semantic alignment and pixel reconstruction.

(3) Extensive experiments demonstrate the superior performance achieved by the proposed EvoTok on both visual understanding and generation benchmarks, and its effectiveness has been comprehensively validated.

## 2 Related Work

### 2.1 Image Tokenizer for Understanding

Image tokenization is a fundamental component of multimodal large language models (MLLMs) [3, 8, 35, 39], which require converting images into compact visual tokens that can be processed together with text tokens. A common design is to adopt pretrained vision encoders that produce semantically rich visual representations. For instance, CLIP [49] learns image representations aligned with text through large-scale contrastive pretraining, making it a widely adopted visual tokenizer for MLLMs [39]. Self-supervised visual encoders such as DINOv2 [47] have also been explored due to their strong performance on region-level and dense visual understanding tasks [44]. While these continuous tokenizers excel at capturing abstract concepts, their continuous latent spaces present significant hurdles for unified modeling alongside discrete text tokens in autoregressive frameworks. To bridge this gap, recent studies have attempted to discretize

continuous semantic features [16, 33] or directly employ discrete, reconstruction-based tokenizers [36] like VQVAE [59].

### 2.2 Image Tokenizer for Generation

In visual generation, image tokenizers aim to encode images into compact latent tokens that preserve detailed spatial information for high-quality reconstruction and generation. The seminal VQVAE [59] introduced discrete tokenization by mapping continuous latent features to a learnable codebook, enabling images to be represented as sequences of discrete tokens. Such vector-quantized tokenizers are widely adopted due to their discrete latent space and compatibility with autoregressive and masked generative models [6, 54, 56]. Subsequent improvements further enhanced reconstruction fidelity. VQGAN [66] introduced adversarial and perceptual losses to improve synthesis quality, while later works explored more expressive quantization strategies such as residual quantization [30]. Despite their strong reconstruction capability, these tokenizers are primarily optimized for pixel-level reconstruction and tend to focus on low-level visual details, often lacking strong semantic alignment with language representations. Consequently, models built upon such tokenizers often struggle with visual understanding tasks.

### 2.3 Unified Image Tokenizer

The pursuit of native multimodal systems has driven recent efforts to unify visual understanding and generation within a single architecture [42, 61–63]. A key challenge in this direction lies in designing an image tokenizer that can simultaneously support semantic reasoning and pixel-level generation. Existing unified tokenization strategies generally fall into two paradigms. One line of work adopts an ***entangled*** representation design, where both understanding and generation objectives are optimized on the same visual features. For example, models such as ViLA-U [63] and UniTok [45] integrate image reconstruction losses [66] with language alignment objectives [49] based on quantized features. Although such designs allow both tasks to share a unified representation space, they suffer from an inherent optimization conflict between semantic alignment and fine-grained reconstruction, ultimately limiting the upper-bound performance of both tasks.

To solve this, another line of work attempts to explicitly ***decouple*** semantic and pixel representations. Several strategies have been explored to achieve this goal. Some methods separate these two representations across different feature layers within the tokenizer, allowing early layers to capture pixel-level details while deeper layers focus on semantic abstractions [34, 53]. Other approaches introduce dedicated branches to independently model semantic and reconstruction features [12, 17]. Additionally, some works employ separate codebooks to store semantic and pixel tokens, explicitly partitioning the latent space for different objectives [12, 53]. While these designs alleviate optimization interference between tasks, the resulting representations often become overly independent, weakening the intrinsic consistency between visual structure and semantic information.

# 3 Method

In this section, we introduce EvoTok, a unified image tokenizer for visual understanding and generation that learns both semantic-level and pixel-level representations through residual evolution within a shared latent space. We first present the overall framework (Sec. 3.1), followed by its integration into visual understanding (Sec. 3.2) and generation (Sec. 3.3) tasks.

## 3.1 Unified Image Tokenizer

We propose **EvoTok**, which represents an image as a residual latent evolution trajectory within a shared feature space. This design simultaneously achieves task-decoupled features for understanding and generation, while maintaining consistency between pixel-level and semantic-level representations. Along with the evolution process, earlier residual stages capture structural and perceptual features for low-level reconstruction, while deeper stages progressively accumulate and refine prior features into high-level semantic tokens. By sharing this trajectory across both tasks, EvoTok achieves a truly unified representation where pixels and semantics co-evolve, naturally reconciling the demands of generation and understanding within a single latent space. The overview of the proposed **EvoTok** is presented in Fig. 2.

**Residual Evolution within Shared Latent Space.** We propose the residual evolution strategy to tokenize an image $I$ into task-decoupled and cross-task consistent representations, maintaining pixel-level and semantic-level features for visual generation and understanding. Given an initial feature $\mathbf{f}$ extracted by the shared encoder $\mathcal{E}$, a cascaded residual vector quantization composed of $L$ stages is performed following the RQ-VAE [30], which is formulated as:

$$(\mathbf{k}_1, \ldots, \mathbf{k}_L) = \mathcal{RQ}(\mathbf{f}; \{\mathcal{C}_i\}_{i=1}^{L}), \tag{1}$$

where $\mathcal{C}_i$ is the codebook for stage $i$. Setting the initial input $\mathbf{r}_0$ of the residual quantization process as the continuous feature $\mathbf{r}_0 = \mathbf{f}$, each stage recursively quantizes the residual as:

$$\mathbf{k}_i = \mathcal{Q}(\mathbf{r}_{i-1}; \mathcal{C}_i), \quad \mathbf{r}_i = \mathbf{r}_{i-1} - \mathbf{e}_i(\mathbf{k}_i), \quad i = 1, \ldots, L. \tag{2}$$

Here, $\mathbf{e}_i$ is the embedding table of $\mathcal{C}_i$, and $\mathcal{Q}$ is the standard vector quantization process formulated as:

$$\mathcal{Q}(\mathbf{z}; \mathcal{C}) = \arg\min_{k \in [K]} \|\mathbf{z} - \mathbf{e}(k)\|_2^2, \tag{3}$$

in which $K = |\mathcal{C}|$ denotes the codebook size.

Along the residual quantization process, we propose the *pixel-to-semantic* progression: early residuals capture perceptual structure for high-fidelity reconstruction, while the accumulated trajectory converges to high-level semantics.

