SemHiTok: A Unified Image Tokenizer via Semantic-Guided Hierarchical Codebook for Multimodal Understanding and Generation

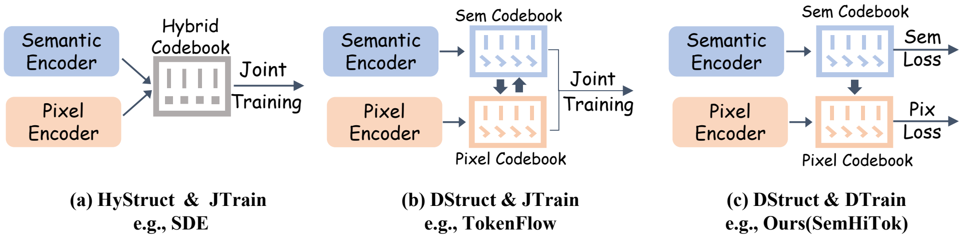
Anonymous authors
Paper under double-blind review

Abstract

In this paper, we introduce **SemHiTok**, a unified image **Tok**enizer via **Sem**antic-Guided **Hi**erarchical codebook that provides consistent discrete representations for multimodal understanding and generation. Recently, unified image tokenizers have sparked exploration within the research community, which is designed to capture high-level semantic features for understanding and retaining low-level pixel features for generation. Previous works attempt to train a unified image tokenizer by combining loss for semantic distillation and pixel reconstruction. However, due to the differing levels of features prioritized by multimodal understanding and generation, joint training methods face significant challenges in achieving a good trade-off. SemHiTok addresses this challenge through a novel semantic-guided hierarchical codebook, which builds pixel sub-codebooks on a pretrained semantic codebook. This design decouples the semantic and pixel in terms of structure and training strategy, enabling the tokenizer to capture pixel features while retaining its ability to comprehend high-level semantic information. Our experiments demonstrate that SemHiTok achieves SOTA performance in image reconstruction and multimodal understanding under the LLaVA-v1.5 setting. Further, we develop a unified MLLM with SemHiTok, which exhibits superior performance across multimodal understanding and generation tasks. For understanding, SemHiTok achieves impressive performance on most benchmarks. For generations, our model achieves SOTA performance on MJHQ30K in unified MLLMs. Our code and models will be open source.

1 Introduction

In recent years, autoregressive models have achieved great success in natural language processing and have been extended to the multimodal understanding domain, demonstrating immense potential. This triggers researchers' interest in unified multimodal understanding and generation by employing a single autoregressive framework. To achieve a unified multimodal large model, the key challenge is designing a tokenizer suitable for both multimodal generation and understanding tasks.



(a) HyStruct & JTrain
e.g., SDE

(b) DStruct & JTrain
e.g., TokenFlow

(c) DStruct & DTrain
e.g., Ours(SemHiTok)

Figure 1: Illustration of other tokenizers and SemHiTok. **HyStruct**: Using a single model to extract information at different levels; **DStruct**: Using different models to extract information at various levels; **JTrain**: Using a joint optimization training strategy; **DTrain**: Adopt a phased optimization training strategy.

However, there is a vast gap in the visual information required for these two tasks. For instance, models from the CLIP (Radford et al., 2021a; Sun et al., 2023; Zhai et al., 2023) family, commonly used in multimodal understanding tasks, tend to lose visual pixel information. On the contrary, the VQGAN (Yu et al., 2021a; Zhu et al., 2024a) family models, often used in autoregressive generation tasks, lack the ability to extract semantic features for multimodal understanding tasks. This leads to poor performance when a single tokenizer is applied to a unified MLLM (Wu et al., 2024b; Jin et al., 2023; Li et al., 2024b). In light of the aforementioned issues, some recent work has attempted to incorporate a semantic learning branch into the original VQGAN training pipeline, aiming to obtain a unified tokenizer via joint optimization. VILA-U (Wu et al., 2024c) employs a straightforward combination of semantic alignment loss and pixel reconstruction loss, which allows the model to capture both low-level and high-level information. Nevertheless, the reliance on a hybrid structure and joint optimization often drives the tokenizer toward a suboptimal solution. Furthermore, some recent works (Xie et al., 2025; Qu et al., 2024) have built upon it with further improvements. As shown in Fig.1(a), while SDE (Xie et al., 2025) further decouples the encoders, the remaining hybrid codebook continues to impede the model's optimization. TokenFlow (Qu et al., 2024) uses shared mapping to decouple the semantic branch and pixel branch while maintaining the consistency of the codebook index, but joint training still affects the final performance.

