WinTok: A Win-Win Hybrid Tokenizer via Decomposing Visual Understanding and Generation with Transferable Tokens

Yiwei Guo*[1], Shaobin Zhuang*[3], Zhipeng Huang[2], Canmiao Fu[2]
Chen Li[2], Jing LYU[2], and Yali Wang†[1]

[1] Shenzhen Key Laboratory of Computer Vision and Pattern Recognition, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[2] WeChat Vision, Tencent Inc.
[3] Shanghai Jiao Tong University

Abstract. Building a unified visual tokenizer is essential for bridging the gap between visual understanding and generation. Yet existing approaches struggle with the inherent conflict between these tasks, as a single token space is forced to support both high-level semantic abstraction and low-level pixel reconstruction. We propose **WinTok**, a concise hybrid tokenizer that achieves a win-win performance by explicitly decoupling the two objectives. WinTok supplements pixel tokens with a set of learnable semantic tokens, effectively mitigating cross-task interference without incurring the computational overhead of dual tokenizers. To further enhance understanding capability, we introduce an asymmetric token distillation mechanism: the semantic tokens are guided by pretrained semantic embeddings from any visual foundation model, enabling them to inherit strong discriminative power while maintaining flexibility. Across **10** challenging benchmarks, WinTok delivers consistent improvements in reconstruction, understanding, and generation. Trained on only 50M open-source data, WinTok surpasses the strong baseline UniTok by **11.2%** in classification accuracy and achieves a competitive reconstruction rFID of **0.41**, despite using substantially less training data. Code is released at https://github.com/WeChatCV/WinTok.

Keywords: Unified tokenizer · Hybrid encoding · Transferable tokens

1 Introduction

The emergence of large language models (LLMs) has fundamentally reshaped the landscape of natural language processing by introducing a unified next-token prediction paradigm [3,9,80]. Building upon this principle, recent efforts have extended the unified autoregressive framework to jointly model visual understanding and generation within a single multimodal architecture [38,75,77,78,93,108]. However, these unified models face a persistent dilemma: Despite these advances,

* Work done during interns at WeChat Vision, Tencent Inc. † Corresponding Author.

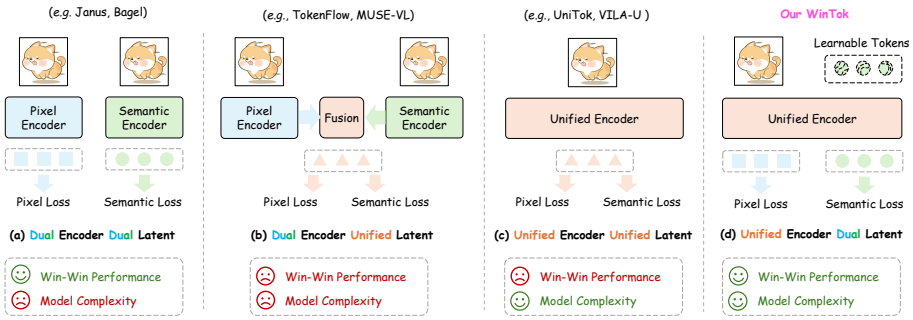


Fig. 1: Comparsion of different tokenization paradigms.(a) Dual encoders with separate representations obtain plausible performance at the cost of model complexity. (b-c) Previous unified tokenizers face representation conflict due to contradictory optimization objectives, thus leading to performance trade-offs. (d) Our WinTok decomposes visual understanding and generation with learnable tokens, achieving a win-win performance with reduced complexity.

unified multimodal models confront an inherent tension: visual understanding and visual generation require tokens with distinct granularities and representational forms. Visual understanding favors high-level continuous tokens that capture abstract and semantically rich representations for image comprehension [56, 57, 85]. In contrast, visual generation relies on low-level discrete tokens to enable precise and high-fidelity pixel synthesis [29, 74, 89].

To reconcile these divergent requirements, most existing unified multi-modal models typically employ a separate visual tokenizer for each task [19, 88, 110], as shown in Fig. 1 (a). Semantic encoders [67, 81, 103] extract continuous tokens for understanding, while pixel encoders [29, 82, 99] generate discrete tokens for generation. However, this would introduce significant complexity, without achieving fundamental model unification. To tackle this core issue, recent efforts have been made to construct a unified tokenizer. One type is the dual encoder with fusion (*e.g.*, shared mapping or MLP) [66, 94], as shown in Fig. 1 (b), which inherit the complexities of dual architectures. Alternatively, encoder unification is preferable [60, 105], as shown in Fig. 1 (c). But unified encoders force a single set of visual tokens to handle both high-level semantic abstraction and low-level pixel reconstruction, leading to performance trade-offs between understanding and generation due to optimization conflict [53, 73, 76, 90, 102].

To address this conflict, we introduce WinTok, which is a hybrid tokenizer that achieves a win-win performance, by decomposing visual understanding and generation with transferable tokens. As illustrated in Fig. 1 (d), WinTok comprises two types of tokens, including pixel and semantic tokens. On one hand, the unified encoder receives the input image to generate the pixel tokens. Since these pixel tokens often contain rich local details, they are suitable for visual generation. On the other hand, we introduce a set of learnable tokens as extra input to the unified encoder to obtain semantic tokens. To effectively equip these learnable tokens with high-level semantics, we introduce an asymmetric

token distillation paradigm in our WinTok. Specifically, we leverage any visual foundation model [67, 81, 103] as the semantic teacher to extract well-pretrained semantic tokens of the input image. Then, we treat them as semantic supervision to guide the optimization of learnable tokens. We refer to this as asymmetric token distillation, since semantic knowledge is transferred from the input image to learnable tokens. Through this token-level knowledge transfer paradigm, the learnable tokens progressively inherit and internalize the representational capabilities of foundation models, thereby acquiring enhanced capacity to model high-level visual semantics.

The interaction between the two types of tokens in the unified encoder enables them to adaptively integrate complementary contextual information, optimizing local details for generation and global semantics for understanding. Extensive experiments demonstrate that WinTok outperforms pioneering unified tokenizers across 10 mainstream benchmarks for visual reconstruction, understanding, and generation. Specifically, on ImageNet-1K [25] validation set, WinTok achieves an **82.0%** Top-1 accuracy, surpassing the strong counterpart UniTok [60] by **11.2%**, while maintaining a competitive reconstruction quality with an rFID of **0.41** using significantly fewer data and a reduced codebook size. Moreover, WinTok demonstrates leading performance on downstream multimodal comprehension and visual generation, outperforming UniTok on POPE [51] by 3.3% and delivering a better performance on GenEval (0.76 vs. 0.59). These results indicate that WinTok effectively balances the needs of both visual understanding and generation tasks. We believe that WinTok offers a promising direction for future research in unified multimodal modeling.

2 Related Work

2.1 Visual Tokenizer for Understanding

To extend the capability of large language models (LLMs) to comprehend visual content, numerous multimodal large language models (MLLMs) have been proposed [21, 48, 56, 57, 85]. These methods typically employ a language-aligned semantic visual tokenizer [67, 81, 103] to extract continuous visual tokens that encapsulate high-level semantic information from images. However, these tokenizers are primarily designed for visual understanding tasks, and thus may not effectively capture the fine-grained details necessary for faithful image reconstruction or generation [73, 76].