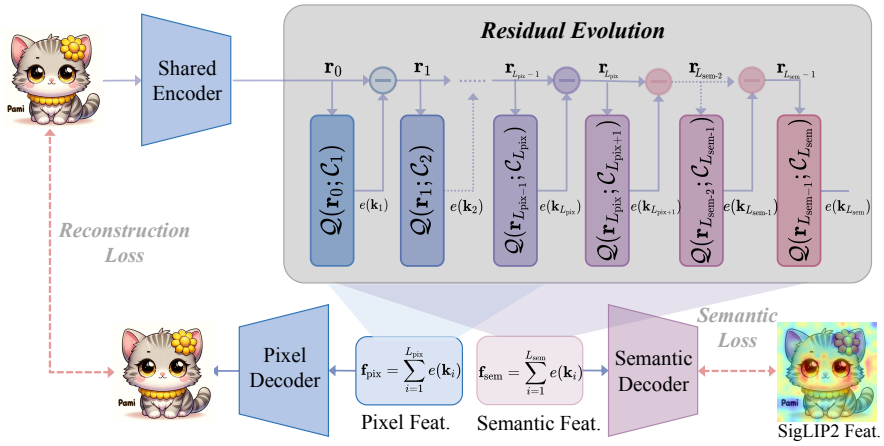



**Fig. 2: The overview of the proposed EvoTok.** An image is encoded into a sequence of residual tokens within a shared latent space, forming a residual evolution trajectory. Tokens from earlier stages preserve **pixel-level** details for reconstruction, while tokens from deeper stages, formed by progressively accumulating tokens from full stages, capture **semantic-level** features aligned with those extracted by the pre-trained SigLIP2 [58]. Within this residual evolution trajectory, low-level pixel features and high-level semantic representations ***co-evolve***, enabling a decoupled and consistent representations to support both visual understanding and generation.

We define two pivotal features along the residual trajectory. The pixel-level feature $\mathbf{f}_{\text{pix}}$ is computed as the partial sum of earlier stages, while the semantic feature $\mathbf{f}_{\text{sem}}$ is the cumulative sum over the full trajectory:

$$\mathbf{f}_{\text{pix}} = \sum_{i=1}^{L_{\text{pix}}} \mathbf{e}_i(\mathbf{k}_i), \quad \mathbf{f}_{\text{sem}} = \sum_{i=1}^{L_{\text{sem}}} \mathbf{e}_i(\mathbf{k}_i), \tag{4}$$

where $L_{\text{pix}}$ and $L_{\text{sem}}$ denote the number of earlier residual stages used for the pixel-level representation and the total number of residual stages in the trajectory, respectively. The pixel feature $\mathbf{f}_{\text{pix}}$ preserves dominant structures and dense perceptual details for generation, and the semantic feature $\mathbf{f}_{\text{sem}}$ represents refined semantics aligned with SigLIP2 [58].

**Unified Training Objective.** To enable the features extracted along the evolution trajectory to simultaneously support visual generation and understanding, we propose a unified training objective that enforces both pixel-level reconstruction and semantic alignment. In the *pixel-to-semantic* residual evolution process, the pixel-level feature $\mathbf{f}_{\text{pix}}$ is decoded by a pixel decoder for reconstruction, and the semantic feature $\mathbf{f}_{\text{sem}}$ is decoded by a semantic decoder to align with the feature extracted by SigLIP2 model. The total training objective is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pix}} + \mathcal{L}_{\text{sem}} + \mathcal{L}_{\text{VQ}}, \tag{5}$$

where the semantic loss $\mathcal{L}_{\text{sem}}$ is defined as the cosine similarity between the decoded semantic feature and target SigLIP2 embeddings. The pixel-level loss consists of reconstruction loss $\mathcal{L}_{\text{R}}$, perceptual loss $\mathcal{L}_{\text{P}}$, and adversarial loss $\mathcal{L}_{\text{G}}$:

$$\mathcal{L}_{\text{pix}} = \mathcal{L}_{\text{R}} + \lambda_{\text{P}}\mathcal{L}_{\text{P}} + \lambda_{\text{G}}\mathcal{L}_{\text{G}}. \tag{6}$$

In addition, the residual codebooks are trained with a standard VQ loss [66]:

$$\mathcal{L}_{\text{VQ}} = \sum_{d=1}^{L} \|\mathbf{f} - \text{sg}[\widehat{\mathbf{f}}_d]\|_2^2, \tag{7}$$

where $\widehat{\mathbf{f}}_d = \sum_{i=1}^{d} \mathbf{e}_i(\mathbf{k}_i)$ is the $d$-th layer feature from the residual trajectory, sg$[\cdot]$ denotes stop-gradient.

### 3.2 Visual Understanding with EvoTok

With the unified visual tokenizer aligned with language representations, we build a MLLM for visual understanding following the LLaVA [39] paradigm. Specifically, EvoTok encodes an image into $H \times W \times L_{\text{sem}}$ discrete codes, where $L_{\text{sem}}$ denotes the depth of the residual trajectory. For each spatial location, the semantic feature $\mathbf{f}_{\text{sem}}$ is obtained by aggregating residual embeddings along the trajectory as in Eq.(4). These semantic features are decoded by the semantic decoder and projected into the language embedding space through an MLP projector, which are then fed into the LLM together with text tokens. The model is trained with the standard autoregressive language modeling objective:

$$\mathcal{L}_{\text{CE}} = -\sum_{t} \log p(y_t \mid y_{<t}, \mathbf{v}), \tag{8}$$

where $\mathbf{v}$ denotes the visual tokens derived from EvoTok.

### 3.3 Visual Generation with EvoTok

For image generation, the LLM autoregressively predicts the pixel residual trajectory codes produced by the unified tokenizer. Following ViLA-U [63] and UniTok [45], we employ a depth RQ-Transformer [30] head to generate $L_{\text{pix}}$ codes for each spatial token. Specifically, given the hidden representation $\mathbf{h}_s$ from the LLM at spatial position $s$, the RQ head autoregressively predicts the residual codes $(\hat{\mathbf{k}}_{s,1}, \ldots, \hat{\mathbf{k}}_{s,L_{\text{pix}}})$ along the depth dimension. At each depth level $i \in \{1, \ldots, L_{\text{pix}}\}$, the code is sampled from a categorical distribution:

$$\hat{\mathbf{k}}_{s,i} \sim P(\mathbf{k}_{s,i} \mid \mathbf{h}_s, \hat{\mathbf{k}}_{s,<i}), \tag{9}$$

where a dedicated classifier is used for each codebook level $\mathcal{C}_i$.

