
TokLIP: Marry Visual Tokens to CLIP for Multimodal Comprehension and Generation

Haokun Lin^{*1,2,4}, Teng Wang^{*1}, Yixiao Ge^{†1}, Yuying Ge¹, Zhichao Lu², Ying Wei³, Qingfu Zhang², Zhenan Sun⁴, Ying Shan¹

¹ ARC Lab, Tencent PCG ² City University of Hong Kong
³ Zhejiang University ⁴ NLPR & MAIS, Institute of Automation, CAS

^{*}Equal Contribution [†]Corresponding Author

Abstract

Pioneering token-based works such as Chameleon and Emu3 have established a foundation for multimodal unification but face challenges of high training computational overhead and limited comprehension performance due to a lack of high-level semantics. In this paper, we introduce TokLIP, a visual tokenizer that enhances comprehension by semanticizing vector-quantized (VQ) tokens and incorporating CLIP-level semantics while enabling end-to-end multimodal autoregressive training with standard VQ tokens. TokLIP integrates a low-level discrete VQ tokenizer with a ViT-based token encoder to capture high-level continuous semantics. Unlike previous approaches (e.g., VILA-U) that discretize high-level features, TokLIP disentangles training objectives for comprehension and generation, allowing the direct application of advanced VQ tokenizers without the need for tailored quantization operations. Our empirical results demonstrate that TokLIP achieves exceptional data efficiency, empowering visual tokens with high-level semantic understanding while enhancing low-level generative capacity, making it well-suited for autoregressive Transformers in both comprehension and generation tasks. The code and models are available at <https://github.com/TencentARC/TokLIP>.

1 Introduction

A unified autoregressive model with end-to-end next multimodal token prediction is considered the way toward the ChatGPT moment for the multimodal world. Pioneering works, such as Chameleon [50] and Emu3 [58], train a single Transformer with vector-quantized (VQ) visual tokens and word tokens on multimodal sequences. Though scalable, the models suffer from expensive computational overhead for training to converge and weak comprehension performance.

VILA-U *et al.* [61, 42, 71] highlight that the issue lies in the fact that VQ tokens lack high-level semantics, which CLIP [43] possesses and has proven effective for LLaVA [34, 35]-like frameworks designed solely for comprehension purposes. They attempt to solve the problem via discretizing high-level semantical features, that is, training a continuous-to-discrete visual tokenizer with both reconstruction and text alignment objectives, illustrated in Figure 1 (a). However, the conflict between two distinct objectives and the information loss of visual semantics caused by quantization operation pose new challenges for properly unifying multimodal comprehension and generation.

The evidence suggests that a) the objectives of multimodal comprehension and generation should be disentangled due to varying semantic demands, and b) improved quantization techniques yield superior results. To this end, we argue that the proper approach to meet such principles is to *semanticize VQ tokens* rather than discretize CLIP features. In this paper, we introduce **TokLIP**, a discrete-to-continuous visual tokenizer as shown in Figure 1 (b) that enables end-to-end autoregressive

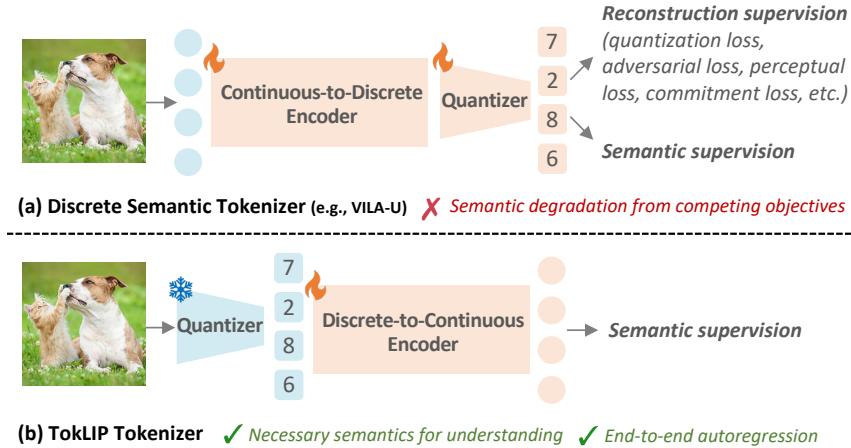


Figure 1: **Visual tokenizers for unified multimodal comprehension and generation.** (a) Previous discrete semantic tokenizers quantize high-level features under entangled reconstruction and semantic supervision. (b) In contrast, our TokLIP disentangles the dual objectives by semanticizing low-level VQ tokens.

training of a single Transformer with sequences of multimodal discrete tokens, while simultaneously incorporating CLIP-level semantics to enhance multimodal comprehension.

Specifically, the TokLIP tokenizer consists of an off-the-shelf low-level VQ tokenizer, such as VQGAN [14, 47], and a ViT-based token encoder with causal attention. An image is first discretized at a low level, after which the visual code embeddings are fed into the token encoder to capture high-level continuous semantics. The token encoder is pre-trained with only semantic supervision, specifically through text alignment and CLIP distillation, and is subsequently equipped with the autoregressive Transformer for end-to-end multimodal token prediction.

TokLIP offers three key advantages: (1) It effectively disentangles the training objectives for comprehension and generation across different semantic levels, eliminating the need for complex training strategies to balance these objectives. (2) It integrates CLIP-level semantics into Emu3-like unified autoregressive models, enhancing comprehension performance while preserving end-to-end multimodal autoregression on plain VQ tokens. (3) It leverages the strengths of VQ-based methods, enabling the direct use of state-of-the-art VQ tokenizers without addressing information loss from quantization training at the semantic level.

We empirically validate the effectiveness of TokLIP by comparing it with other tokenizers and integrating it with large language models (LLMs) for autoregressive multimodal comprehension and generation. Our analysis reveals three key findings: TokLIP demonstrates exceptional image representation capabilities, utilizing less than 20% of the pretraining data required by VILA-U, while significantly outperforming established tokenizers such as VILA-U and QLIP in zero-shot ImageNet classification. Additionally, TokLIP retains the original reconstruction capabilities of the low-level VQ tokenizer without necessitating specific optimization for generation and quantization. (2) *Multimodal comprehension at significantly reduced training cost:* When integrated with LLMs for comprehension tasks, TokLIP achieves comparable results using only 3% of the training data required by SynerGen-VL. This reveals that TokLIP unleashes the potential of discrete token-based multimodal LLMs to efficiently deliver advanced comprehension performance similar to that of the LLaVA series. (3) *Synergistic autoregressive image generation:* We demonstrate that high-level semantic features complement low-level code embeddings to enhance image generation, underscoring TokLIP’s ability to integrate diverse abstraction levels for robust multi-modal outputs.