2.2 Visual Tokenizer for Generation

In the visual generation domain, mainstream approaches [12, 29, 64, 68, 79] utilize reconstruction-oriented visual tokenizers [42, 82, 99] to map images into a compact latent space, which reduces computational overhead and modeling complexity. For example, diffusion models [61, 64, 68] employ VAE [42, 43] to encode images into continuous tokens while autoregressive models [12, 74, 79]

leverage VQVAE [82, 99] to convert images into discrete tokens. More recently, some works [8, 14, 106, 107] have explored to transfer vision foundation models (VFMs) [63, 67, 81] as visual tokenizers for generative tasks. Others introduce 1D-tokenizers [2, 49, 100] for better token efficiency, or multi-codebook quantization [5, 39] for better codebook learning. Nonetheless, these reconstruction-oriented tokenizers may not effectively capture high-level semantic information essential for understanding tasks [66, 90].

2.3 Unified Visual Tokenizer

Early unified multimodal models (UMMs) [19,24,88] typically adopt separate visual tokenizers for understanding and generation tasks, which leads to increased model complexity and training costs. To bridge this gap, recent efforts have focused on unified visual tokenizers. Some methods combine semantic and pixel encoders via late-fusion, targeting either distinct codebook learning [66,94] or hierarchical feature integration [22,53]; however, this paradigm remains cumbersome. Alternatively, single-encoder approaches yield unified representations through explicit semantic alignment [91, 105], shared latent optimization [50, 60, 76], or dual-stream balancing [73,102]. While structurally simpler, they often struggle to balance conflicting training objectives. Though contemporary work VQRAE [27] addresses this via a hybrid tokenizer, it relies on an elaborate two-stage training scheme. In contrast, WinTok mitigates these obstacles by introducing transferable tokens to decompose the conflicting objectives, enabling a harmonious balance between visual understanding and generation.

3 Method

In this section, we first provide the preliminary knowledge on typical tokenizers used for visual understanding and generation. Motivated by representation conflict between these tasks, we then elaborate on our WinTok framework, incorporating asymmetric token distillation.

3.1 Preliminary

Semantic Tokenizers for Visual Understanding. Semantic tokenizers convert raw pixels into a compact sequence of high-level semantic tokens for visual understanding. The scaling of data and models in recent foundation models has demonstrated remarkable discriminative power by utilizing vision transformers [26] alongside various self-supervised learning and multimodal alignment strategies [34]. For example, MAE [35] learns scalable visual representations through masked image modeling, while DINOv2 [63] acquires semantic features via self-distillation across augmented views. More recently, contrastive-based methods [67, 81, 103] learn language-grounded visual representations through large-scale image-text alignment objectives. These models are considered effective semantic tokenizers because they leverage continuous semantic tokens to address downstream understanding tasks [1, 56, 57].

Pixel Tokenizers for Visual Generation. To facilitate auto-regressive visual generation, several pixel tokenizers have been developed using Vector-Quantized Variational Autoencoders (VQVAEs) [29,82,99]. These tokenizers generally consist of an encoder, a quantizer, and a decoder. Given an input image, the encoder converts it into a set of continuous latent tokens. The quantizer then transforms these continuous tokens into discrete visual tokens by identifying their nearest embeddings within a learnable codebook. Finally, the decoder reconstructs the image from the quantized tokens. By combining pixel reconstruction loss with codebook learning loss [74,99], these tokenizers achieve high-fidelity reconstruction and generation.

Conflict between Understanding and Generation. As discussed above, visual understanding and generation typically depend on different tokenizers with different visual granularities and formulations to process images. Hence, it would be inappropriate to use pixel tokenizer for understanding and use semantic tokenizer for generation. We further conduct an experiment on ImageNet [25] to validate this observation using two state-of-the-art tokenizers, *i.e.*, SigLIP2 [81] as the semantic tokenizer, and WeTok [109] as the pixel tokenizer. As shown in Tab. 1, SigLIP2 exhibits strong capability in understanding (*e.g.*, classification accuracy) but deteriorates in reconstruction (*e.g.*, rFID). In contrast, WeTok excels in reconstruction quality but shows significant declines in understanding performance. These findings underscore that adapting existing visual tokenizers for both understanding and generation is suboptimal due to the conflicting goals. This fundamental conflict hinders the development of a unified tokenizer [60, 66].

Table 1: Comparison of adapting different visual tokenizers for understanding and reconstruction.

Visual Tokenizer	Type	rFID ↓	Acc.↑
SigLIP2 [81]	Semantic	0.82	81.7
WeTok [109]	Pixel	0.64	18.3
WinTok	Hybrid	**0.41**	**82.0**

3.2 WinTok

Based on the above discussion, a natural question arises: *Is it feasible to build a unified tokenizer while decomposing visual understanding and generation?* To answer this question, we introduce WinTok, a hybrid tokenizer incorporated with transferable tokens to decompose visual understanding and generation and achieve a win-win performance.

Unified Encoder with Learnable Tokens. As depicted in Fig. 2, WinTok employs a unified encoder $\mathcal{E}$, structured as a Vision Transformer (ViT) [26]. The input image $\mathbf{X}$ is first patchified into N non-overlapping patches, which are subsequently converted into pixel tokens via a patch embedding layer $\mathcal{P}$:

$$\mathbf{P}^o = \mathcal{P}(\mathbf{X}) = \{\mathbf{P}^o_1, ..., \mathbf{P}^o_N\} \tag{1}$$

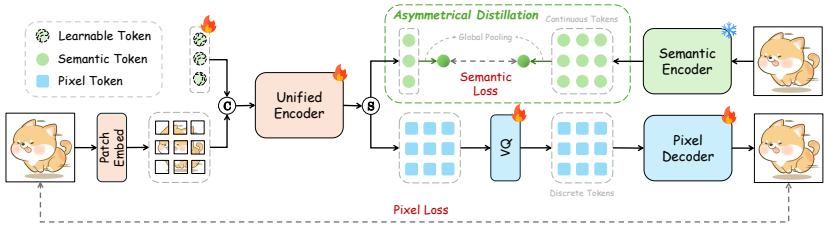


Fig. 2: Overview of our WinTok. Our WinTok adopts a hybrid tokenization paradigm integrated with learnable tokens. The separate token sets are optimized for visual understanding and generation respectively, with asymmetric token distillation for semantic tokens and image reconstruction for pixel tokens. Therefore, we mitigate the representation conflict while maintaining a unified tokenizer.

However, a singular token set proves inadequate for addressing both visual understanding and generation due to the inherent conflict between high-level semantic abstractions and low-level pixel reconstructions. To overcome this, we incorporate an additional set of M learnable tokens for task decomposition:

$$\mathbf{S}^o = \{\mathbf{S}^o_1, ..., \mathbf{S}^o_M\}. \tag{2}$$

More specifically, we leverage pixel tokens for visual generation, since these tokens contain rich local details from the input images. Alternatively, we leverage learnable tokens for visual understanding, since these tokens are the global vectors that can be optimized to summarize the global semantic of the input images. To make these two types of tokens work collaboratively, we concatenate them along the sequence dimension and input the combined tokens to our unified encoder: $\mathbf{P} \oplus \mathbf{S} = \mathcal{E}(\mathbf{P}^o \oplus \mathbf{S}^o)$. This arrangement enables the retention of distinct representations for each token type, fostering contextual integration for cooperative functionality. The subsequent sections detail the supervision methods for these two token sets.