To reconstruct the image, the predicted codes are mapped back to their feature embeddings $\{\mathbf{e}_i(\hat{\mathbf{k}}_{s,i})\}_{i=1}^{L_{\text{pix}}}$. These embeddings are accumulated along the residual trajectory to retrieve the pixel-level feature $\hat{\mathbf{f}}_{\text{pix},s}$ as defined in Eq. (4):

$$\hat{\mathbf{f}}_{\text{pix},s} = \sum_{i=1}^{L_{\text{pix}}} \mathbf{e}_i(\hat{\mathbf{k}}_{s,i}). \tag{10}$$

The resulting feature map $\hat{\mathbf{f}}_{\text{pix}}$ is then processed by the pixel decoder to generate the final image. During training, the generation model is optimized using a cross-entropy loss over the ground-truth residual codes:

$$\mathcal{L}_{\text{gen}} = -\sum_{s}\sum_{i=1}^{L_{\text{pix}}} \log P(\mathbf{k}_{s,i} \mid \mathbf{h}_s, \mathbf{k}_{s,<i}). \tag{11}$$

## 4 Experiments

In this section, we first introduce the experimental setup (Sec. 4.1), including the datasets, implementation details, and evaluation metrics. We then present quantitative and qualitative results on reconstruction (Sec. 4.2), understanding, and generation benchmarks (Sec. 4.3). Finally, we conduct ablation studies to analyze the effectiveness of the proposed pixel-to-semantic evolution (Sec. 4.4).

### 4.1 Experimental Setup

**Datasets.** For 256 resolution, we utilize ***fully open-sourced data without any re-caption or distillation*** across all of the training stages of our model. Specifically, EvoTok is trained on CC12M [7] and ImageNet-1K [13] trainset, and evaluated on ImageNet-1K valset. For multimodal understanding, LLaVA-Pretrain-558$k$ [39] is used for pretraining, and 13M data sampled from LLaVA-OneVision [31], LLaVA-NeXT-780$k$ [38], Cambrian-10M [57], and HoneyBee [4] are used for supervised finetuning (SFT). For image generation, we sample 22M data from ImageNet1K-QwenImage [19], BLIP3o-Pretrain-Short [8], LLaVA-OneVision-1.5 [1], ShareGPT-4o-Image [9] and OpenGPT-4o-Image [11]. For the 384 resolution variant, we train the model for multimodal understanding using an 8M internal dataset. While previous unified tokenizers [17, 35, 45] are often trained on billion-scale image–text pairs such as COYO-700M [5] or DataComp-1B [15], our model is trained on a significantly smaller scale and still achieves even stronger performance.

**Implementation Details.** We employ SigLip2-large-patch16-256 [58] as our vision encoder architecture and the teacher model. Residual quantization [30] is adopted to produce the codes, with each codebook containing 32,768 entries for each depth. All images are resized to 256×256 and encoded into 16×16×$L$ tokens, where the first $L_{\text{pix}} = 4$ depths of tokens preserve pixel-level features, while the full $L_{\text{sem}} = L = 16$ depths accumulated into semantic-level representations.


**Fig. 3: Images reconstruction at a resolution of 256× 256 on ImageNet-1K.**

The tokenizer is trained for $500k$ steps with a global batch size of 256, and the learning rate is set to $1e^{-4}$. For MLLM, we choose Qwen-2.5-7B-Instruct [55] as the language model and a depth transformer as in [63] as the generation head. All training experiments are performed on 16 GPUs. At inference, classifer-free guidance [22] is applied for image generation with a scale factor of 7.5.

**Evaluation Metrics.** We evaluate reconstruction quality using Fréchet Image Distance (FID) [21] on ImageNet-1K validation set [13]. For multimodal understanding, we conduct evaluations across a comprehensive suite of vision-language benchmarks, including SEEDBench [32], MM-Vet [68], GQA [24], AI2D [29], RealWorldQA [64], MMMU [69], MMBench [40], MME [14] and MME-P (MME-Perception). Besides, the image generation capability is evaluated on GenAI-Bench [26], GenEval [18], and DPG [23]. Following [17, 50, 61], we report our GenEval results using GPT-4o as a rewriter.

### 4.2 Unified Image Tokenizer

We benchmark EvoTok on the ImageNet-1K validation set to evaluate its reconstruction capability. Table 1 reports the commonly used metric rFID. Despite training on only 13M images, EvoTok achieves a strong rFID of **0.43** at 256 resolution. While AToken [43] and UniTok [45] report 0.38 and 0.41 respectively, a fully fair comparison is challenging due to significant differences in training scale and data (e.g., UniTok relies on the DataComp-1B [15] corpus with 1.28B image-text pairs). Furthermore, when further trained at 384 resolution, EvoTok achieves an even stronger rFID of **0.25**. Qualitative reconstruction results on ImageNet-1K are shown in Fig. 3.

### 4.3 Unified Understanding and Generation

**Understanding Performance.** Table 2 reports the performance achieved by the proposed EvoTok and other leading VLMs on diverse visual understanding benchmarks. EvoTok achieves leading performance among discrete unified methods. Specifically, our model secures the best results on **7** of 9 evaluated benchmarks, i.e., SEEDBench, GQA, AI2D, RealWorldQA, MMMU, MMBench, and

**Table 1: Comparison of the reconstruction FID on ImageNet-1K.** rFID is measured at 256×256 resolution. **Bold** and underline indicate the best and the second best performance, respectively. †: indicates model trained on ***billion-scale*** datasets such as COYO-700M [5] or DataComp-1B [15].

| Model | Resolution | Code Shape | Codebook Size | rFID↓ |
|---|---|---|---|---|
| ***Reconstruction Only*** | | | | |
| LLaMAGen [54] | 256 | 16×16 | 16,384 | 2.19 |
| RQVAE [30] | 256 | 16×16×4 | 16,384 | 3.20 |
| VQGAN-LC [71] | 256 | 16×16 | 16,384 | 2.62 |
| FlexTok [2] | 256 | 1-256 | 64,000 | 1.45 |
| IBQ [52] | 256 | 16×16 | 252,144 | 1.00 |
| ***Unified Tokenizer*** | | | | |
| TokenFlow-B† [17] | 224 | 680 | 32,768 | 1.37 |
| TokenFlow-L† [17] | 256 | 680 | 32,768 | 0.63 |
| VILA-U† [63] | 256 | 16×16×4 | 16,384 | 1.80 |
| SemHiTok [12] | 256 | 16×16 | 196,608 | 1.16 |
| UniTok† [45] | 256 | 16×16 | 65,536 | 0.41 |
| DualToken [53] | 256 | 16×16×4 | - | 0.52 |
| AToken-So/D [43] | 256 | 16×16 | 32,768 | <u>0.38</u> |
| Tar [20] | 384 | {729, 169, 81} | 65,536 | 2.60 |
| EvoTok (Ours) | 256 | 16×16×4 | 131,072 | 0.43 |
| EvoTok (Ours) | 384 | 27×27×4 | 131,072 | **0.25** |