2 Related Works

Vector-quantized tokenizers. Vector quantization (VQ) is effective in processing different modalities, especially images [55, 14, 65, 25, 4, 57, 44, 52], compressing high-dimensional raw data into discrete token sequences for efficient learning. The pioneering work VQ-VAE [55] proposes to discretize continuous representations by assigning them to the closest vectors in a learnable code-

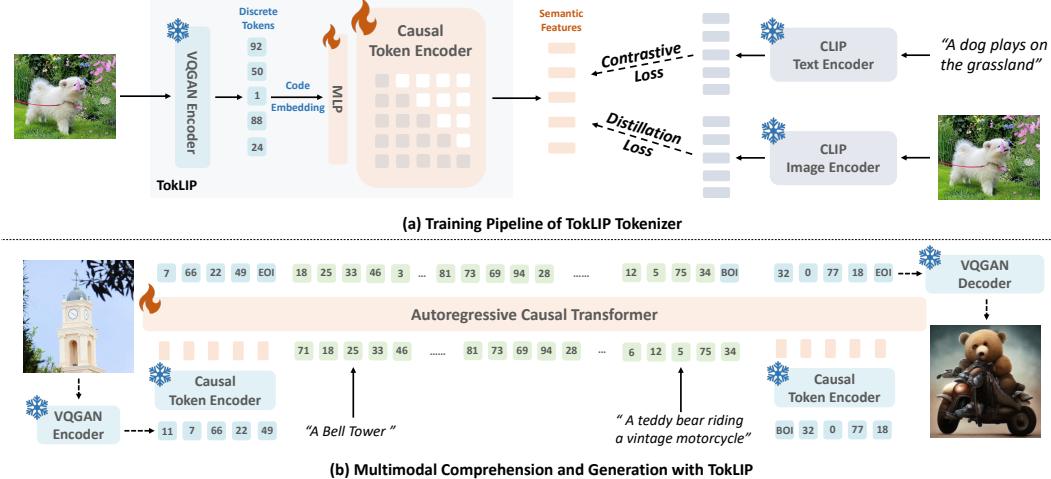


Figure 2: **Overview of TokLIP.** (a) Given an image, it is first quantized by an off-the-shelf VQGAN encoder before being fed into the causal token encoder equipped with a multi-layer perceptron. The token encoder is initialized from a pre-trained bidirectional CLIP vision encoder and semanticizes the discrete visual codes under a data-efficient learning strategy with contrastive loss and distillation loss. (b) The pretrained TokLIP is further employed in the framework of LLMs to tokenize images for end-to-end multimodal autoregressive learning. Given the semantic knowledge in the pretrained token encoder, we achieve advanced vision-language comprehension results with few computational overhead. Furthermore, TokLIP enables autoregressive image generation on plain VQ tokens.

book. VQGAN [14] incorporates adversarial and perceptual losses to further improve generation quality. Recent studies, such as LlamaGen [47], MAGVIT2 [38, 66] and VAR [51], demonstrate the great potential of VQ tokenizers in autoregressive image generation. Such tokenizers are naturally suited for generation tasks due to capturing low-level patterns, however, they struggle to provide necessary semantics for multimodal comprehension tasks, as evidenced in [59, 62]. In this work, we present TokLIP, which semanticizes VQ tokens to enhance multimodal comprehension while maintaining end-to-end multimodal autoregressive learning on plain VQ tokens when integrated into large multimodal models.

Large multimodal models with unified comprehension and generation. Exploring unified modeling of multimodal comprehension and generation in the framework of large language models has gained significant attention. Despite various attempts, the core challenge remains addressing the essential conflicts of comprehension and generation tasks. There are four categories of approaches to meet the challenge: (1) *External diffusion*: [19, 48, 53, 20, 49, 60] utilize external pretrained diffusion models linked through semantic features for image generation, resulting in independent processing of comprehension and generation. (2) *Internal diffusion*: [62, 72, 46] facilitate visual-language comprehension and generation within a single model; however, their use of denoising steps during image generation tasks prevents true fusion of images and text. (3) *Dual encoders*: [59, 39, 7] employ a continuous semantic encoder for comprehension and a discrete encoder for generation, which incurs increased computational overhead and results in independent processing of these tasks. (4) *Unified tokenizer*: [33, 50, 58] pioneered this line of research by training a single autoregressive Transformer with discrete image VQ tokens and text tokens. [42, 61, 71] further improve the multimodal comprehension performance by proposing new discrete visual tokenizers with high-level semantics.

It is evident that only the "unified tokenizer" category can effectively integrate comprehension and generation tasks, along with text and image tokens, into a single autoregressive model. However, state-of-the-art methods in this area face challenges due to conflicting training objectives—namely, reconstruction losses for generation and text-alignment losses for comprehension—as well as information loss from quantization training. This complexity makes stable training particularly difficult. We contend that it is more proper to semanticize low-level discrete visual tokens, as our proposed TokLIP, rather than to discretize high-level continuous visual features, as done in previous works.

3 Method

TokLIP is a discrete-to-continuous tokenizer tailored for integrating multimodal comprehension and generation tasks into a single autoregressive model. In this section, we outline the TokLIP architecture and training details in Section 3.1, explore its application to multimodal comprehension and generation under the framework of LLMs in Section 3.2, and conclude with a discussion and comparison of its main differences with existing tokenizers in Section 3.3.

3.1 TokLIP Tokenizer

Pioneering works [50, 58] in multimodal unification adopt VQ-based low-level tokenizers to encode images for both comprehension and generation tasks. These fidelity-oriented tokenizers primarily capture basic image features, such as contours and patterns, making them ideal for generation tasks. However, visual understanding tasks require high-level semantic representations to enhance interpretability and reasoning capabilities, as evidenced in [34, 73, 2]. The lack of such semantic knowledge in VQ-based tokenizers limits their effectiveness in multimodal understanding.