Semantic Token: Asymmetric Distillation. Since the enhanced tokens $\mathbf{S}$ derive from the randomly-initialized learnable tokens $\mathbf{S}^o$, we optimize $\mathbf{S}$ to capture high-level semantics for visual understanding. Specifically, we propose an asymmetric token distillation technique where we use a visual foundation model $\mathcal{T}$ as the semantic teacher, and extract K semantic tokens of the input image:

$$\mathbf{T} = \mathcal{T}(\mathbf{X}) = \{\mathbf{T}_1, ..., \mathbf{T}_K\}. \tag{3}$$

Since the foundation model is well-pretrained with large-scale data, these tokens $\mathbf{T}$ contain discriminative semantic knowledge to understand the input image. As such, we propose to transfer such knowledge from semantic tokens $\mathbf{T}$ to learnable tokens $\mathbf{S}$. Asymmetry not only lies in the knowledge, but may also exists in the token numbers. Therefore, we adopt global pooling on both of them to obtain their corresponding global vectors: $\mathbf{s} = \text{Pool}(\mathbf{S}), \quad \mathbf{t} = \text{Pool}(\mathbf{T})$. This process is optimized using a cosine similarity loss:

$$\mathcal{L}_{sem} = 1 - \frac{\langle \mathbf{s}, \mathbf{t} \rangle}{||\mathbf{s}||_2 \cdot ||\mathbf{t}||_2}. \tag{4}$$

Table 2: Comparison of reconstruction quality and semantic capability on 256 × 256 ImageNet-1K validation set. "Capacity" denotes the theoretical combinations of code entries. * indicates the results are obtained by linear probing. WinTok achieves state-of-the-art classification accuracy while being competitive in reconstruction, with significantly less data compared to other unified tokenizers.

Method	Type	Ratio	Training Data	Codebook Size	Capacity	rFID (↓)	Accuracy (↑)
Semantic Tokenizer							
CLIP-L/14 [67]	Continuous	-	WIT400M	-	-	-	75.5
Dinov2-L [63]	Continuous	-	LVD142M	-	-	-	86.3*
SigLIP2-So/16 [81]	Continuous	-	WebLI10B	-	-	-	83.4
Reconstruction-oriented Tokenizer							
SD-VAE 1.x [68]	Continuous	8	OI1B	-	-	1.22	-
SD-VAE 2.x [68]	Continuous	8	Mix6B	-	-	0.70	-
SDXL-VAE [65]	Continuous	8	-	-	-	0.67	-
SD-VAE 3.5 [28]	Continuous	8	-	-	-	0.19	-
FLUX-VAE [44]	Continuous	8	-	-	-	0.18	-
VA-VAE [98]	Continuous	16	IN-1K	-	-	0.28	-
RAE (SigLIP2) [107]	Continuous	16	IN-1K	-	-	0.53	79.1*
VFM-VAE [8]	Continuous	16	IN-1K	-	-	0.52	-
LlamaGen [74]	Discrete	16	IN-1K	16384	2^{14}	2.19	-
Open-Magvit2 [59]	Discrete	16	IN-1K	262144	2^{18}	1.17	-
VFMTok [106]	Discrete	-	IN-1K	16384	2^{14}	0.89	69.4*
WeTok [109]	Discrete	16	IN-1K	-	2^{32}	0.61	-
MGVQ [39]	Discrete	16	IN-1K	2048 × 8	2^{88}	0.49	-
Unified Tokenizer							
UniLIP [76]	Continuous	16	BP-32M	-	-	0.79	-
UniFlow (SigLIP2) [102]	Continuous	16	IN-1K	-	-	0.62	-
VILA-U [91]	Discrete	16	CY700M	16384	2^{14}	1.80	73.3
QLIP-L [105]	Discrete	16	DC1B	-	-	1.46	79.1
DualToken [73]	Discrete	16	CC12M	-	-	0.54	81.6
TokenFlow [66]	Discrete	16	LA+CY700M	32768	2^{15}	1.37	-
SemHiTok [22]	Discrete	16	Mix70M	16384 × 12	2^{14}	1.16	-
TokLIP-L [53]	Discrete	16	Mix80M	16384	2^{14}	2.19	80.0
UniTok [60]	Discrete	16	DC1B	4096 × 8	2^{96}	0.41	70.8
VQRAE [27]	Hybrid	16	BP-32M	16384	2^{14}	1.31	-
WinTok	Hybrid	16	Mix50M	4096 × 4	2^{48}	0.41	82.0

Pixel Tokens: Image Reconstruction. The enhanced tokens $\mathbf{P}$ emerge from the pixel tokens $\mathbf{P}^o$ and are utilized for visual generation. Specifically, we employ a vector-quantization module $\mathcal{Q}$ to convert $\mathbf{P}$ into discrete pixel tokens via a learnable codebook:

$$\mathbf{Q} = \mathcal{Q}(\mathbf{P}) = \{\mathbf{Q}_1, ..., \mathbf{Q}_N\}. \tag{5}$$

These quantized tokens then serve as the input to a decoder $\mathcal{D}$ to reconstruct the original image:

$$\hat{\mathbf{X}} = \mathcal{D}(\mathbf{Q}). \tag{6}$$

We optimize the quantizer and the decoder using a combination of pixel reconstruction and codebook losses [74, 99]:

$$\mathcal{L}_{rec} = ||\mathbf{X} - \hat{\mathbf{X}}||_2 + ||\mathrm{g}[\mathbf{P}] - \mathbf{Q}||_2 + \beta||\mathbf{P} - \mathrm{g}[\mathbf{Q}]||_2, \tag{7}$$

where $\mathrm{g}[\cdot]$ denotes the stop-gradient operation and β is a hyperparameter to balance the loss terms. Additionally, we incorporate a perceptual loss $\mathcal{L}_{per}$ [104] and an adversarial loss $\mathcal{L}_{adv}$ [41] to enhance the visual quality of reconstructed im-

Table 3: Comparison of visual reconstruction on ImageNet-1K and MS-COCO 2017 validation set. Images are resized to 256 × 256 for evaluation. Our WinTok achieves competitive performance compared to UniTok even with significantly less training data and a reduced codebook size.

Method	Ratio	MS-COCO 2017			Imagenet-1K		
		rFID↓	PSNR↑	SSIM↑	rFID↓	PSNR↑	SSIM↑
Show-o [93]	16	9.26	20.90	0.59	3.50	21.34	0.59
Open-MAGVIT2-I-PT [59]	16	7.93	22.21	0.62	2.55	22.21	0.62
LlamaGen [74]	16	8.40	20.28	0.55	2.47	20.65	0.54
WeTok [109]	16	6.55	21.99	0.63	1.58	22.38	0.62
BSQ [97]	16	-	-	-	3.81	24.12	0.66
QLIP-B [105]	16	-	-	-	3.21	23.16	0.63
TokenFlow [66]	16	-	-	-	1.37	21.41	0.69
UniTok [60]	16	4.05	23.92	0.73	0.41	24.31	0.71
WinTok	16	4.24	23.83	0.72	0.41	24.23	0.71

ages. Thus, the optimization of pixel tokens is based on $\mathcal{L}_{pix} = \mathcal{L}_{rec} + \lambda_{per}\mathcal{L}_{per} + \lambda_{adv}\mathcal{L}_{adv}$, where λ_{per} and λ_{adv} are hyperparameters.