MME, while ranking second-best on MM-Vet and MME-P. Besides, our model demonstrates superior reasoning and comprehensive perception capability. Notably, in reasoning-heavy benchmarks like AI2D and MMMU, EvoTok outperforms other discrete models by a large margin (i.e., 76.2 and 45.9, respectively). Furthermore, on the comprehensive MME test, our model (1895.1) significantly surpasses TokenFlow-B (1660.4). By scaling to 384 resolution, EvoTok further establishes new state-of-the-art results across most benchmarks, elevating AI2D to 81.4 and MM-Vet to 56.8. These results indicate that the pixel-to-semantic evolution ensures strong semantic representations, making EvoTok a competitive unified tokenizer for visual understanding.

**Generation Performance.** To evaluate the visual generation capability, we compare our model with other leading methods, including generation-only specialists and unified series, on three standard benchmarks: GenAI-Bench [26], GenEval [18], and DPG [23]. As reported in Table 3, EvoTok demonstrates strong performance, achieving the highest score of **86.36** on DPG, along with an overall score of 0.75 on GenEval and 0.87 on GenAI-Bench (Basic). Notably, our model significantly outperforms both unified baselines and specialized diffusion models in complex compositional tasks, particularly in Position (0.69) and Color Attribution (0.62). These results demonstrate that while low-level pixel features successfully evolve into high-level semantics for understanding, the residual trajectory simultaneously preserves granular generative priors, establishing EvoTok as a highly balanced and versatile unified tokenizer. Fig. 4

**Table 2: Quantitative results on visual understanding benchmarks.** We collect benchmarks including: SEEDBench [32], MM-Vet [68], GQA [24], AI2D [29], RealWorldQA [64], MMMU [69], MMBench [40], MME [14] and MME-P (MME-Perception). Three main types of methods are compared: continuous understanding methods (*top*), continuous unified methods (*middle*) and discrete unified methods (*bottom*). The best and the second best results among discrete unified methods are highlighted in **bold** and underline respectively.

| Model | LLM | Res. | SEEDB | MM-Vet | GQA | AI2D | RWQA | MMMU | MMB | MME | MME-P |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ***Understanding Only (Continuous)*** | | | | | | | | | | | |
| Qwen-VL-Chat [3] | Qwen-7B | 448 | 57.7 | - | 57.5 | - | - | - | - | 1848.3 | 1478.5 |
| VILA [35] | LLaMA-2-7B | 336 | 61.1 | 34.9 | 62.3 | - | - | - | 68.9 | 1533.0 | - |
| ShareGPT4V [10] | Vicuna-7B | 336 | 69.7 | 37.6 | 63.3 | 58.0 | 54.9 | 37.2 | 68.8 | 1943.8 | 1567.4 |
| LLaVA-v1.5 [37] | Vicuna-1.5-13B | 336 | 68.1 | 36.1 | 63.3 | 61.1 | 55.3 | 36.4 | 67.7 | 1826.7 | 1531.3 |
| ***Unified Understanding and Generation (Continuous)*** | | | | | | | | | | | |
| Janus [62] | DeepSeek-1.3B | 384 | - | 34.3 | 59.1 | - | - | - | - | 1338.0 | - |
| Unified-IO 2 [42] | 6.8B from scratch | 384 | 61.8 | - | - | - | - | - | 71.5 | - | - |
| LaVIT [27] | LLaMA-7B | 224 | - | - | 46.8 | - | - | - | - | - | - |
| ILLUME [60] | Vicuna-7B | - | - | 37.0 | - | 71.4 | - | 38.2 | 75.1 | - | 1445.3 |
| Show-o2 [65] | Qwen-2.5-7B-Inst. | 432 | 69.8 | - | 63.1 | 78.6 | - | 48.9 | 79.3 | - | 1620.5 |
| TUNA [41] | Qwen-2.5-7B-Inst. | 512 | 74.7 | - | 63.9 | 79.3 | 66.1 | 49.8 | - | - | 1641.5 |
| ***Unified Understanding and Generation (Discrete)*** | | | | | | | | | | | |
| QLIP [70] | Vicuna-1.5-7B | 392 | - | 33.3 | **61.8** | - | - | - | - | 1498.3 | - |
| VILA-U [63] | LLaMA-7B | 384 | 59.0 | 33.5 | 60.8 | - | - | - | - | 1401.8 | - |
| TokLIP [34] | Qwen2.5-7B-Inst. | 384 | 70.4 | 29.8 | 59.5 | - | - | 43.1 | 67.6 | - | 1488.4 |
| EMU3 [61] | 8B from scratch | - | 68.2 | 37.2 | 60.3 | 70.0 | 57.4 | 31.6 | 58.5 | - | - |
| Tar [20] | Qwen-2.5-7B-Inst. | 384 | 73.0 | - | 61.3 | - | - | 39.0 | 74.4 | 1926.0 | **1571.0** |
| AToken-So/D [43] | SlowFast-LLaVA-1.5-7B | 1536 | - | - | - | 81.2 | **68.8** | **48.7** | - | - | - |
| EvoTok (Ours) | Qwen-2.5-7B-Inst. | 384 | **75.6** | **56.8** | 50.2 | **81.4** | 65.8 | 47.6 | **77.3** | **1938.6** | 1504.7 |
| TokenFlow-B [17] | Vicuna-13B | 224 | 60.4 | 22.4 | 59.3 | 54.2 | 49.4 | 34.2 | 55.3 | 1660.4 | 1353.6 |
| TokenFlow-L [17] | Vicuna-13B | 256 | 62.6 | 27.7 | 60.3 | 56.6 | 49.2 | 34.4 | 60.3 | 1622.9 | 1365.4 |
| VILA-U [63] | LLaMA-7B | 256 | 56.3 | 27.7 | 58.3 | - | - | - | - | 1336.2 | - |
| SemHiTok [12] | Qwen2.5-7B | 256 | 62.9 | - | 60.3 | - | - | - | - | - | 1355.8 |
| UniTok [45] | LLaMA-2-7B | 256 | - | 33.9 | 61.1 | - | - | - | - | 1448.0 | - |
| DualToken [53] | LLaMA-2-7B | 256 | **71.8** | **40.5** | - | - | - | 45.8 | **74.9** | 1502.7 | - |
| EvoTok (Ours) | Qwen-2.5-7B-Inst. | 256 | **71.8** | 39.9 | **61.8** | **76.2** | **62.0** | **45.9** | **74.9** | **1895.1** | **1455.9** |

shows qualitative image generation results of EvoTok. The model demonstrates strong visual generation ability and produces high-quality images across diverse visual domains, including portraits, landscapes, objects, and artistic paintings.