Overall architecture. To solve the problem, we explore the possibility of enriching VQ-based tokenizers [14, 47] with high-level semantic knowledge. CLIP [43, 70], trained on hundreds of millions of image-text pairs, has demonstrated remarkable success in visual understanding. To this end, we propose TokLIP, a framework that integrates fidelity encoders with CLIP vision encoders to simultaneously preserve both generation and comprehension capabilities in a single tokenizer, as illustrated in Figure 2 (a). Specifically, given an input image $I \in \mathbb{R}^{H \times W \times 3}$, we first quantize it using a VQGAN encoder \mathcal{E}_{vq} to obtain the code embeddings of discrete tokens $x_{\text{vq}} = \mathcal{E}_{\text{vq}}(I) \in \mathbb{R}^{(\frac{H \cdot W}{p \cdot p}) \times \hat{d}}$, where p is the downsample ratio and \hat{d} is the dimension of VQGAN codebook embeddings. The code embeddings x_{vq} are further mapped to CLIP input feature dimension d via a multi-layer perception (MLP) layer before being fed into a token encoder \mathcal{E}_{tok} which shares the same architecture as CLIP’s ViT encoder but uses causal attention. To conclude, the TokLIP tokenizer encodes the input image into semantic features $x_{\text{toklip}} \in \mathbb{R}^{(\frac{H \cdot W}{p \cdot p}) \times d}$ from discrete to continuous, denoted as

$$x_{\text{toklip}} = \mathcal{E}_{\text{tok}}(\text{MLP}(x_{\text{vq}})) = \mathcal{E}_{\text{tok}}(\text{MLP}(\mathcal{E}_{\text{vq}}(I))). \quad (1)$$

Our objective is to train the token encoder \mathcal{E}_{tok} along with the MLP layer on large-scale image-text data to effectively learn semantic representations from discrete token inputs.

Causal attention enables end-to-end multimodal autoregression. In the original CLIP architecture [43], the vision Transformer encoder employs bidirectional attention to capture the rich semantic relationships among image patches and facilitate a holistic understanding of the input. However, in the context of multimodal autoregressive learning, the bidirectional attention mechanism hinders the training paradigm of end-to-end next token prediction, where each token can only be predicted by attending to its preceding tokens. To address this discrepancy, we adopt causal attention for our token encoder, which enables end-to-end next multimodal token prediction when further equipping TokLIP with LLMs as shown in Figure 2 (b). In such a way, image and text signals, as well as comprehension and generation tasks, can be truly fused in a single causal Transformer.

Training objectives. We empirically find that directly training a causal ViT presents performance degradation compared to the original bidirectional ones, as the learning process struggles to adapt to this more constrained attention mechanism. We tailor a training strategy to mitigate the performance degradation. First, instead of training the token encoder from scratch, we initialize it with a pretrained bidirectional CLIP vision encoder. Our ablation studies, detailed in Section 4.5, show that inheriting knowledge from a well-trained CLIP model significantly reduces the performance gap between causal and bidirectional models. Next, we combine contrastive loss $\mathcal{L}_{\text{contra}}$ with distillation loss $\mathcal{L}_{\text{distill}}$ as the training objectives. Specifically, given an image-text pair (x_i, y_i) in a mini-batch (\mathbb{X}, \mathbb{Y}) , we obtain the image and text representations $z_{\text{img}}^i, z_{\text{text}}^i$ from pretrained CLIP teacher encoders. We denote the [CLS] token of x_{toklip}^i (Eq. (1)) as z_{toklip}^i . The contrastive loss and distillation loss are formulated as follows:

$$\begin{aligned} \mathcal{L}_{\text{contra}}(x_i, y_i) &= \text{InfoNCE}(z_{\text{toklip}}^i, z_{\text{text}}^i), \\ \mathcal{L}_{\text{distill}}(x_i) &= \text{MSE}(z_{\text{toklip}}^i, z_{\text{img}}^i), \end{aligned} \quad (2)$$

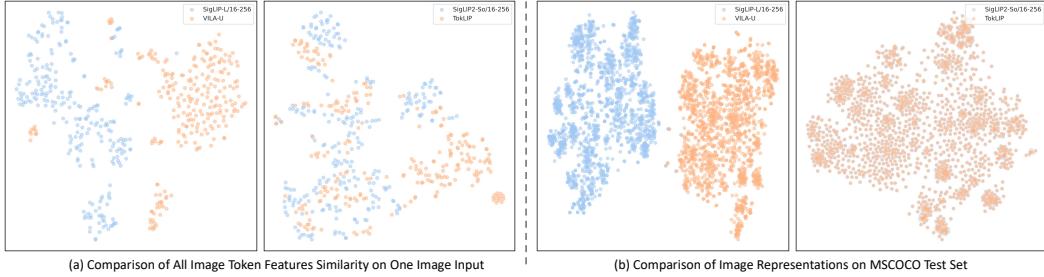


Figure 3: Empirical Comparison of Feature Representations. (a) t-SNE visualization of all tokenized features from VILA-U (left) and TokLIP (right), compared with patch features from respective SigLIP encoders for a single input image. (b) t-SNE visualization of 1500 image representations from VILA-U (left) and TokLIP (right), compared with representations from the SigLIP encoders. The representations are achieved using an attention pooler for cross-modal retrieval tasks. *TokLIP* features remain closely aligned with *SigLIP* across both levels.

where the overall training objective is $\mathcal{L} = \mathcal{L}_{\text{contra}} + \mathcal{L}_{\text{distill}}$ that enables the effective and data-efficient training of TokLIP tokenizer. Only the MLP layer and the token encoder are trainable. By freezing the VQGAN encoder, we decouple the reconstruction capability from the TokLIP training process, avoiding potential objective conflicts and easing the overall optimization.

Extensibility. We emphasize that TokLIP is not a fixed model architecture but a continuously improvable framework that can be enhanced with advanced VQ-based tokenizers and powerful CLIP encoders. For instance, we can replace VQGAN with superior tokenizers such as MAGVIT-v2 [66], Cosmos [1] or large-scale pre-trained tokenizers as utilized in Emu3 [58]. Additionally, we can substitute CLIP encoders with more effective ViT-based semantic encoders like SigLIP [70], AIM [13], or DFN [15]. Notably, TokLIP can be further extended by combining it with a diffusion head, following the success of GPT-4o [6, 40], to generate higher-quality and high-resolution images.

3.2 Multimodal Autoregression with TokLIP

The pretrained TokLIP tokenizer enables end-to-end multimodal autoregressive learning in a single causal Transformer, as demonstrated in Figure 2 (b), which unleashes the great potential of multimodal comprehension and generation applications. Specifically, during the autoregressive training of the Transformer, TokLIP is frozen to tokenize images, where the output image tokens are fed into the Transformer together with the text tokens. Only the cross-entropy loss is employed here for multimodal token prediction. In the paper, we take **image-to-text** and **text-to-image** as typical tasks for verifying the effectiveness of TokLIP in multimodal comprehension and generation applications, respectively. We concatenate the low-level output tokens from VQGAN with the high-level output features from the causal token encoder for feature fusion and mutual enhancement.

3.3 Discussion

Discretize CLIP or semanticize VQ? Another line of work aims to train a semantic tokenizer by quantizing high-level image features from the CLIP encoder into discrete tokens [61, 71]. These methods typically rely on joint training with both reconstruction and semantic losses. The key distinction between these approaches and TokLIP is that the former adopts a pipeline that transitions from continuous to discrete representations (*i.e.*, discretize high-level CLIP features), while TokLIP follows a discrete-to-continuous way (*i.e.*, semanticize low-level VQ tokens).