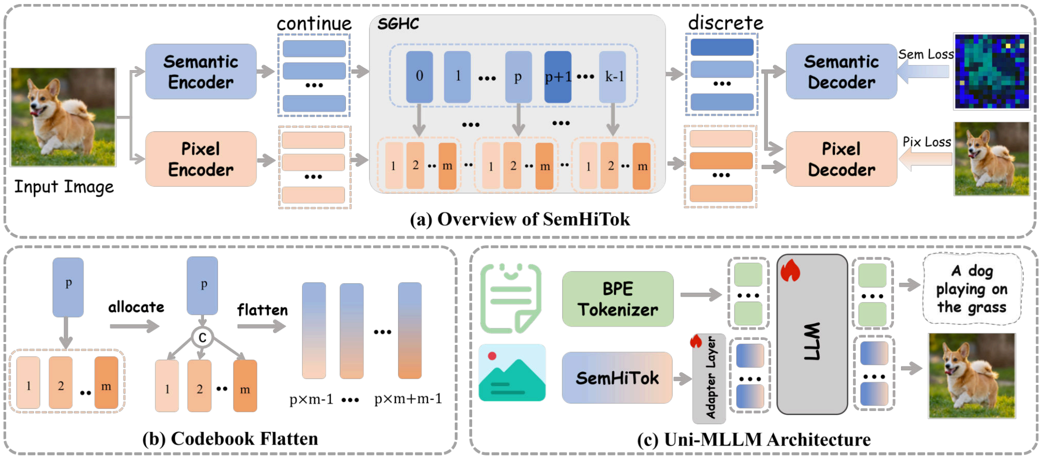
A straightforward approach is to use CLIP and VQGAN to extract semantic and pixel information, respectively, and the concatenation of these two token sequences is then used as a unified representation. Janus (Wu et al., 2024a) introduces a dual-encoder method that separates encoders for understanding and generation tasks to address this conflict, but this increases the complexity of handling mixed tasks and does not fundamentally resolve the feature conflict challenge. However, this leads to a doubling of the token sequence count or multiplicative expansion in vocabulary size, depending on whether the concatenation is applied along the length or dimension. These limitations underscore a fundamental challenge in the field: *How to balance semantic-level and pixel-level information effectively, without compromising the ease of integration into MLLM frameworks?*

To address this challenge, we propose **SemHiTok**, a unified image tokenizer that provides consistent feature representations for multimodal understanding and generation tasks through a unique hierarchical codebook design. Inspired by the observation that image patches with the same semantic code tend to have similar pixel features, we introduce a novelty hierarchical codebook which uses a sub-codebook to model the pixel-level space associated with each semantic code, named Semantic-Guided Hierarchical Codebook(SGHC). Unlike existing approaches, SemHiTok supports a stage-wise training paradigm where each stage exclusively optimizes specific hierarchy level features, allowing us to achieve a better trade-off between semantic and pixel feature extraction. In addition, SemHiTok can be seamlessly integrated into existing MLLMs following the next-token paradigm through a simple codebook flattening operation.

Our contributions can be summarized as follows: **(1)**: A novel unified tokenizer that achieves a trade-off between semantic and pixel information, demonstrating outstanding performance in both image reconstruction and multimodal understanding tasks. **(2)**: We develop a unified MLLM architecture that demonstrates superior performance across both multimodal understanding and generation tasks, validating its versatility. **(3)**: Our approach further pushes the performance boundary of unified discrete MLLMs, enabling improved scalability and representation capacity within next-token prediction frameworks.