### 4.4 Ablation Studies

**Ablation on the Depth of Pixel-to-Semantic Evolution.** We verify our core design by varying the depth allocation for pixel reconstruction $L_{\text{pix}}$ and semantic abstraction $L_{\text{sem}}$. Specifically, we compare three representative configurations: **1)** an *entangled feature space* where pixel and semantic using equal depth ($L_{\text{pix}} = L_{\text{sem}} = 4$), **2)** a *semantic-to-pixel evolution* where semantic abstraction is performed in earlier stages followed by pixel reconstruction ($L_{\text{pix}} = 16, L_{\text{sem}} = 4$), and **3)** the proposed *pixel-to-semantic evolution* where low-level pixel features are first modeled and progressively evolved into semantic representations ($L_{\text{pix}} = 4, L_{\text{sem}} = 16$).

As shown in Table 4, the results reveal two critical insights regarding the unified latent space: **First**, the ***entangled feature space*** (*Row 1*) leads to degraded performance on both reconstruction and semantic benchmarks, indicating that directly mixing pixel and semantic features within the same trajectory causes severe representation interference. **Second**, the ***semantic-to-pixel***

**Table 3: Quantitative results on visual generation benchmarks.** We compare two types of baselines: generation-only specialist models (*top*) and unified MLLMs (*bottom*). EvoTok achieves strong performance on both widely used benchmarks, GenAI-Bench [26] and GenEval [18].

| Model | Type | GenAI-Bench ↑ | | GenEval ↑ | | | | | | | DPG ↑ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Basic | Advanced | **Overall** | Single Obj. | Two Obj. | Counting | Colors | Position | Color Attri. | |
| ***Geneartion Only*** | | | | | | | | | | | |
| SD v2.1 [51] | Diffusion | 0.78 | 0.62 | 0.50 | 0.98 | 0.51 | 0.44 | 0.85 | 0.07 | 0.17 | - |
| SD v1.5 [51] | Diffusion | - | - | 0.43 | 0.97 | 0.38 | 0.35 | 0.76 | 0.04 | 0.06 | - |
| SDXL [48] | Diffusion | 0.83 | 0.63 | 0.55 | 0.98 | 0.74 | 0.39 | 0.85 | 0.15 | 0.23 | 74.65 |
| DALL-E 3 [50] | Diffusion | - | - | 0.67 | 0.96 | 0.87 | 0.47 | 0.83 | 0.43 | 0.45 | 83.50 |
| ***Unified Understanding and Generation*** | | | | | | | | | | | |
| Janus [62] | AR | - | - | 0.61 | 0.97 | 0.68 | 0.30 | 0.84 | 0.46 | 0.42 | - |
| ILLUME [60] | AR | 0.75 | 0.60 | 0.61 | **0.99** | 0.86 | **0.45** | 0.71 | 0.39 | 0.28 | - |
| QLIP-B/16 [70] | AR | - | - | 0.48 | 0.91 | 0.59 | 0.22 | 0.80 | 0.17 | 0.24 | 78.17 |
| TokenFlow-XL [17] | AR | - | - | 0.63 | 0.93 | 0.72 | **0.45** | 0.82 | 0.45 | 0.42 | 73.38 |
| EMU3 [61] | AR | - | - | 0.66 | **0.99** | 0.81 | 0.42 | 0.80 | 0.49 | 0.45 | 80.60 |
| VILA-U [63] | AR | 0.76 | 0.64 | - | - | - | - | - | - | - | - |
| SemHiTok [12] | AR | 0.83 | 0.65 | - | - | - | - | - | - | - | 83.59 |
| UniTok [45] | AR | 0.85 | **0.67** | - | - | - | - | - | - | - | 83.45 |
| DualToken-7B [53] | AR | - | - | **0.75** | - | - | - | - | - | - | - |
| EvoTok (Ours) | AR | **0.87** | 0.66 | **0.75** | **0.99** | **0.91** | 0.44 | **0.87** | **0.69** | **0.62** | **86.36** |

**Table 4: Ablation on unified feature trajectories.** Three different unified strategies are compared: 1) entangled pixel and semantic modeling, 2) semantic-to-pixel evolution, and 3) our proposed pixel-to-semantic evolution. Our method achieves the most balanced performance across all tasks.

| $L_{\text{pix}}$ | $L_{\text{sem}}$ | Reconstruction | | | Understanding | | | Generation | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | rFID ↓ | PSNR ↑ | SSIM ↑ | SEEDB ↑ | GQA ↑ | MME ↑ | GenAI (*Bas.*) ↑ | GenAI (*Adv.*) ↑ | GenEval ↑ |
| 4 | 4 | 0.66 | 22.96 | 0.76 | 62.7 | 58.0 | 1668.9 | 0.70 | 0.59 | 0.64 |
| 16 | 4 | **0.44** | **24.39** | **0.79** | 64.6 | 60.9 | 1731.6 | 0.71 | 0.58 | 0.60 |
| 4 | 16 | 0.55 | 23.60 | 0.77 | **67.1** | **61.3** | **1793.5** | **0.75** | **0.62** | **0.67** |

***evolution*** (*Row 2*) achieves the best reconstruction quality but yields noticeably weaker results on both understanding and generation tasks, indicating that high-level semantics cannot be effectively regressed from pixel-heavy tokens.

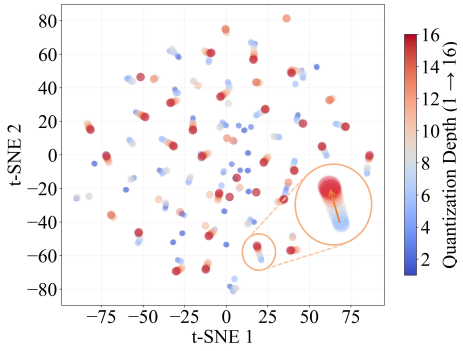
In contrast, the proposed ***pixel-to-semantic evolution*** (*Row 3*), where low-level pixel features progressively evolve into high-level semantic representations, produces the most balanced performance across reconstruction, understanding, and generation. These results demonstrate that the pixel-to-semantic evolution effectively reconciles the representational requirements of pixel-level generation and semantic-level understanding. By establishing this cascaded multi-stage progression, EvoTok mitigates the typical performance degradation seen in entangled architectures, achieving a unique synergy between generative quality and semantic understanding capability.