Intuitively, TokLIP properly disentangles the training objectives of comprehension and generation at different semantic levels (*i.e.*, low-level discrete tokens for generation and high-level continuous features for comprehension), alleviating the accuracy loss caused by objective conflicts in previous methods. Furthermore, VQ is known to be difficult to optimize due to the complicated training objectives. In TokLIP, we can directly exploit the state-of-the-art VQ tokenizers and inherit their generation capabilities without tailoring strategies for quantization training.

To better understand the advantages of disentangled training objectives, we conduct a feature analysis for comparison, as shown in Figure 3, demonstrating that TokLIP learns better semantics than VILA-U. We primarily compare the feature similarity between tokenizers and their initialized SigLIP [70, 54] encoders. Specifically, we extract the normalized features $x_{\text{toklip}}/x_{\text{vila-u}}$ and x_{siglip} for a single input image and compute the cosine similarity between $x_{\text{toklip}}/x_{\text{vila-u}}$ and x_{siglip} . Our results show that TokLIP achieves a similarity score of 0.83, outperforming VILA-U, which scores 0.75. We further visualize the normalized features using t-SNE in Figure 3 (a). On the left, the VILA-U features are far from the SigLIP-ViT-L/16 features, indicating that the feature spaces differ significantly. In contrast, TokLIP features are much closer to SigLIP2-So/16. These results suggest that the training process in VILA-U drastically alters the feature representations compared to the semantic encoder. We attribute this to the conflict between the reconstruction and semantic objectives and the information loss during the quantization process.

Additionally, we analyze the image representations used for retrieval and classification tasks, which are achieved by pooling image features x_{siglip} through an attention pooling head in SigLIP [70]. We randomly select 1,500 images from the MSCOCO test set and visualize these representations using t-SNE in Figure 3 (b). Notably, the representations from VILA-U are relatively distant from SigLIP, while TokLIP shares nearly the same distribution as SigLIP2. This aligns with TokLIP’s stronger retrieval performance demonstrated in Section 4.2. These analyses further show that our pipeline learns more similar semantics to pre-trained CLIP ViTs, thus achieving superior comprehension performance.

Comparison with CLIP Encoders. Pretrained on large-scale image-text pairs, CLIP [43, 54, 31] has achieved remarkable success in capturing high-level semantic features of images, making it highly effective for feature extraction in multimodal large language models (LLMs). Numerous studies [34, 73, 2, 37, 32] leverage CLIP’s continuous features by combining them with text tokens in LLMs and training on vision-language data to enhance performance in comprehension tasks. However, while CLIP excels at encoding high-level semantics, it often overlooks low-level structures and textual details, which are crucial for visual generation tasks. Models such as SEED-X [20], DreamLLM [12], and MetaMorph [53] incorporate CLIP’s continuous features with text tokens for autoregressive prediction but still rely on diffusion models [41] for generation inference to compensate for the lack of low-level details. In contrast, TokLIP proposes a unified tokenizer that integrates both low-level features and high-level semantics, enabling direct tokenization with CLIP encoders and allowing for more effective handling of visual generation tasks with VQGAN decoder.

4 Experiments

4.1 Experimental Setup

Implementation details. The TokLIP architecture integrates the text-to-image VQGAN tokenizer from LlamaGen [47] and a token encoder with ViT-So400M architecture. The downsample ratio is 16. Input resolutions are 256×256 for the base model and 384×384 for the large variant, resulting in a 256 and 576 tokens accordingly. We initialized the weights of token encoder from SigLIP2 [54] as a good basic for semantic perception. A two-layer MLP bridges the VQGAN and CLIP-based token encoder, mapping 8-dimensional VQGAN features as the input of token encoder. The TokLIP-large model trains on 80 million image-text pairs from Capfusion [67] and CC12M [5] and LAION-high-resolution, while the TokLIP-base model augments this with subsets of LAION400M [45], yielding a combined 125 million samples. We filter low-resolution images to ensure effective learning.

For multimodal comprehension, we utilize TokLIP as the input encoder and pass the output tokens through a 2-layer MLP to Qwen2.5-7B-Instruct [64]. Our training process consists of two stages. In Stage 1, we jointly train the MLP and the LLM on 4.5 million images with detailed captions for multimodal alignment, using data sourced from LLaVA-OneVision [26]. In Stage 2, we perform instruction tuning using the LLaVA-v1.5-mix-665K dataset [35]. For visual generation, we follow LlamaGen [47], training class-to-image generation models on ImageNet. We replace LlamaGen’s text-to-image VQGAN tokenizer with TokLIP.

Evaluation details. We evaluate the understanding capacity of our TokLIP on ImageNet classification [11] and MSCOCO 5K retrieval [8] tasks. For multimodal comprehension, we rigorously

Table 1: Comparative analysis of tokenizers on multimodal comprehension tasks. We adopt the LLaVA-v1.5 [35] training framework and combine tokenizers with Qwen2.5-7B-Instruct [64], training on LLaVA-v1.5 data. [‡] denotes the results are reported from the original paper.

Model	Data	Res.	POPE↑	MME-P↑	SEED↑	GQA↑	MMMU↑
Discrete Tokenizer (VQGAN)	LLaVA-v1.5 data	256	65.6	716.8	35.0	39.8	36.6
Discrete Semantic Tokenizer (VILA-U) [‡]	MMC4+ShareGPT4V(7M)	256	83.9	1336.2	56.3	48.3	-
TokLIP (Ours)	LLaVA-v1.5 data	256	81.2	1346.8	59.8	57.4	40.2
Discrete Tokenizer (VQGAN)	LLaVA-v1.5 data	384	69.2	773.7	34.6	41.2	34.8
TokLIP (Ours)	LLaVA-v1.5 data	384	82.7	1410.2	65.2	59.3	42.1

Table 2: Comparison with state-of-the-art tokenizers and CLIP encoders on zero-shot ImageNet classification and MSCOCO 5K retrieval tasks. TR refers to image-to-text retrieval, while IR denotes text-to-image retrieval.