2 Method

The main objective of **SemHiTok** is to establish a simple and unified image tokenizer for multimodal understanding and generation. In this model, the image is transformed into discrete tokens that contain semantic information and pixel information. We begin with a semantic codebook training recipe that reconstructs semantic features extracted from a language image pre-training model (Zhai et al., 2023; Wei et al., 2022), and point out the semantic codebook's poor texture feature representation in section 2.1. In section 2.2, we conduct a preliminary discussion and observation. Building on this observation, we introduce Semantic-Guided Hierarchical Codebook(SGHC), incorporating texture information while perfectly inheriting the semantic information of the semantic codebook, to enable pixel reconstruction enablement. Finally, we introduce the application of SemHiTok on unified MLLM in section 2.3. The overview framework is shown in Fig.2.



(a) Overview of SemHiTok

(b) Codebook Flatten

(c) Uni-MLLM Architecture

Figure 2: (a) SemHiTok is structurally composed of two branches: **semantic branch** and **pixel branch**. The **semantic branch** is trained following the VQKD (Wei et al., 2022), where the semantic codebook is learned through semantic loss. We propose semantic-guided hierarchical codebook(SGCH) composed of multiple pixel sub-codebooks, in which each pixel sub-codebook is in a one-to-one correspondence with a semantic code. The selection of pixel sub-codebook is indexed by the semantic code from semantic quantization. To enable a unified discrete representation, we concatenate the quantized semantic and pixel features along the channel dimension and feed the result into the pixel decoder for reconstruction. (b) Each semantic code is allocated to the corresponding pixel sub-codebook, and their features are concatenated along the dimension. (c) An illustration of the unified MLLM framework.

2.1 Semantic Codebook Training

For multimodal understanding, using a text-aligned visual encoder (Zhai et al., 2023; Radford et al., 2021a; Sun et al., 2023; Wang et al., 2024b) as an image tokenizer can accelerate convergence and improve performance. However, these text-aligned visual encoders typically output continuous semantic features. In this work, to achieve a unified visual tokenizer, we first train a semantic codebook to quantize the continuous semantic feature following VQKD (Wei et al., 2022).

Given an image $X^{H\times W\times 3}$, the semantic encoder $\mathcal{E}_{\text{sem}}$ extract continuous semantic features:

$$Z_{sem} = \mathcal{E}_{\text{sem}}(X) \in \mathbb{R}^{h\times w\times d_{sem}} \tag{1}$$

Where $\mathcal{E}_{\text{sem}}$ is a frozen text-aligned image encoder, *e.g.*, CLIP (Radford et al., 2021a) or SigLIP (Zhai et al., 2023). Then Z_{sem} are transformed into discrete feature space $\mathcal{C}_{sem} = \{c_1, c_2, ..., c_K\} \in \mathbb{R}^{K\times d_{sem}}$ through quantization function $\mathcal{Q}_{sem}(*)$. The quantization process $\mathcal{Q}_{sem}(*)$ is as follows:

$$Z_{q_{sem}}, I_{q_{sem}} = \underset{k\in\{1,...,K\}}{\arg\min} \; \|Z_{sem} - \mathcal{C}_{sem}[k]\| \tag{2}$$

Where $I_{q_{sem}} \in [\mathcal{C}_{sem}]^{h\times w}$ is quantized index, $Z_{q_{sem}}$ is discrete feature indexed from $\mathcal{C}_{sem}$. Finally, semantic decoder $\mathcal{D}_{sem}$ maps $Z_{q_{sem}}$ to raw semantic feature space $\hat{Z}_{sem}$. The $\mathcal{D}_{sem}$ are end-to-end trainable by minimizing semantic distill loss:

$$L_{sem} = 1 - \cos(Z_{q_{sem}}, \hat{Z}_{sem}) + |Z_{q_{sem}} - \hat{Z}_{sem}| \tag{3}$$