**Analysis of the Latent Evolution Trajectory.** Fig. 5a provides qualitative evidence of the shared latent space in EvoTok. The t-SNE visualization reveals a continuous progression along the residual stack, transitioning from pixel-level perception (blue) at shallower depths to high-level semantics (red) at deeper stages. The trajectory of each token (as indicated by the magnified inset)

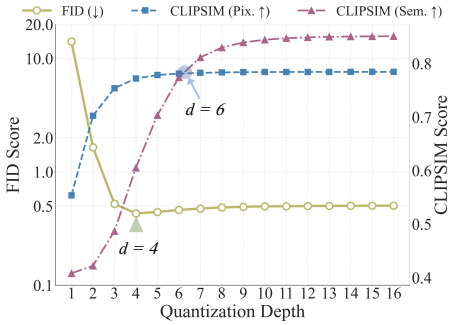

**Fig. 4: Images generated with our unified EvoTok at a resolution of 256× 256.** Our model generates high-quality images spanning diverse visual domains, including portraits, landscapes, objects, artistic paintings, etc.

demonstrates that instead of maintaining decoupled representations, EvoTok effectively unifies generative and discriminative features within a single, evolving latent path. This observation confirms that the residual quantization layers progressively refine low-level visual representations into high-level semantic concepts, providing a unified representation for both understanding and generation.

**Analysis of Feature Refinement Along the Evolution Trajectory.** We quantitatively analyze the representation evolution across the residual stack in Fig. 5b. Specifically, we monitor three key metrics along the trajectory: **1)** rFID of the reconstructed image; **2)** $\text{CLIPSIM}_{\text{pix}}$, calculated as the semantic similarity between the reconstructed image and the original input; and **3)** $\text{CLIPSIM}_{\text{sem}}$,




**(a) Latent Evolution Trajectory.** t-SNE visualization reveals a continuous transition from pixel-level perception (blue) to high-level semantics (red) within a single, unified latent space, bridging the gap between generation and understanding.




**(b) Feature Refinement Along the Evolution Trajectory.** At early depths (1–4), representations capture fine-grained pixel structures for reconstruction. As depth increases (8–16), the cumulative residuals progressively encode abstract concepts, achieving seamless alignment with high-level semantic features.

**Fig. 5: Analysis of the Unified Latent Space.**

which measures the semantic similarity between decoded semantic features and the GT semantic features. These metrics allow us to observe how visual information evolves between explicit pixel reconstruction and semantic latent encoding.

The empirical trends reveal a clear functional divergence. At early depths ($d \leq 4$), $\text{CLIPSIM}_{\text{pix}}$ and rFID improves rapidly, indicating that the initial residuals are highly efficient at capturing the primary semantic structures visible at the pixel level. However, beyond $d = 4$, $\text{CLIPSIM}_{\text{pix}}$ enters a plateau, suggesting a representational bottleneck where further pixel-level refinement no longer yields additional semantic gain. In contrast, $\text{CLIPSIM}_{\text{sem}}$ continues to ascend significantly. This decoupling of trends demonstrates that the later residual stages successfully transcend the limitations of raw pixel alignment, shifting their capacity toward encoding higher-order semantic concepts.

**Effectiveness of Feature Decoupling Strategy.** Table 5 compares the impact of four decoupling strategies: **1)** *entangled features* (Fig. 1(a), e.g., ViLA-U [63]) directly shares the same pixel and semantic features, implemented as the configuration in Table 4 with $L_{\text{pix}} = L_{\text{sem}} = 4$; **2)** *Layer-wise decoupled features from a single encoder* (Fig. 1(b), e.g., DualToken [53]) separates semantic and pixel representations by extracting them from different depths of a shared vision encoder, whereas our method derives both representations from the top-layer features without splitting intermediate layers; **3)** *Encoder-level decoupled features* (Fig. 1(b), e.g., TokenFlow [17]) decouples pixel and semantic representations using two independent visual encoders, whereas our method uses a shared encoder and evolves both features within a unified latent space. and **4)** our pixel-to-semantic evolution features (Fig. 1(c)).

As shown in Table 5, entangling pixel and semantic features (0.66) or using separate encoder branches (0.68) leads to inferior reconstruction quality, while decoupling features from different layers of a single encoder improves perfor-

**Table 5: Ablation on different feature decoupling method.**

| Decouple Method | rFID ↓ |
|---|---|
| 1) No decouple | 0.66 |
| 2) Different feature layer | 0.59 |
| 3) Different branch | 0.68 |
| 4) Ours | **0.55** |

**Table 6: Ablation on quantization methods.**

| Quantization | rFID ↓ |
|---|---|
| FSQ [46] | 3.17 |
| LFQ [67] | 2.13 |
| PQ [25] | 0.98 |
| RQ [30] | **0.55** |

mance (0.59), our pixel-to-semantic evolution achieves the best rFID of 0.55. This demonstrates that our approach preserves fine-grained visual details more effectively, compared to rigid decoupling or fully entangled representations.

**Ablation on Quantization Methods.** To evaluate the impact of different quantization strategies on the unified representation, we compare our adopted Residual Quantization (RQ) with other vector quantization variants (FSQ, LFQ, PQ). As shown in Table 6, RQ achieves the best reconstruction rFID of 0.55. This is attributed to its cascaded residual structure, which inherently enables the *stage-by-stage* pixel-to-semantic evolution. In contrast, other variants quantize dimensions or subspaces independently, failing to support the progressive feature evolution required for our unified modeling.

## 5 Conclusion

This paper presents EvoTok, a unified image tokenizer designed to support both visual understanding and generation. Existing approaches typically either entangle semantic and pixel representations in a shared feature space or overly decouple them into independent branches, which limits the effectiveness of unified multimodal modeling. To address this issue, EvoTok represents images as a residual evolution trajectory within a shared latent space, where pixel-level representations progressively evolve into higher-level semantic abstractions through cascaded residual quantization. This design enables task-decoupled representations for generation and understanding while preserving their intrinsic consistency. Extensive experiments demonstrate that EvoTok achieves strong performance on both visual generation and understanding benchmarks, despite being trained on significantly less data than prior unified models. We hope the proposed residual evolution paradigm can offer new insights into tokenizer design and facilitate the development of more capable unified multimodal systems.