Training Objectives. The overall loss function for optimizing WinTok combines the losses from both token types:

$$\mathcal{L} = \mathcal{L}_{sem} + \mathcal{L}_{pix}, \tag{8}$$

Through asymmetric token distillation, the learnable tokens evolve to encapsulate the semantic strengths of the foundation model for visual understanding, while pixel tokens learn to capture local details for visual generation. Consequently, our WinTok achieves a win-win performance for both tasks with a hybrid tokenization framework.

WinTok for Downstream Tasks. WinTok's hybrid structure allows for the effective application of semantic and pixel tokens in understanding and generation tasks, respectively. Therefore, we further integrate WinTok into a pre-trained LLM to excavate its potential in downstream tasks. For multimodal understanding, the unified encoder $\mathcal{E}$ extracts continuous semantic tokens $\mathbf{S}$ from the input image. Then these continuous visual tokens are projected into the textual embedding space with a learnable linear layer, and integrated with text tokens for comprehension. For visual generation, discrete pixel tokens $\mathbf{Q}$ are extracted using the unified encoder and quantizer. Subsequently, the LLM learns the joint distribution between these discrete visual tokens given text tokens as condition. We also introduce an autoregressive head to facilitate multi-code prediction as in previous work [46,60,91]. During inference, the LLM autoregressively samples discrete visual tokens, which are then decoded into images with the decoder $\mathcal{D}$.

4 Experiment

4.1 Implementation Details

Tokenizer Setup. In our experiments, we adopt ViT-based architectures for both the encoder and the decoder. The encoder is initialized with SigLIP2-

Table 4: Evaluation on multimodal understanding benchmarks. WinTok achieves superior performance compared to other unified tokenizers integrated with LLMs or UMMs. For instance, we outperform TokenFlow-L by 14.6% on MMBench and UniTok by 3.3% on POPE. WinTok† denotes using only 64 semantic tokens.

Method	LLM	Token Type	Res.	POPE	GQA	TextVQA	MME-P	MMBench	MM-Vet
Understanding Only MLLM									
InstructBLIP [23]	Vicuna-7B	2D-Continuous	224	-	49.2	50.7	-	-	26.2
ShareGPT4V [18]	Vicuna-7B	2D-Continuous	336	-	63.3	60.4	1567.4	68.8	37.6
LLaVA-v1.5 [56]	Vicuna-7B	2D-Continuous	336	85.9	62.0	58.2	1510.7	64.3	30.5
Qwen2.5-VL [4]	Qwen2.5-7B	2D-Continuous	dynamic	-	-	84.9	-	83.5	67.1
InternVL2.5 [20]	InternLM2.5-7B	2D-Continuous	dynamic	90.6	-	79.1	-	84.6	62.8
LLaVA-OneVision [47]	Qwen2-7B	2D-Continuous	384	-	-	-	1580.0	80.8	57.5
Unified Multimodal Model									
LWM [55]	LLaMA2-7B	2D-Discrete	256	75.2	44.8	18.8	-	-	9.6
Show-o [93]	Phi-1.5-1.3B	2D-Discrete	256	80.0	58.0	-	1097.2	-	-
Liquid [89]	Gemma-7B	2D-Discrete	512	81.1	58.4	42.4	1119.0	-	-
Emu3 [87]	8B (from scratch)	2D-Discrete	512	85.2	60.3	64.7	1243.8	58.5	37.2
Janus-Pro [19]	DeepSeek-LLM-7B	2D-Continuous	384	87.4	62.0	-	1567.1	79.2	50.0
LaVIT [40]	LLaMA-7B	2D-Continuous	224	-	46.8	-	-	58.0	-
SEED-X [32]	LLaMA2-13B	2D-Continuous	448	84.1	49.1	-	1457.0	70.1	43.0
ILLUME [84]	Vicuna-7B	2D-Continuous	224	88.5	-	72.1	1445.3	65.1	37.0
VARGPT [110]	Vicuna-7B	2D-Continuous	256	85.9	-	-	1488.8	67.6	-
BLIP3-o [15]	Qwen2.5VL-7B-Instruct	2D-Continuous	dynamic	-	-	83.1	1682.6	83.5	66.6
Unified Tokenizer w/ LLM									
VILA-U [91]	LLaMA2-7B	2D-Discrete	256	83.9	58.3	48.3	1336.2	-	27.7
UniTok [60]	LLaMA2-7B	2D-Discrete	256	83.2	61.1	51.6	1448.0	-	33.9
MUSE-VL [94]	Qwen2.5-7B	2D-Discrete	256	-	-	-	-	72.1	-
TokLIP [53]	Qwen2.5-7B-Instruct	1D-Continuous	384	84.9	57.0	-	1496.6	76.9	-
TokenFlow-L [66]	Vicuna-13B	2D-Discrete	256	85.0	60.3	54.1	1365.4	60.3	27.7
SemHiTok [22]	Qwen2.5-7B-Instruct	2D-Discrete	256	83.4	60.3	-	1449.0	72.3	30.5
VQRAE [27]	Vicuna-13B	2D-Continuous	256	85.1	63.4	46.5	1491.1	65.5	-
WinTok†	Qwen3-8B	1D-Continuous	256	84.1	58.5	47.1	1370.4	60.2	25.2
WinTok	Qwen3-8B	1D-Continuous	256	86.5	62.4	55.2	1552.0	74.9	34.6

So400M [81], while the decoder is trained from scratch. The number of learnable tokens is set to 256 by default. Since the quantization operation is non-differentiable, we employ the straight-through estimator [6] to for proper gradient backpropagation. We adopt Multi-codebook Quantization (MCQ) [60] and select SigLIP2-So400M [81] as the default semantic teacher. We use 50M images randomly sampled from open-source datasets [10,13,25,31,70,83,86] to train our WinTok. The tokenizer is trained for 5 epochs with global batch size of 256 and learning rate of 2e-4 with warm-up and cosine decay schedule. All experiments are conducted on H20 GPUs with PyTorch. More details are provided in the supplementary material.

UMM Setup. We initialize our unified multimodal model with Qwen3-8B [96]. For pre-training stage, we incorporate approximately 80M image-text pairs from [15, 31, 69, 72]. We further finetune the model using 6M instruction-following data from [47, 56] for multimodal understanding, along with 4M synthetic data generated by FLUX.2-klein [45] and Z-Image-Turbo [11] for visual generation.

Evaluation Metrics. We evaluate WinTok on ImageNet-1K [25] validation set using rFID and Top-1 classification accuracy for state-of-the-art comparison. For visual reconstruction, we further report rFID, PSNR, and SSIM on ImageNet-1K and MS-COCO 2017 [54] validation set. For multimodal understanding, we evaluate on a wide range of benchmarks, including POPE [51], GQA [37], TextVQA [71], MME-P [30], MMBench [58], and MM-Vet [101]. For visual generation, we evaluate on GenEval [33] and DPG-Bench [36].

Table 5: Comparison of visual generation on GenEval and DPG-Bench.