Model	Dataset	Res.	IN Top1↑	COCO TR@1↑	COCO IR@1↑
EVA02-B/16 [16]	Merged-2B	224	74.7	58.74	42.15
CLIP-L/14 [43]	WIT400M	224	75.5	56.34	36.51
MetaCLIP-L/14 [63]	MetaCLIP400M	224	76.2	60.00	43.81
VILA-U [61]	COYO700M	256	73.3	62.56	45.37
QLIP [71]	DataComp1B	256	74.3	55.62	38.91
TokLIP	Mix 125M	256	76.4	64.06	48.46
OpenCLIP-ViT-L/14 [9]	LAION2B	336	75.3	63.36	46.51
SigLIP-B/16 [70]	WEBIL10B	384	76.5	67.74	49.90
SigLIP-B/16 [70]	WEBIL10B	512	79.1	68.72	50.55
VILA-U [61]	COYO700M	384	78.0	-	-
QLIP [71]	DataComp1B	384	79.1	60.86	43.00
TokLIP	Mix 80M	384	80.0	68.00	52.87

evaluate our model on several zero-shot vision-language benchmarks, which include POPE [29], MME-P (MME-Perception) [17], MMB (MMBench) [36], SEED (SEED-Bench Img) [27], GQA [21], MMMU [69], and MM-Vet [68], covering a diverse range of tasks that test both visual and linguistic understanding in a variety of real-world settings. For visual generation, we mainly report Fréchet inception distance (FID) on the ImageNet 256x256 benchmark.

4.2 Comparison with Other Tokenizers

Comprehension capacity. We compare TokLIP with state-of-the-art tokenizers [61, 71] and CLIP models [9, 43, 70] across both image classification and cross-modal retrieval tasks, shown in Table 2. It can be observed that TokLIP outperforms several SOTA CLIP-based encoders. For instance, TokLIP achieves an impressive 76.4% accuracy on ImageNet with a 256 input resolution, surpassing MetaCLIP [63] and EVA02 [16]. These results demonstrate that TokLIP effectively inherits semantic understanding from the pre-trained teacher and demonstrates competitive performance on downstream tasks. Additionally, at both 256x256 and 384x384 input resolutions, TokLIP outperforms VILA-U [61] and QLIP [71], particularly excelling in the retrieval task. TokLIP achieves 68.00% TR@1 and 52.87% IR@1, which are 7.14% and 9.87% higher than QLIP, respectively, at a 384 resolution. This suggests that TokLIP captures more semantic information compared to methods that discretize high-level image features. Notably, VILA-U is trained on COYO-700M [3], and QLIP uses DataComp1B [18], both of which involve significantly more resource-intensive training processes than ours. Future work will focus on enhancing TokLIP’s performance through the utilization of higher-quality training data [15, 63].

Enhanced vision tokenizer for MLLMs. To assess the effectiveness of different tokenizer architectures integrated into MLLMs, we compare TokLIP with VILA-U [61] and VQGAN [47] on multimodal comprehension tasks. Specifically, we use the LLaVA-v1.5 [35] framework to evaluate the understanding performance of VQGAN and TokLIP, with VILA-U results reported from the original paper. We train on the LLaVA-Pretrain-558K dataset for MLP pretraining and LLaVA-v1.5

Table 3: Comparison with state-of-the-art MLLMs on multimodal comprehension benchmarks. “Cmp.” and “Gen.” denote “comprehension” and “generation”, respectively. Models using external pretrained diffusion models are marked with \dagger .

Model	#Param	Res.	POPE \uparrow	MME-P \uparrow	MMB \uparrow	SEED \uparrow	GQA \uparrow	MMMU \uparrow	MM-Vet \uparrow
<i>Cmp. Only</i>									
LLaVA-v1.5-Phi-1.5 [62]	1.3B	256	84.1	1128.0	-	-	56.5	30.7	-
LLaVA-Phi [74]	2.7B	336	85.0	1335.1	59.8	-	-	-	28.9
LLaVA-v1.5 [35]	7B	336	85.9	1510.7	64.3	58.6	62.0	35.4	31.1
InstructBLIP [10]	7B	224	-	-	36.0	53.4	49.2	-	26.2
Qwen-VL-Chat [2]	7B	448	-	1487.5	60.6	58.2	57.5	-	-
IDEFICS-9B [24]	8B	224	-	-	48.2	-	38.4	-	-
InstructBLIP [10]	13B	224	78.9	1212.8	-	-	49.5	-	25.6
<i>Cmp. and Gen. Continuous</i>									
DreamLLM \dagger [12]	7B	224	-	-	-	-	-	-	36.6
LaVIT \dagger [22]	7B	224	-	-	58.0	-	46.8	-	-
MetaMorph \dagger [53]	8B	384	-	-	75.2	71.8	-	-	-
NExT-GPT \dagger [60]	13B	224	-	-	-	57.5	-	-	-
SEED-X \dagger [20]	13B	448	84.1	1457.0	70.1	66.5	49.1	35.6	43.0
ILLUME [56]	7B	224	88.5	1445.3	65.1	72.9	-	38.2	37.0
Janus [59]	1.5B	384	87.0	1338.0	69.4	63.7	59.1	30.5	34.3
<i>Cmp. and Gen. Discrete</i>									
Show-o [62]	1.3B	256	73.8	948.4	-	-	48.7	25.1	-
Show-o [62]	1.3B	512	80.0	1097.2	-	-	58.0	26.7	-
LWM [33]	7B	256	75.2	-	-	-	44.8	-	9.6
D-Dit [30]	2.0B	256	84.0	1124.7	-	-	59.2	-	-
Emu3-Chat [58]	8B	512	85.2	1244.0	58.5	68.2	60.3	31.6	37.2
Chameleon [50]	7B	256	-	-	-	-	-	22.4	8.3
Orthus [23]	7B	256	79.6	1265.8	-	-	52.8	28.2	-
TokenFlow-B [42]	13B	224	84.0	1353.6	55.3	60.4	59.3	34.2	22.4
TokenFlow-L [42]	13B	256	85.0	1365.4	60.3	62.6	60.3	34.4	27.7
SynerGen-VL [28]	2.4B	512	85.3	1381.0	53.7	62.0	59.7	34.2	34.5
VILA-U [61]	7B	256	83.9	1336.2	-	56.3	48.3	-	27.7
VILA-U [61]	7B	384	85.8	1401.8	-	59.0	60.8	-	33.5
TokLIP (Ours)	7B	384	84.1	1488.4	67.6	70.4	59.5	43.1	29.8

mix-665K for instruction tuning. The comprehensive evaluations are presented in Table 1. Our results demonstrate that TokLIP consistently outperforms VQGAN across all metrics and on nearly all benchmarks. For 256 input resolution, TokLIP achieves a clear performance advantage over both VILA-U and VQGAN. Notably, VILA-U is trained on 7M vision-language pairs, significantly more than the LLaVA-v1.5 dataset. Despite these variations in training data size, input resolution changes have a negligible impact on VQ-based tokenizers’ performance, pointing to the inherent limitations in scaling their comprehension capacity. These findings further support our hypothesis that: (1) discrete tokenizers struggle with comprehension tasks due to their lack of semantic depth, and (2) discrete semantic tokenizers face performance degradation due to conflicting training objectives. Overall, our results highlight TokLIP’s superior ability to capture semantic-level understanding, making it a more effective solution for multimodal comprehension tasks involving discrete inputs.