## Acknowledgements

This work was supported by the National Natural Science Foundation of China (62506229), Natural Science Foundation of Shanghai under 25ZR1402268, and Shanghai QiYuan Innovation Foundation.

## References


1. An, X., Xie, Y., Yang, K., Zhang, W., Zhao, X., Cheng, Z., Wang, Y., Xu, S., Chen, C., Wu, C., Tan, H., Li, C., Yang, J., Yu, J., Wang, X., Qin, B., Wang, Y., Yan, Z., Feng, Z., Liu, Z., Li, B., Deng, J.: Llava-onevision-1.5: Fully open framework for democratized multimodal training. In: arXiv (2025)
2. Bachmann, R., Allardice, J., Mizrahi, D., Fini, E., Kar, O.F., Amirloo, E., El-Nouby, A., Zamir, A., Dehghan, A.: Flextok: Resampling images into 1d token sequences of flexible length (2025), https://arxiv.org/abs/2502.13967
3. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966 **1**(2), 3 (2023)
4. Bansal, H., Sachan, D.S., Chang, K.W., Grover, A., Ghosh, G., Yih, W.t., Pasunuru, R.: Honeybee: Data recipes for vision-language reasoners. arXiv preprint arXiv:2510.12225 (2025)
5. Byeon, M., Park, B., Kim, H., Lee, S., Baek, W., Kim, S.: Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset (2022)
6. Chang, H., Zhang, H., Jiang, L., Liu, C., Freeman, W.T.: Maskgit: Masked generative image transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11315–11325 (2022)
7. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3558–3568 (2021)
8. Chen, J., Xu, Z., Pan, X., Hu, Y., Qin, C., Goldstein, T., Huang, L., Zhou, T., Xie, S., Savarese, S., et al.: Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. arXiv preprint arXiv:2505.09568 (2025)
9. Chen, J., Cai, Z., Chen, P., Chen, S., Ji, K., Wang, X., Yang, Y., Wang, B.: Sharegpt-4o-image: Aligning multimodal models with gpt-4o-level image generation (2025), https://arxiv.org/abs/2506.18095
10. Chen, L., Li, J., Dong, X., Zhang, P., He, C., Wang, J., Zhao, F., Lin, D.: Sharegpt4v: Improving large multi-modal models with better captions. In: European Conference on Computer Vision. pp. 370–387. Springer (2024)
11. Chen, Z., Bai, X., Shi, Y., Fu, C., Zhang, H., Wang, H., Sun, X., Zhang, Z., Wang, L., Zhang, Y., Wan, P., Zhang, Y.F.: Opengpt-4o-image: A comprehensive dataset for advanced image generation and editing (2025), https://arxiv.org/abs/2509.24900
12. Chen, Z., Wang, C., Chen, X., Xu, H., Huang, R., Zhou, J., Han, J., Xu, H., Liang, X.: Semhitok: A unified image tokenizer via semantic-guided hierarchical codebook for multimodal understanding and generation. arXiv preprint arXiv:2503.06764 (2025)
13. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
14. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., et al.: Mme: A comprehensive evaluation benchmark for multimodal large language models. arXiv preprint arXiv:2306.13394 (2023)
15. Gadre, S.Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., Marten, R., Wortsman, M., Ghosh, D., Zhang, J., et al.: Datacomp: In search of the next generation of multimodal datasets. Advances in Neural Information Processing Systems **36**, 27092–27112 (2023)

16. Ge, Y., Ge, Y., Zeng, Z., Wang, X., Shan, Y.: Planting a seed of vision in large language model. arXiv preprint arXiv:2307.08041 (2023)
17. Geyer, M., Bar-Tal, O., Bagon, S., Dekel, T.: Tokenflow: Consistent diffusion features for consistent video editing. arXiv preprint arXiv:2307.10373 (2023)
18. Ghosh, D., Hajishirzi, H., Schmidt, L.: Geneval: An object-focused framework for evaluating text-to-image alignment. Advances in Neural Information Processing Systems **36** (2024)
19. Han, J.: Imagenet1k-t2i-qwenvl-qwenimage (2025), `https://huggingface.co/datasets/csuhan/ImageNet1K-T2I-QwenVL-QwenImage`
20. Han, J., Chen, H., Zhao, Y., Wang, H., Zhao, Q., Yang, Z., He, H., Yue, X., Jiang, L.: Vision as a dialect: Unifying visual understanding and generation via text-aligned representations. Advances in Neural Information Processing Systems **38**, 158430–158459 (2026)
21. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
22. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
23. Hu, X., Wang, R., Fang, Y., Fu, B., Cheng, P., Yu, G.: Ella: Equip diffusion models with llm for enhanced semantic alignment. arXiv preprint arXiv:2403.05135 (2024)
24. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6700–6709 (2019)
25. Jegou, H., Douze, M., Schmid, C.: Product quantization for nearest neighbor search. IEEE transactions on pattern analysis and machine intelligence **33**(1), 117–128 (2010)
26. Jiang, D., Ku, M., Li, T., Ni, Y., Sun, S., Fan, R., Chen, W.: Genai arena: An open evaluation platform for generative models. arXiv preprint arXiv:2406.04485 (2024)
27. Jin, Y., Xu, K., Chen, L., Liao, C., Tan, J., Huang, Q., Chen, B., Lei, C., Liu, A., Song, C., et al.: Unified language-vision pretraining in llm with dynamic discrete visual tokenization. arXiv preprint arXiv:2309.04669 (2023)
28. Jin, Y., Li, J., Gu, T., Liu, Y., Zhao, B., Lai, J., Gan, Z., Wang, Y., Wang, C., Tan, X., et al.: Efficient multimodal large language models: A survey. Visual Intelligence **3**(1), 27 (2025)
29. Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., Farhadi, A.: A diagram is worth a dozen images. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14. pp. 235–251. Springer (2016)
30. Lee, D., Kim, C., Kim, S., Cho, M., Han, W.S.: Autoregressive image generation using residual quantization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11523–11532 (2022)
31. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., Li, C.: Llava-onevision: Easy visual task transfer. Transactions on Machine Learning Research (2024)
32. Li, B., Wang, R., Wang, G., Ge, Y., Ge, Y., Shan, Y.: Seed-bench: Benchmarking multimodal llms with generative comprehension. arXiv preprint arXiv:2307.16125 (2023)
33. Liao, N., Shi, B., Zhang, X., Cao, M., Yan, J., Tian, Q.: Rethinking visual prompt learning as masked visual token modeling. Artificial Intelligence p. 104417 (2025)