Method	# Params	GenEval				DPG-Bench			
		Two Obj.	Counting	Color Attri.	Overall↑	Entity	Attribute	Relation	Overall↑
Diffusion-based Model									
SDv1.5 [68]	0.9B	0.38	0.35	0.06	0.43	74.23	75.39	73.49	63.18
SDXL [65]	2.6B	0.74	0.39	0.23	0.55	82.43	80.91	86.76	74.65
PixArt-α [17]	0.6B	0.50	0.44	0.07	0.48	79.32	78.60	82.57	71.11
PixArt-Σ [16]	0.6B	-	-	-	-	82.89	88.94	86.59	80.54
Hunyuan DiT [52]	1.5B	-	-	-	-	80.59	88.01	74.36	78.87
DALLE3 [7]	-	0.87	0.47	0.45	0.67	89.61	88.39	90.58	83.50
SD3-Medium [28]	2B	0.94	0.72	0.60	0.74	91.01	88.83	80.70	84.08
SANA-1.5 [92]	4.8B	0.93	0.86	0.65	0.81	-	-	-	84.70
Autoregressive-based Model									
Chameleon [77]	7B	-	-	-	0.39	-	-	-	-
LlamaGen [74]	0.8B	0.34	0.21	0.04	0.32	75.43	76.17	84.76	64.84
Janus [88]	1.3B	0.68	0.30	0.42	0.61	87.38	87.70	85.46	79.68
Harmon [90]	1.5B	0.86	0.66	0.48	0.76	-	-	-	-
Transfusion [108]	7B	-	-	-	0.63	-	-	-	-
UniTok [60]	7B	0.73	0.38	0.41	0.59	88.27	88.05	88.45	81.18
TokenFlow [66]	13B	0.66	0.40	0.26	0.55	79.22	81.29	85.22	73.38
WinTok	8B	0.88	0.75	0.55	0.76	89.51	88.27	89.78	83.36

4.2 Comparison with State-of-The-Art Methods

Tokenizer. As summarized in Tab. 2, our WinTok demonstrates superior performance in both reconstruction quality and classification accuracy compared to leading visual tokenizers. In terms of reconstruction quality, our WinTok surpasses all previous discrete reconstruction-oriented tokenizers as well as most unified tokenizers. Notably, WinTok achieves a comparable rFID to the state-of-the-art UniTok [60], employing considerably less training data and a reduced codebook size. Moreover, WinTok achieves 100% codebook usage even with a high capacity. Regarding semantic representation capabilities, WinTok outperforms all prior unified tokenizers, exceeding the specialized semantic tokenizer CLIP-L/14 [67] by 6.5%. These findings highlight the efficacy of our hybrid tokenization design, which incorporates transferable tokens and asymmetric token distillation to effectively decompose understanding and generation tasks.

Visual Reconstruction. As shown in Tab. 3, our WinTok achieves impressive reconstruction quality on 256 $\times$ 256 ImageNet-1K and MS-COCO 2017 validation set. Notably, WinTok is competitive with the state-of-the-art unified tokenizer UniTok while trained on significantly less data (50M vs. 1B) and a reduced codebook capacity (2^{48} vs. 2^{96}).

Multimodal Understanding. As presented in Tab. 4, when integrated with an LLM, WinTok achieves competitive performance across various multimodal understanding benchmarks. Our model outperforms several pioneering UMMs, including SEED-X [32] and Liquid [89] on POPE and GQA. Moreover, we achieve 86.5% on POPE and 55.2 % on TextVQA, surpassing UniTok [60] by 3.3% and 3.6%, respectively. Even with only 64 semantic tokens representing the image, WinTok can still obtain satisfactory comprehension results, outperforming VILA-U [91] on MME-P and SemHiTok [22] on POPE. Overall, WinTok shows consistent improvements across all benchmarks, indicating its potential for multimodal understanding tasks.

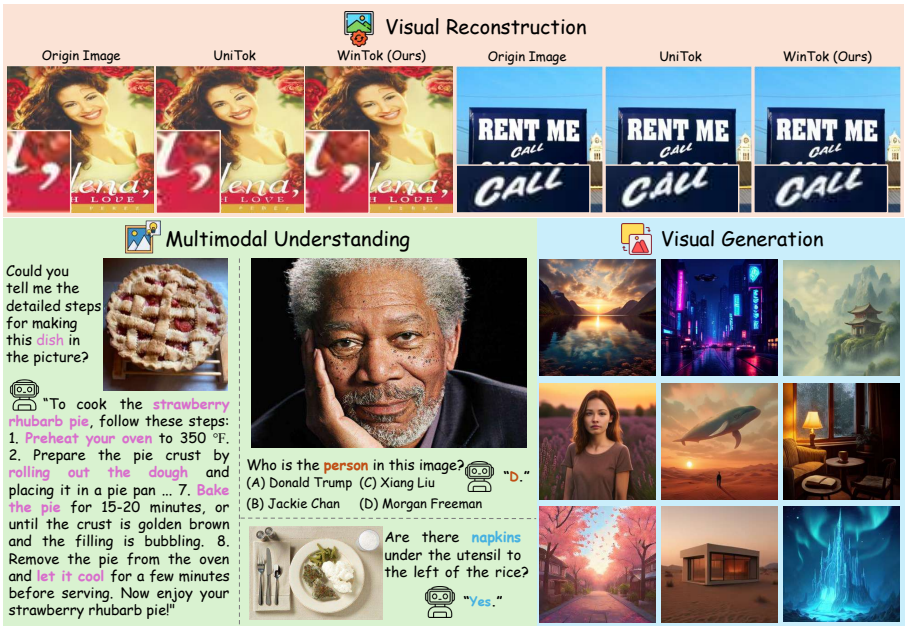
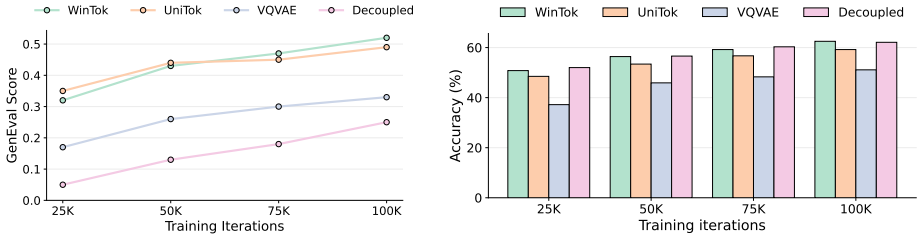


Fig. 3: Qualitative results demonstrating the superior performance of our WinTok on downstream applications.

Visual Generation. As summarized in Tab. 5, our multimodal model consistently achieves competitive or even superior performance compared to state-of-the-art diffusion and autoregressive-based models. Notably, our unified multimodal model outperforms representative autoregressive systems such as Chameleon [77], LlamaGen [74], and Janus [88], while remaining competitive with large-scale diffusion experts trained on billions of image–text pairs. Furthermore, compared with recent approaches that incorporate unified tokenizers into large language models, *i.e.*, UniTok [60] and TokenFlow [66], WinTok consistently achieves better performance across both benchmarks. These results highlight the robustness and effectiveness of WinTok as the visual tokenizer within a unified multimodal framework, particularly for complex and compositional text-to-image generation tasks. Moreover, the above observations further confirm that hybrid tokenization substantially benefits downstream unified modeling.

Visualization Fig. 3 presents qualitative results of WinTok on visual reconstruction, multimodal understanding, and visual generation. Our WinTok not only preserves fine-grained details for high-quality reconstruction, but also effectively captures global semantic information for accurate multimodal understanding. Moreover, it enables the generative model to produce diverse and realistic images. These results substantiate the efficacy of our approach in balancing the requirements of both understanding and generation tasks via decomposition with transferable tokens.