4.3 Multimodal Comprehension

Main results. We present a comprehensive evaluation of TokLIP in comparison to state-of-the-art models in both understanding-only MLLMs and unified understanding-and-generation MLLMs. Illustrated in Table 3, TokLIP, as a discrete tokenizer, achieves competitive performance when compared to existing methods that rely on continuous inputs from CLIP vision encoders. For instance, TokLIP performs on par with LLaVA-v1.5 across most benchmarks, while outperforming it on the SEED Bench, MMB, and MMMU datasets. These results demonstrate that TokLIP effectively learns semantic features from discrete visual tokens, significantly enhancing multimodal understanding

capabilities. By leveraging the pre-trained CLIP model, TokLIP successfully incorporates high-level semantic information, bridging the gap between visual tokens and CLIP-level semantics.

In the context of unified models with discrete inputs, which represents the most comparable setting, TokLIP substantially outperforms Chameleon and Show-o across all evaluated benchmarks. Moreover, TokLIP surpasses Emu3-chat on most benchmarks, despite utilizing much less training data and operating at a lower input resolution (384 vs. 512). This reinforces the effectiveness of TokLIP in successfully semanticizing VQ codes with CLIP-level semantics, showing that high-quality semantic understanding can be achieved with fewer resources. When compared to VILA-U, TokLIP also delivers superior performance, further supporting our hypothesis that transitioning from discrete to continuous semantics offers a more effective approach, as discussed in Section 3.3. These empirical results highlight TokLIP’s potential as a foundational tokenizer for unified understanding and generation tasks, paving the way for more advanced multimodal models.

Data efficiency. Approaches [58, 62] that directly optimize discrete VQ-based tokenizers typically rely on large-scale datasets. For example, SynerGen-VL [28] is trained on 170 million samples with task-specific prompts for visual understanding, which is often unaffordable for most researchers. Similarly, discrete semantic models [61, 42, 71] like VILA-U require large-scale training sets to discretize high-level semantics effectively. In contrast, we emphasize that TokLIP provides an efficient solution for tokenizer training, achieving strong performance with just 80M image-text pairs and fewer resources for visual understanding. This efficiency is due to our pipeline, which bridges the gap from discrete tokens to continuous representations, enabling effective learning with less data and computational overhead.

4.4 Visual Generation

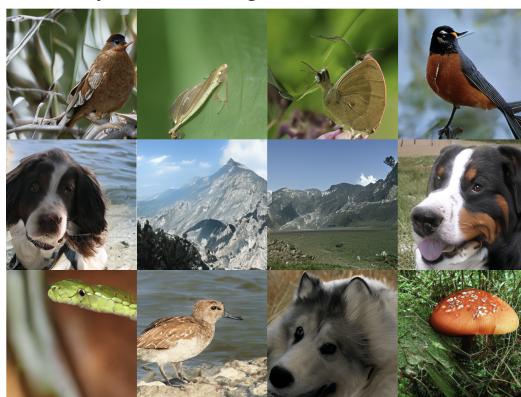
Class-conditional image generation. We follow the training pipeline of LlamaGen [47] to evaluate the generation capacity on ImageNet benchmarks. Specifically, we replace the text-to-image VQGAN tokenizer with our TokLIP tokenizer to train both GPT-Base and GPT-Large models for different epochs, while keeping the training settings and hyperparameters consistent. We compare the discrete VQGAN tokenizer with TokLIP in Table 4. All experiments are conducted with the same sampling configuration: $\text{cfg} = 2.5$, $\text{top-k} = 0$ (all), $\text{top-p} = 1.0$, and $\text{temperature} = 1.0$. Our results show that all models using the TokLIP tokenizer achieve a lower FID than those using the discrete VQGAN tokenizer, demonstrating that the incorporation of high-level semantics enhances image generation. The improvement is less significant at the 256 input image size compared to the 384 resolution, which we attribute to the fact that semantics are more beneficial when dealing with longer image tokens.

Qualitative evalutaion. We show class-conditional generation examples with TokLIP in Figure 4. Images are generated at 384x384 and resized to 256x256 for display. It can be observed that the model generates high-quality, aesthetically pleasing images that effectively capture visual concepts.

Table 4: Performance on class-conditional 256x256 ImageNet 50K benchmark. The downsample ratio is 16.

Tokenizer	Model	Res.	epochs	FID ↓
VQGAN	GPT-B	256	50	13.21
TokLIP	GPT-B	256	50	13.12
VQGAN	GPT-B	384	50	14.48
TokLIP	GPT-B	384	50	12.37
VQGAN	GPT-B	256	300	7.43
TokLIP	GPT-B	256	300	7.29
VQGAN	GPT-B	384	300	9.70
TokLIP	GPT-B	384	300	7.64
VQGAN	GPT-L	256	50	5.96
TokLIP	GPT-L	256	50	5.72
VQGAN	GPT-L	384	50	7.28
TokLIP	GPT-L	384	50	6.19

Figure 4: Visualizations of class-conditional generation by TokLIP using LlamaGen-B framework.



4.5 Ablation Studies

TokLIP training. We explore several training strategies for our TokLIP models. Specifically, we evaluate three key aspects: (1) whether to initialize the vision encoder with pre-trained weights, (2) the method for projecting VQ codes into the CLIP feature space, and (3) the design of the learning objectives. To analyze these strategies, we implement various ablation models based on OpenCLIP ViT-B/16 [9] and pre-train on CC3M dataset [5], listing the results in Table 5.

For **vision initialization**, omitting the pre-trained vision encoder significantly degrades performance, as large-scale pre-training provides the CLIP encoder with a deep understanding of image semantics, enabling it to capture complex patterns and textures. By leveraging pre-trained weights, TokLIP benefits from this rich knowledge, improving its semantic comprehension capabilities.

For **projection methods**, we compare MLP projection with a direct codebook-based mapping of discrete tokens into the CLIP feature space. As shown in Table 5, MLP projection outperforms codebook mapping, likely due to the large codebook size (16,532 entries), which complicates the learning of meaningful representations. A more compact or optimized codebook could enhance performance by focusing on a more relevant set of visual features, a direction we leave for future work.

Regarding **learning objectives**, our results show that distillation from a teacher model significantly enhances TokLIP’s performance. Distilling the [CLS] token features from the teacher improves understanding of image-text relationships, while distilling all tokens does not lead to gains. This suggests that the [CLS] token, representing global input understanding, is most beneficial for distillation. The lack of improvement when distilling all tokens may stem from the differing token dependencies between the causal attention mechanism in TokLIP and the bidirectional dependencies in the teacher model, leading to misalignment in token-level representations.