34. Lin, H., Wang, T., Ge, Y., Ge, Y., Lu, Z., Wei, Y., Zhang, Q., Sun, Z., Shan, Y.: Toklip: Marry visual tokens to clip for multimodal comprehension and generation. arXiv preprint arXiv:2505.05422 (2025)
35. Lin, J., Yin, H., Ping, W., Molchanov, P., Shoeybi, M., Han, S.: Vila: On pretraining for visual language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26689–26699 (2024)
36. Liu, H., Yan, W., Zaharia, M., Abbeel, P.: World model on million-length video and language with blockwise ringattention. arXiv preprint arXiv:2402.08268 (2024)
37. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26296–26306 (2024)
38. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llava-next: Improved reasoning, ocr, and world knowledge (January 2024), https://llava-vl.github.io/blog/2024-01-30-llava-next/
39. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning (2023)
40. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? In: European Conference on Computer Vision. pp. 216–233. Springer (2025)
41. Liu, Z., Ren, W., Liu, H., Zhou, Z., Chen, S., Qiu, H., Huang, X., An, Z., Yang, F., Patel, A., et al.: Tuna: Taming unified visual representations for native unified multimodal models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15740–15751 (2026)
42. Lu, J., Clark, C., Lee, S., Zhang, Z., Khosla, S., Marten, R., Hoiem, D., Kembhavi, A.: Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26439–26455 (2024)
43. Lu, J., Song, L., Xu, M., Ahn, B., Wang, Y., Chen, C., Dehghan, A., Yang, Y.: Atoken: A unified tokenizer for vision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 28701–28711 (2026)
44. Ma, C., Jiang, Y., Wu, J., Yuan, Z., Qi, X.: Groma: Localized visual tokenization for grounding multimodal large language models. In: European Conference on Computer Vision. pp. 417–435. Springer (2024)
45. Ma, C., Jiang, Y., Wu, J., Yang, J., Yu, X., Yuan, Z., Peng, B., Qi, X.: Unitok: A unified tokenizer for visual generation and understanding. arXiv preprint arXiv:2502.20321 (2025)
46. Mentzer, F., Minnen, D., Agustsson, E., Tschannen, M.: Finite scalar quantization: Vq-vae made simple. In: International Conference on Learning Representations. vol. 2024, pp. 51772–51783 (2024)
47. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
48. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
49. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
50. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)

51. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: 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)
52. Shi, F., Luo, Z., Ge, Y., Yang, Y., Shan, Y., Wang, L.: Scalable image tokenization with index backpropagation quantization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16037–16046 (2025)
53. Song, W., Wang, Y., Song, Z., Li, Y., Sun, H., Chen, W., Zhou, Z., Xu, J., Wang, J., Dualtoken, K.Y.: Towards unifying visual understanding and generation with dual visual vocabularies. arXiv preprint arXiv:2503.14324 (2025)
54. Sun, P., Jiang, Y., Chen, S., Zhang, S., Peng, B., Luo, P., Yuan, Z.: Autoregressive model beats diffusion: Llama for scalable image generation. arXiv preprint arXiv:2406.06525 (2024)
55. Team, Q.: Qwen2.5: A party of foundation models (September 2024), https://qwenlm.github.io/blog/qwen2.5/
56. Tian, K., Jiang, Y., Yuan, Z., Peng, B., Wang, L.: Visual autoregressive modeling: Scalable image generation via next-scale prediction. Advances in neural information processing systems **37**, 84839–84865 (2024)
57. Tong, P., Brown, E., Wu, P., Woo, S., Iyer, A.J.V., Akula, S.C., Yang, S., Yang, J., Middepogu, M., Wang, Z., et al.: Cambrian-1: A fully open, vision-centric exploration of multimodal llms. Advances in Neural Information Processing Systems **37**, 87310–87356 (2024)
58. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., Hénaff, O., Harmsen, J., Steiner, A., Zhai, X.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features (2025), https://arxiv.org/abs/2502.14786
59. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017)
60. Wang, C., Lu, G., Yang, J., Huang, R., Han, J., Hou, L., Zhang, W., Xu, H.: Illume: Illuminating your llms to see, draw, and self-enhance. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21612–21622 (2025)
61. Wang, X., Zhang, X., Luo, Z., Sun, Q., Cui, Y., Wang, J., Zhang, F., Wang, Y., Li, Z., Yu, Q., et al.: Emu3: Next-token prediction is all you need. arXiv preprint arXiv:2409.18869 (2024)
62. Wu, C., Chen, X., Wu, Z., Ma, Y., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C., et al.: Janus: Decoupling visual encoding for unified multimodal understanding and generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12966–12977 (2025)
63. Wu, Y., Zhang, Z., Chen, J., Tang, H., Li, D., Fang, Y., Zhu, L., Xie, E., Yin, H., Yi, L., et al.: Vila-u: a unified foundation model integrating visual understanding and generation. arXiv preprint arXiv:2409.04429 (2024)
64. XAI: Realworldqa (2024)
65. Xie, J., Yang, Z., Shou, M.Z.: Show-o2: Improved native unified multimodal models. Advances in Neural Information Processing Systems **38**, 47490–47518 (2026)
66. Yu, J., Li, X., Koh, J.Y., Zhang, H., Pang, R., Qin, J., Ku, A., Xu, Y., Baldridge, J., Wu, Y.: Vector-quantized image modeling with improved vqgan. arXiv preprint arXiv:2110.04627 (2021)
67. Yu, L., Lezama, J., Gundavarapu, N.B., Versari, L., Sohn, K., Minnen, D., Cheng, Y., Gupta, A., Gu, X., Hauptmann, A.G., et al.: Language model beats diffusion-

tokenizer is key to visual generation. In: International Conference on Learning Representations. vol. 2024, pp. 765–783 (2024)
68. Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., Wang, L.: Mm-vet: Evaluating large multimodal models for integrated capabilities. arXiv preprint arXiv:2308.02490 (2023)
69. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9556–9567 (2024)
70. Zhao, Y., Xue, F., Reed, S., Fan, L., Zhu, Y., Kautz, J., Yu, Z., Krähenbühl, P., Huang, D.A.: Qlip: Text-aligned visual tokenization unifies auto-regressive multimodal understanding and generation. arXiv preprint arXiv:2502.05178 (2025)
71. Zhu, L., Wei, F., Lu, Y., Chen, D.: Scaling the codebook size of vq-gan to 100,000 with a utilization rate of 99%. Advances in Neural Information Processing Systems **37**, 12612–12635 (2024)