(a) Text-to-image generation performance. **(b)** Multimodal understanding performance.

Fig. 4: Comparisons of using different tokenization strategies. (a) Generation performance is evaluated on GenEval [33]. (b) Understanding performance is averaged across 4 benchmarks [37, 51, 58, 71].

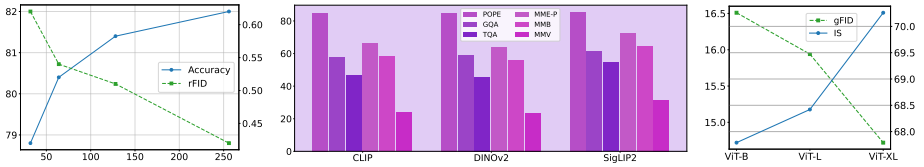
4.3 Comparison with Other Tokenization Strategies

In this section, we compare our hybrid tokenization approach, WinTok, with several representative visual tokenization strategies for unified visual understanding and generation. Specifically, we consider the following baselines: (a) **VQVAE** [82]: a pixel-level discrete representation widely used in autoregressive image generation. We adopt the pretrained VQVAE tokenizer from LlamaGen [74], which converts images into discrete visual tokens for multimodal modeling. (b) **Decoupled tokenization**: a dual-encoder design that employs separate visual representations for understanding and generation, using a semantic encoder (SigLIP2 [81]) for perception and a VQVAE tokenizer for image synthesis. This strategy resembles previous methods [24, 88] but has a slightly different implementation. (c) **Unified tokenizer**: this line of work employs a unified representation that jointly models semantic and pixel-level information. We select UniTok [60] as a representative.

To ensure a fair comparison, we train unified multimodal models with these tokenizers under the same training configuration and data scale. Specifically, we construct a controlled training subset consisting of 10M text-to-image and image-to-text data. All tokenizers are integrated with Qwen3-4B [96]. All models are evaluated on both visual understanding benchmarks [37, 51, 58, 71] and text-to-image generation tasks [33] to comprehensively assess their performance in unified multimodal modeling.

As illustrated in Fig. 4, our approach demonstrates strong performance on downstream text-to-image (T2I) generation and multimodal understanding tasks. In T2I generation, our method converges faster and achieves the best results. Compared to the single-stream unified tokenizer UniTok, our method obtains a greater upper bound. In contrast, the VQVAE-based model, which relies solely on pixel-level representation, underperforms in generation. Moreover, the decoupled strategy performs worst in generation, likely due to the inconsistency in representation space and need large-scale training to achieve better performance.

For multimodal understanding, both WinTok and the decoupled strategy achieve satisfactory results owing to the continuous semantic representation. The discrete unified tokenizer UniTok, despite jointly optimizing pixel reconstruction



(a) Effects of learnable token numbers. **(b)** Comparison on using different semantic teachers. **(c)** Effects of different decoder sizes.

Fig. 5: Ablations on design choices.(a) As the learnable token number increases, both the reconstruction quality and semantic capability improve. (b) Our WinTok demonstrates similar trends when adopting different semantic teachers, while SigLIP2 [81] benefits the most. (c) A larger decoder not only enhances reconstruction quality but also improves generation performance. Implementation details and more comprehensive analyses are provided in the supplementary material.

and semantic alignment losses, performs similarly to VQVAE, suggesting that its representation remains biased toward pixel-level information and fails to fully exploit semantic cues. Overall, our method provides the best trade-off between visual generation and understanding, demonstrating strong and balanced performance across both tasks.

4.4 Ablations

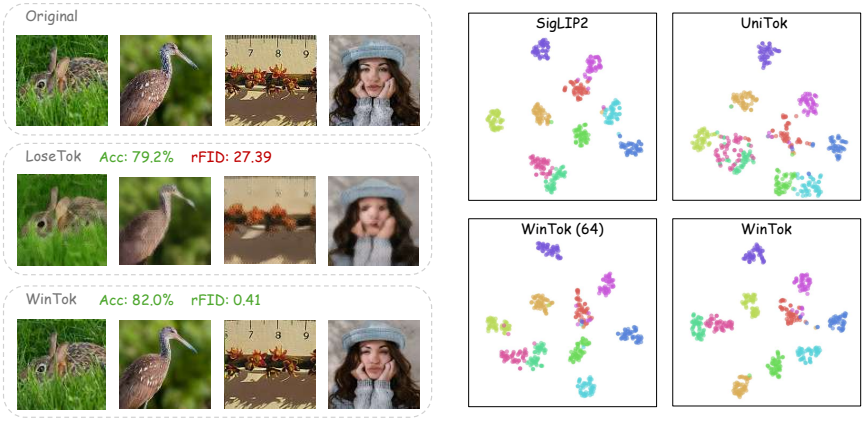
Number of Learnable Tokens. We investigate the impact of learnable token quantity on task performance. As shown in Fig. 5a, increasing the number of learnable tokens consistently improves both reconstruction quality and semantic capability. This improvement arises from the enhanced ability of additional tokens to capture semantic information and facilitate the disentanglement of representations, thereby mitigating conflicts between the two tasks.

Semantic Teacher. To optimize the transfer of semantic knowledge to the randomly initialized semantic tokens, we assess the influence of the chosen semantic teacher model. We examine three representative visual foundation models: CLIP-L/14 [67], DINOv2-L [63], and SigLIP2-So400M [81]. Our results indicate consistent improvements in downstream multimodal understanding tasks across different semantic teachers, with SigLIP2 yielding the best performance due to its superior representation capabilities derived from meticulous pre-training.

Decoder Size. We explore the effects of varying decoder sizes on both reconstruction quality and generation performance by training several WinTok variants. All variants achieved reasonable reconstruction quality, with ViT-B attaining an rFID of 0.60, ViT-L achieving 0.55, and the best-performing ViT-XL achieving 0.54. As illustrated in Fig. 5c, larger decoders not only improve reconstruction quality but also enhance generation performance, aligning with findings in prior studies [2, 95, 107].

4.5 Discussions

What if we design in the other way? To validate our hybrid strategy, we evaluate a reversed variant termed *LoseTok*, where learnable tokens handle re-



(a) Alternative design consideration. **(b)** Clusters of different tokenizers.

Fig. 6: Discussions. (a) Reversing the roles of the two token types results in significant performance degradation, and WinTok is more rationale. (b) WinTok produces more discriminative clusters compared to other tokenizers, even with only 64 tokens representing the global semantics.

construction and pooled pixel tokens manage asymmetric token distillation. As shown in Fig. 6a, this swap severely degrades reconstruction quality while barely affecting semantic capability. This drop in generative performance occurs because learnable tokens lack the capacity to preserve fine-grained spatial details. While this finding appears to contrast with some prior works [2,49], the discrepancy arises from the additional semantic supervision imposed on pixel tokens in our setting. Ultimately, this ablation confirms the WinTok design, highlighting the necessity of strictly disentangling semantic abstraction from pixel-level reconstruction.

Do the transferable tokens truly learn? To evaluate the effectiveness of learnable tokens in capturing transferable semantic representations, we visualize the t-SNE [62] embeddings of the learned semantic tokens alongside the pooled image features produced by pre-trained tokenizers, including SigLIP2 [81] and UniTok [60]. As shown in Fig. 6b, the semantic tokens generated by WinTok form more compact and well-separated clusters than those obtained from the other two tokenizers, despite using only 64 tokens to encode global semantics. This result highlights the superiority of our asymmetric token distillation strategy in effectively transferring semantic knowledge.