For $\mathcal{C}_{sem}$, we adopt EMA (Hunter, 1986) VQ as the semantic codebook. Unlike traditional quantization methods, the EMA VQ is not updated via gradient descent, but instead through an Exponential Moving Average (EMA) algorithm:

$$\mathbf{c}_k^{(t)} = m \cdot \mathbf{c}_k^{(t-1)} + (1-m) \cdot \frac{1}{N_k} \sum_{i=1}^{N_k} \mathbf{z}_i, \tag{4}$$

where $\mathbf{c}_k^t$ denotes the k-th codebook vector at update step t, m is the momentum term, and the update is based on the average of all input vectors $\mathbf{z}_i$ assigned to code $\mathbf{k}$ in the current batch. However,

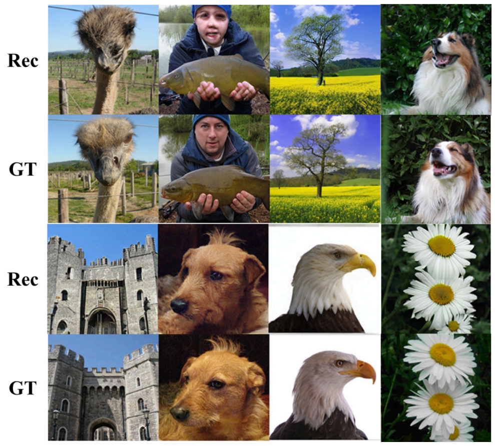


Figure 3: Visualization of reconstruction by quantized semantic features. Semantic codebook tends to produce inaccuracies in the reconstruction of image textures and color information.

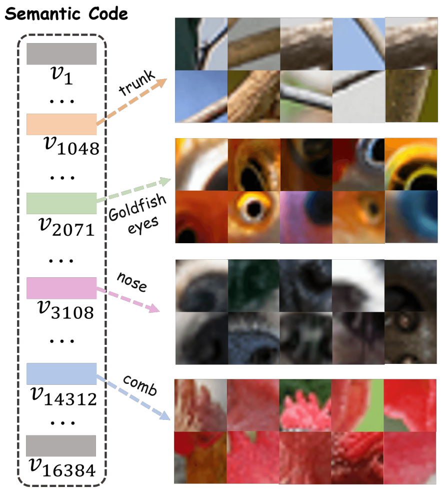


Figure 4: Visualization of semantic code. Each code corresponds to a set of image patches that share similar pixel-level features.

we further conduct an experiment that reconstructs the original pixel from the quantized semantic features extracted by $\mathcal{C}_{sem}$. The reconstructed images exhibited noticeable blurriness and a significant loss of high-frequency details, as shown in Fig. 3. This indicates that the semantic codebook lacks pixel information.

2.2 Pixel Reconstruction Enablement

Discussion. In section 2.1, we demonstrate that semantic code lacks the ability to model pixel information. In order to pixel reconstruction enablement and avoid a reduction of understand ability, a straightforward approach is to add an extra VQGAN (Yu et al., 2021a) model. Semantic codebook extracts discrete semantic tokens for multimodal understanding, and VQGAN extracts discrete texture tokens for generation. The two token sets are concatenated—either dimensionally or sequentially, and passed to the LLM. However, the resulting token inflation or oversized codebook introduces a significant computational burden, limiting its feasibility for MLLMs.

Furthermore, we present the visualization results of the semantic code, as shown in Fig.4. It can be observed that image patches corresponding to the same code exhibit similar pixel features. For example, the code v_{14312} is more likely to be assigned to the rooster comb element in the image. At the same time, the image patches corresponding to these combs exhibit similar pixel features, such as color, patterns, and shapes. Based on this observation, we propose Semantic-Guided Hierarchical Codebook to model the pixel feature space corresponding to each semantic code using a sub-codebook.