Based on these empirical findings, we adopt a training strategy that loads pre-trained vision encoder weights to retain semantic knowledge, employs MLP projection for efficient code mapping, and integrates contrastive learning with knowledge distillation for effective learning.

Table 5: Ablations of different vision initialization, projection methods and learning objectives for training TokLIP on CC3M.

Vision Encoder	Mapping	Training Objective	Top 1 Acc.	Top 5 Acc.
Init	Patch	$\mathcal{L}_{\text{contra}}$	18.50	39.97
Init	MLP	$\mathcal{L}_{\text{contra}}$	15.73	36.81
ViT-B/16	MLP	$\mathcal{L}_{\text{contra}}$	32.93	60.33
ViT-B/16	Soft MLP	$\mathcal{L}_{\text{contra}} + \mathcal{L}_{\text{distill}}$	51.53	79.85
ViT-B/16	Codebook	$\mathcal{L}_{\text{contra}} + \mathcal{L}_{\text{distill}}$	35.17	65.07
ViT-B/16	MLP	$\mathcal{L}_{\text{contra}} + \mathcal{L}_{\text{distill_all}}$	51.04	79.29
ViT-B/16	MLP	$\mathcal{L}_{\text{contra}} + \mathcal{L}_{\text{distill}}$	52.46	80.52

Bidirectional TokLIP. As introduced in Section 3.1, we apply a causal attention module in TokLIP for unified understanding and generation. Our causal attention-based TokLIP demonstrates superior semantic understanding compared to the VILA-U. In this section, we implement a bidirectional attention version using OpenCLIP-ViT-B/16 [9] to investigate whether a bidirectional architecture could further enhance comprehension tasks. We pre-train both the causal and bidirectional models on the CC3M and CC12M datasets with the same training setups. Results in Table 6 show that bidirectional-based models surpass causal-based models on classification and retrieval tasks. This suggests that a bidirectional TokLIP can further improve comprehension capacity for MLLMs, reinforcing the idea that learning semantics from discrete tokens is more effective than discretizing continuous high-level features.

Feature fusion. We explore different fusion functions to complement the low-level tokens from VQGAN with semantic-level features from the causal token encoder. Specifically, we compare the following strategies: (1) direct sum, (2) weighted sum with a learnable parameter, and (3) concatenation within the LlamaGen-Base (384x384) framework. We evaluate all models after training for 50 epochs, generating 10,000 samples for quick evaluation. As shown in Table 7, concatenation

Table 6: Ablations on bidirectional and causal attention in TokLIP.

Vision Encoder	Attention	Dataset	IN Top1↑	COCO TR@1↑	COCO IR@1↑
ViT-B/16	Bidirectional	CC3M	56.32	44.66	32.66
	Causal	CC3M	52.46	41.30	30.13
ViT-B/16	Bidirectional	CC12M	58.99	47.32	33.78
	Causal	CC12M	55.02	42.80	31.06

yields the lowest FID, which we adopt as the final approach. Additionally, all fusion functions outperform the discrete VQGAN, further demonstrating that incorporating high-level semantics enhances visual generation capabilities.

Table 7: Ablations of fusion function for pixel-level features and semantic-level features using LlamaGen-B. The inference setting is top-k = 0 (all), top-p = 1.0, temperature = 1.0 for all experiments.

Method	Input_size	Epochs	Cfg_ratio	Generated_samples	FID ↓
VQGAN	384	50	2.5	10K	17.31
TokLIP_sum	384	50	2.5	10K	16.37
TokLIP_weighted_sum	384	50	2.5	10K	15.46
TokLIP_concatenation	384	50	2.5	10K	14.93

5 Conclusion

In this paper, we introduce TokLIP, a discrete-to-continuous visual tokenizer that enables end-to-end autoregressive training of a single Transformer with sequences of multimodal discrete tokens. Our approach offers three key advantages: it disentangles the training objectives for comprehension and generation, enhances comprehension performance, and efficiently uses state-of-the-art VQ tokenizers. Through empirical validation, we demonstrate that TokLIP achieves exceptional image representation capabilities, reduces training cost, and enhances image generation capability. Our results underscore the potential of TokLIP to integrate diverse abstraction levels for robust multi-modal outputs.

Appendix

A Additional Implementation Details

Hyperparameters for TokLIP. We list detailed hyperparameters for training TokLIP in Table A1.

Table A1: Detailed Hyperparameters for training TokLIP.

Config	TokLIP	TokLIP
Resolution	256x256	384x384
Optimizer	AdamW	
Optimizer momentum	$\beta_1=0.9, \beta_2=0.98$	
LR schedule	CosineLRScheduler	
Weight decay	0.1	
Warmup steps	500	
Base LR	1e-5	1e-5
Batch size	1792	512

Mapping function. For the multi-layer perceptron projection between VQGAN and token encoder, we first map the 8-dimensional features from VQGAN to 4×dimensional hidden states of CLIP features, followed by a GeLU activation layer, and then use another linear layer to map to the CLIP feature dimension. For the codebook approach, we directly establish a large embedding layer to map each code to the CLIP feature space.

B More Experimental Results

B.1 Further Discussion

Effects of all token distillation on bidirectional TokLIP. As analyzed in Section 4.5, distilling features from all tokens does not enhance causal-attention TokLIP due to different attention dependencies. Here, we ablate the all-token distillation loss for training bidirectional-attention TokLIP on CC3M dataset. As shown in Table B2, distilling all tokens significantly improves performance on downstream tasks. This indicates that bidirectional-attention TokLIP shares a more similar architecture with the teacher model and thus inherits more semantic knowledge. This provides a promising approach to better empower discrete tokens with semantics for multimodal understanding tasks.

Table B2: Ablation of all token distillation for bidirectional TokLIP.

Vision Encoder	Attention	Distillation	IN Top1↑	COCO TR@1↑	COCO IR@1↑
ViT-B/16	Bidirectional	Single token	56.32	44.66	32.66
ViT-B/16	Bidirectional	All token	59.47	47.32	35.60

References

- [1] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.
- [2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.
- [3] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Sae-hoon Kim. Coyo-700m: Image-text pair dataset. <https://github.com/kakaobrain/coyo-dataset>, 2022.
- [4] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11315–11325, 2022.
- [5] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 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*, pages 3558–3568, 2021.
- [6] Sixiang Chen, Jinbin Bai, Zhuoran Zhao, Tian Ye, Qingyu Shi, Donghao Zhou, Wenhao Chai, Xin Lin, Jianzong Wu, Chao Tang, et al. An empirical study of gpt-4o image generation capabilities. *arXiv preprint arXiv:2504.05979*, 2025.
- [7] 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.
- [8] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.
- [9] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2818–2829, 2023.
- [10] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.
- [11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [12] Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jian-jian Sun, Hongyu Zhou, Haoran Wei, et al. Dreamllm: Synergistic multimodal comprehension and creation. *arXiv preprint arXiv:2309.11499*, 2023.
- [13] Alaaeldin El-Nouby, Michal Klein, Shuangfei Zhai, Miguel Angel Bautista, Alexander Toshev, Vaishaal Shankar, Joshua M Susskind, and Armand Joulin. Scalable pre-training of large autoregressive image models. *arXiv preprint arXiv:2401.08541*, 2024.
- [14] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.
- [15] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. *arXiv preprint arXiv:2309.17425*, 2023.
- [16] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *Image and Vision Computing*, 149:105171, 2024.