5 Conclusion

This work presents WinTok, a hybrid tokenizer that balances visual understanding and generation by using learnable tokens for global semantic distillation and pixel tokens for local detail reconstruction. Through such hybrid encoding, WinTok can provide effective and flexible representations for downstream unified

modeling. While experiments demonstrate its superior performance as a versatile foundation for unified multimodal models, its current generalization is constrained by a modest 50M-sample training dataset and a lack of downstream architectural exploration beyond Qwen3-8B. Future research will focus on scaling the training corpus to billion-scale levels and co-designing novel unified architectures to fully leverage WinTok's unique hybrid representations.

Acknowledgements

This work was supported by Guangdong Science and Technology Program (Grant No. 2024TQ08X365).

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems **35**, 23716–23736 (2022) 4
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. In: Forty-second International Conference on Machine Learning (2025) 4, 13, 14
3. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023) 1
4. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025) 9
5. Bai, Z., Gao, J., Gao, Z., Wang, P., Zhang, Z., He, T., Shou, M.Z.: Factorized visual tokenization and generation. arXiv preprint arXiv:2411.16681 (2024) 4
6. Bengio, Y., Léonard, N., Courville, A.: Estimating or propagating gradients through stochastic neurons for conditional computation. arXiv preprint arXiv:1308.3432 (2013) 9
7. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al.: Improving image generation with better captions. Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf **2**(3), 8 (2023) 10
8. Bi, T., Zhang, X., Lu, Y., Zheng, N.: Vision foundation models can be good tokenizers for latent diffusion models. arXiv preprint arXiv:2510.18457 (2025) 4, 7
9. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020) 1
10. 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) 9

11. Cai, H., Cao, S., Du, R., Gao, P., Hoi, S., Hou, Z., Huang, S., Jiang, D., Jin, X., Li, L., et al.: Z-image: An efficient image generation foundation model with single-stream diffusion transformer. arXiv preprint arXiv:2511.22699 (2025) 9
12. 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) 3
13. 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) 9
14. Chen, B., Bi, S., Tan, H., Zhang, H., Zhang, T., Li, Z., Xiong, Y., Zhang, J., Zhang, K.: Aligning visual foundation encoders to tokenizers for diffusion models. arXiv preprint arXiv:2509.25162 (2025) 4
15. 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
16. Chen, J., Ge, C., Xie, E., Wu, Y., Yao, L., Ren, X., Wang, Z., Luo, P., Lu, H., Li, Z.: Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In: European Conference on Computer Vision. pp. 74–91. Springer (2024) 10
17. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., et al.: Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. arXiv preprint arXiv:2310.00426 (2023) 10
18. 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) 9
19. Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C.: Janus-pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811 (2025) 2, 4, 9
20. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024) 9
21. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24185–24198 (2024) 3
22. 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) 4, 7, 9, 10
23. Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P.N., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. Advances in neural information processing systems **36**, 49250–49267 (2023) 9
24. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., et al.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025) 4, 12
25. 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) 3, 5, 9

26. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: The Ninth International Conference on Learning Representations (2021) 4, 5
27. Du, S., Guo, J., Li, B., Cui, S., Xu, Z., Luo, Y., Wei, Y., Gai, K., Wang, X., Wu, K., et al.: Vqrae: Representation quantization autoencoders for multimodal understanding, generation and reconstruction. arXiv preprint arXiv:2511.23386 (2025) 4, 7, 9
28. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024) 7, 10
29. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021) 2, 3, 5
30. Fu, C., Zhang, Y.F., Yin, S., Li, B., Fang, X., Zhao, S., Duan, H., Sun, X., Liu, Z., Wang, L., et al.: Mme-survey: A comprehensive survey on evaluation of multimodal llms. arXiv preprint arXiv:2411.15296 (2024) 9
31. 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) 9
32. Ge, Y., Zhao, S., Zhu, J., Ge, Y., Yi, K., Song, L., Li, C., Ding, X., Shan, Y.: Seed-x: Multimodal models with unified multi-granularity comprehension and generation. arXiv preprint arXiv:2404.14396 (2024) 9, 10
33. Ghosh, D., Hajishirzi, H., Schmidt, L.: Geneval: An object-focused framework for evaluating text-to-image alignment. Advances in Neural Information Processing Systems **36**, 52132–52152 (2023) 9, 12
34. Gui, J., Chen, T., Zhang, J., Cao, Q., Sun, Z., Luo, H., Tao, D.: A survey on self-supervised learning: Algorithms, applications, and future trends. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(12), 9052–9071 (2024) 4
35. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022) 4
36. 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) 9
37. 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) 9, 12
38. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024) 1
39. Jia, M., Yin, W., Hu, X., Guo, J., Guo, X., Zhang, Q., Long, X.X., Tan, P.: Mgvq: Could vq-vae beat vae? a generalizable tokenizer with multi-group quantization. arXiv preprint arXiv:2507.07997 (2025) 4, 7
40. Jin, Y., Xu, K., Chen, L., Liao, C., Tan, J., Huang, Q., CHEN, B., Song, C., ZHANG, D., Ou, W., et al.: Unified language-vision pretraining in llm with dynamic discrete visual tokenization. In: The Twelfth International Conference on Learning Representations (2024) 9

41. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019) 7
42. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013) 3
43. Kingma, D.P., Welling, M., et al.: An introduction to variational autoencoders. Foundations and Trends® in Machine Learning **12**(4), 307–392 (2019) 3
44. Labs, B.F.: Flux. `https://github.com/black-forest-labs/flux` (2024) 7
45. Labs, B.F.: FLUX.2: Frontier Visual Intelligence. `https://bfl.ai/blog/flux-2` (2025) 9
46. 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) 8
47. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024) 9
48. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023) 3
49. Li, X., Qiu, K., Chen, H., Kuen, J., Gu, J., Raj, B., Lin, Z.: Imagefolder: Autoregressive image generation with folded tokens. arXiv preprint arXiv:2410.01756 (2024) 4, 14
50. Li, Y., Qian, R., Pan, B., Zhang, H., Huang, H., Zhang, B., Tong, J., You, H., Du, X., Gan, Z., et al.: Manzano: A simple and scalable unified multimodal model with a hybrid vision tokenizer. arXiv preprint arXiv:2509.16197 (2025) 4
51. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 292–305 (2023) 3, 9, 12
52. Li, Z., Zhang, J., Lin, Q., Xiong, J., Long, Y., Deng, X., Zhang, Y., Liu, X., Huang, M., Xiao, Z., et al.: Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. arXiv preprint arXiv:2405.08748 (2024) 10
53. 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) 2, 4, 7, 9
54. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014) 9
55. 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) 9
56. 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) 2, 3, 4, 9
57. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023) 2, 3, 4
58. 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 (2024) 9, 12