- [17] 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. *arXiv preprint arXiv:2306.13394*, 2023.
- [18] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36:27092–27112, 2023.
- [19] Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. Making llama see and draw with seed tokenizer. *arXiv preprint arXiv:2310.01218*, 2023.
- [20] Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*, 2024.
- [21] Drew A Hudson and Christopher D Manning. 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*, pages 6700–6709, 2019.
- [22] Yang Jin, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, Bin Chen, Chenyi Lei, An Liu, Chengru Song, Xiaoqiang Lei, et al. Unified language-vision pretraining with dynamic discrete visual tokenization. *arXiv preprint arXiv:2309.04669*, 2023.
- [23] Siqi Kou, Jiachun Jin, Chang Liu, Ye Ma, Jian Jia, Quan Chen, Peng Jiang, and Zhijie Deng. Orthus: Autoregressive interleaved image-text generation with modality-specific heads. *arXiv preprint arXiv:2412.00127*, 2024.
- [24] Hugo Laurençon, Daniel van Strien, Stas Bekman, Leo Tronchon, Lucile Saulnier, Thomas Wang, Siddharth Karamcheti, Amanpreet Singh, Giada Pistilli, Yacine Jernite, and et al. Introducing idefics: An open reproduction of state-of-the-art visual language model, 2023. URL <https://huggingface.co/blog/idefics>.
- [25] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11523–11532, 2022.
- [26] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.
- [27] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.
- [28] Hao Li, Changyao Tian, Jie Shao, Xizhou Zhu, Zhaokai Wang, Jinguo Zhu, Wenhan Dou, Xiaogang Wang, Hongsheng Li, Lewei Lu, et al. Synergen-vl: Towards synergistic image understanding and generation with vision experts and token folding. *arXiv preprint arXiv:2412.09604*, 2024.
- [29] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.
- [30] Zijie Li, Henry Li, Yichun Shi, Amir Barati Farimani, Yuval Kluger, Linjie Yang, and Peng Wang. Dual diffusion for unified image generation and understanding. *arXiv preprint arXiv:2501.00289*, 2024.
- [31] Haokun Lin, Haoli Bai, Zhili Liu, Lu Hou, Muyi Sun, Linqi Song, Ying Wei, and Zhenan Sun. Mope-clip: Structured pruning for efficient vision-language models with module-wise pruning error metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27370–27380, 2024.

- [32] Haokun Lin, Haobo Xu, Yichen Wu, Jingzhi Cui, Yingtao Zhang, Linzhan Mou, Linqi Song, Zhenan Sun, and Ying Wei. Duquant: Distributing outliers via dual transformation makes stronger quantized llms. *Advances in Neural Information Processing Systems*, 37:87766–87800, 2024.
- [33] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with ringattention. *arXiv e-prints*, pages arXiv–2402, 2024.
- [34] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.
- [35] 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.
- [36] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024.
- [37] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhusuo Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.
- [38] Zhuoyan Luo, Fengyuan Shi, Yixiao Ge, Yujiu Yang, Limin Wang, and Ying Shan. Open-magvit2: An open-source project toward democratizing auto-regressive visual generation. *arXiv preprint arXiv:2409.04410*, 2024.
- [39] Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Liang Zhao, et al. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. *arXiv preprint arXiv:2411.07975*, 2024.
- [40] OpenAI. Addendum to gpt-4o system card: 4o image generation, 2025. URL <https://openai.com/index/gpt-4o-image-generation-system-card-addendum/>. Accessed: 2025-04-02.
- [41] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.
- [42] Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K Du, Zehuan Yuan, and Xinglong Wu. Tokenflow: Unified image tokenizer for multimodal understanding and generation. *arXiv preprint arXiv:2412.03069*, 2024.
- [43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [44] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.
- [45] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.
- [46] Weijia Shi, Xiaochuang Han, Chunting Zhou, Weixin Liang, Xi Victoria Lin, Luke Zettlemoyer, and Lili Yu. Llamafusion: Adapting pretrained language models for multimodal generation. *arXiv preprint arXiv:2412.15188*, 2024.
- [47] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

- [48] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Emu: Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*, 2023.
- [49] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14398–14409, 2024.
- [50] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.
- [51] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024.
- [52] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2025.
- [53] Shengbang Tong, David Fan, Jiachen Zhu, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. Metamorph: Multimodal understanding and generation via instruction tuning. *arXiv preprint arXiv:2412.14164*, 2024.
- [54] 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.
- [55] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.
- [56] Chunwei Wang, Guansong Lu, Junwei Yang, Runhui Huang, Jianhua Han, Lu Hou, Wei Zhang, and Hang Xu. Illume: Illuminating your llms to see, draw, and self-enhance. *arXiv preprint arXiv:2412.06673*, 2024.
- [57] Luting Wang, Yang Zhao, Zijian Zhang, Jiashi Feng, Si Liu, and Bingyi Kang. Image understanding makes for a good tokenizer for image generation. *Advances in Neural Information Processing Systems*, 37:31015–31035, 2025.
- [58] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.
- [59] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*, 2024.
- [60] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023.
- [61] Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024.
- [62] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024.
- [63] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. *arXiv preprint arXiv:2309.16671*, 2023.

- [64] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.
- [65] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021.
- [66] Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, et al. Language model beats diffusion-tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023.
- [67] Qiying Yu, Quan Sun, Xiaosong Zhang, Yufeng Cui, Fan Zhang, Yue Cao, Xinlong Wang, and Jingjing Liu. Capsfusion: Rethinking image-text data at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14022–14032, 2024.
- [68] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.
- [69] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024.
- [70] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023.
- [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. *arXiv preprint arXiv:2408.11039*, 2024.
- [73] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.
- [74] Yichen Zhu, Minjie Zhu, Ning Liu, Zhicai Ou, Xiaofeng Mou, and Jian Tang. Llava-phi: Efficient multi-modal assistant with small language model. *arXiv preprint arXiv:2401.02330*, 2024.