59. Luo, Z., Shi, F., Ge, Y., Yang, Y., Wang, L., Shan, Y.: Open-magvit2: An open-source project toward democratizing auto-regressive visual generation. arXiv preprint arXiv:2409.04410 (2024) 7, 8
60. 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) 2, 3, 4, 5, 7, 8, 9, 10, 11, 12, 14
61. Ma, N., Goldstein, M., Albergo, M.S., Boffi, N.M., Vanden-Eijnden, E., Xie, S.: Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In: European Conference on Computer Vision. pp. 23–40. Springer (2024) 3
62. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008) 14
63. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2023) 4, 7, 13
64. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023) 3
65. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. In: The Twelfth International Conference on Learning Representations (2024) 7, 10
66. Qu, L., Zhang, H., Liu, Y., Wang, X., Jiang, Y., Gao, Y., Ye, H., Du, D.K., Yuan, Z., Wu, X.: Tokenflow: Unified image tokenizer for multimodal understanding and generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2545–2555 (2025) 2, 4, 5, 7, 8, 9, 10, 11
67. 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) 2, 3, 4, 7, 10, 13
68. 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) 3, 7, 10
69. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in neural information processing systems **35**, 25278–25294 (2022) 9
70. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2556–2565 (2018) 9
71. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8317–8326 (2019) 9, 12
72. Singla, V., Yue, K., Paul, S., Shirkavand, R., Jayawardhana, M., Ganjdanesh, A., Huang, H., Bhatele, A., Somepalli, G., Goldstein, T.: From pixels to prose: A large dataset of dense image captions. arXiv preprint arXiv:2406.10328 (2024) 9

73. Song, W., Wang, Y., Song, Z., Li, Y., Sun, H., Chen, W., Zhou, Z., Xu, J., Wang, J., Yu, K.: Dualtoken: Towards unifying visual understanding and generation with dual visual vocabularies. arXiv preprint arXiv:2503.14324 (2025) 2, 3, 4, 7
74. 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) 2, 3, 5, 7, 8, 10, 11, 12
75. Sun, Q., Yu, Q., Cui, Y., Zhang, F., Zhang, X., Wang, Y., Gao, H., Liu, J., Huang, T., Wang, X.: Emu: Generative pretraining in multimodality. In: The Twelfth International Conference on Learning Representations (2024) 1
76. Tang, H., Xie, C., Bao, X., Weng, T., Li, P., Zheng, Y., Wang, L.: Unilip: Adapting clip for unified multimodal understanding, generation and editing. arXiv preprint arXiv:2507.23278 (2025) 2, 3, 4, 7
77. Team, C.: Chameleon: Mixed-modal early-fusion foundation models. arXiv preprint arXiv:2405.09818 (2024) 1, 10, 11
78. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023) 1
79. 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) 3
80. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023) 1
81. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025) 2, 3, 4, 5, 7, 9, 12, 13, 14
82. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017) 2, 3, 4, 5, 12
83. Wang, A.J., Mao, D., Zhang, J., Han, W., Dong, Z., Li, L., Lin, Y., Yang, Z., Qin, L., Zhang, F., et al.: Textatlas5m: A large-scale dataset for dense text image generation. arXiv preprint arXiv:2502.07870 (2025) 9
84. 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) 9
85. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024) 2, 3
86. Wang, S., Li, X., Li, J., Wang, G., Sun, X., Zhu, B., Qiu, H., Yu, M., Shen, S., Zhang, T., et al.: Faceid-6m: A large-scale, open-source faceid customization dataset. arXiv preprint arXiv:2503.07091 (2025) 9
87. 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) 9
88. 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) 2, 4, 10, 11, 12

89. Wu, J., Jiang, Y., Ma, C., Liu, Y., Zhao, H., Yuan, Z., Bai, S., Bai, X.: Liquid: Language models are scalable and unified multi-modal generators. arXiv preprint arXiv:2412.04332 (2024) 2, 9, 10
90. Wu, S., Zhang, W., Xu, L., Jin, S., Wu, Z., Tao, Q., Liu, W., Li, W., Loy, C.C.: Harmonizing visual representations for unified multimodal understanding and generation. arXiv preprint arXiv:2503.21979 (2025) 2, 4, 10
91. 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) 4, 7, 8, 9, 10
92. Xie, E., Chen, J., Zhao, Y., Yu, J., Zhu, L., Wu, C., Lin, Y., Zhang, Z., Li, M., Chen, J., et al.: Sana 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer. arXiv preprint arXiv:2501.18427 (2025) 10
93. Xie, J., Mao, W., Bai, Z., Zhang, D.J., Wang, W., Lin, K.Q., Gu, Y., Chen, Z., Yang, Z., Shou, M.Z.: Show-o: One single transformer to unify multimodal understanding and generation. In: The Thirteenth International Conference on Learning Representations (2025) 1, 8, 9
94. Xie, R., Du, C., Song, P., Liu, C.: Muse-vl: Modeling unified vlm through semantic discrete encoding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 24135–24146 (2025) 2, 4, 9
95. Xiong, T., Liew, J.H., Huang, Z., Feng, J., Liu, X.: Gigatok: Scaling visual tokenizers to 3 billion parameters for autoregressive image generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18770–18780 (2025) 13
96. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025) 9, 12
97. Yang, H., Duan, L., Chen, Y., Li, H.: Bsq: Exploring bit-level sparsity for mixed-precision neural network quantization. In: International Conference on Learning Representations (2021) 8
98. Yao, J., Yang, B., Wang, X.: Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15703–15712 (2025) 7
99. 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. In: International Conference on Learning Representations (2022) 2, 3, 4, 5, 7
100. Yu, Q., Weber, M., Deng, X., Shen, X., Cremers, D., Chen, L.C.: An image is worth 32 tokens for reconstruction and generation. Advances in Neural Information Processing Systems **37**, 128940–128966 (2024) 4
101. 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. In: International Conference on Machine Learning. pp. 57730–57754. PMLR (2024) 9
102. Yue, Z., Zhang, H., Zeng, X., Chen, B., Wang, C., Zhuang, S., Dong, L., Du, K., Wang, Y., Wang, L., et al.: Uniflow: A unified pixel flow tokenizer for visual understanding and generation. arXiv preprint arXiv:2510.10575 (2025) 2, 4, 7
103. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11975–11986 (2023) 2, 3, 4
104. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018) 7

105. 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) 2, 4, 7, 8
106. Zheng, A., Wen, X., Zhang, X., Ma, C., Wang, T., Yu, G., Zhang, X., Qi, X.: Vision foundation models as effective visual tokenizers for autoregressive image generation. arXiv preprint arXiv:2507.08441 (2025) 4, 7
107. Zheng, B., Ma, N., Tong, S., Xie, S.: Diffusion transformers with representation autoencoders. arXiv preprint arXiv:2510.11690 (2025) 4, 7, 13
108. Zhou, C., YU, L., Babu, A., Tirumala, K., Yasunaga, M., Shamis, L., Kahn, J., Ma, X., Zettlemoyer, L., Levy, O.: Transfusion: Predict the next token and diffuse images with one multi-modal model. In: The Thirteenth International Conference on Learning Representations (2025) 1, 10
109. Zhuang, S., Guo, Y., Fu, C., Huang, Z., Tian, Z., Wang, F., Zhang, Y., Li, C., Wang, Y.: Wetok: Powerful discrete tokenization for high-fidelity visual reconstruction. arXiv preprint arXiv:2508.05599 (2025) 5, 7, 8
110. Zhuang, X., Xie, Y., Deng, Y., Liang, L., Ru, J., Yin, Y., Zou, Y.: Vargpt: Unified understanding and generation in a visual autoregressive multimodal large language model. arXiv preprint arXiv:2501.12327 (2025) 2, 9