Semantic-Guided Hierarchical Codebook (SGHC). The SGHC consists of a pretrained semantic codebook and several sub-codebooks, where each sub-codebook corresponds to a semantic code of the semantic codebook, as shown in figure 2 (a). Specially, given the pre-trained semantic codebook $\mathcal{C}_{sem} = \{c_1, c_2, ..., c_K\} \in \mathbb{R}^{K \times d_{sem}}$, the pixel codebook $\mathcal{C}_{pix} = \{\mathcal{C}^1_{pix}, \mathcal{C}^2_{pix}, \dots, \mathcal{C}^K_{pix}\} \in \mathbb{R}^{K \times m \times d_{pix}}$, where $\mathcal{C}^k_{pix} \in \mathbb{R}^{m \times d_{pix}}$ is k_{th} semantic code's pixel sub-codebook, m is sub-codebook size. At first, the semantic codebook quantizes X to a discrete semantic token Z_{sem} and a semantic codebook index I_{sem}. In parallel, pixel encoder $\mathcal{E}_{pix}$ extract continuous pixel features $Z_{pix} = \mathcal{E}_{pix}(X)$. For the quantization process of the pixel codebook, the corresponding pixel sub-codebook is selected based on the quantization result of the semantic codebook. For instance, given image patch i, its semantic quantization codebook index k and continuous pixel feature Z^i_{pix}, SGHC selects pixel sub-codebook $\mathcal{C}^k_{pix}$ to quantize Z^i_{pix}. The process is as follows:

$$Z^i_{q_{pix}}, I^i_{q_{pix}} = \underset{j \in \{1, \dots, m\}}{\arg\min} \|Z^i_{pix} - \mathcal{C}^k_{pix}[j]\| \tag{5}$$

Table 1: Comparison of reconstruction quality on the ImageNet-50k validation set. ‡: quantizer uses residual quantization (RQ), where the total Code Shape is multiplied by RQ depth. †: quantizer uses multiple codebooks and product quantization.

Method	Res	Code Shape	Codebook Size	rFID↓
Only Reconstruction				
LlamaGen (Sun et al., 2024a)	256	16×16	16,384	2.19
RQVAE (Lee et al., 2022)	256	$16 \times 16 \times 4$ ‡	16,384	3.20
VQGAN-LC (Zhu et al., 2024a)	256	16×16	100,000	2.62
IBQ (Shi et al., 2024)	256	16×16	16,384	1.37
IBQ (Shi et al., 2024)	256	16×16	262,144	1.00
FQGAN (Bai et al., 2024)	256	$16 \times 16 \times 2$	$16,384 \times 2$†	0.94
Unified				
VILA-U (Wu et al., 2024c)	256	16x16x4‡	16,384	1.80
SDE(MUSE-VL) (Xie et al., 2025)	256	16×16	32,768	2.26
TokenFlow (Qu et al., 2024)	256	680	32,768	1.37
TokLIP (Lin et al., 2025)	256	16×16	16,384	2.19
QLIP-B (Lin et al., 2025)	256	16×16	2^{28}	3.21
UniTok (Ma et al., 2025)	256	16×16	$16,384 \times 4$†	0.39
SemHiTok(ours)	256	16×16	196,608	1.16
SemHiTok(ours)	384	27×27	196,608	0.66

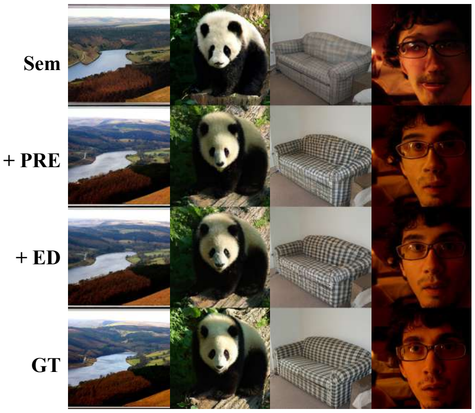


Figure 5: Visualized reconstruction results from the ablation of key modules. PRE brings about a significant improvement in reconstruction quality. Moreover, the Enhance Decoder(ED) further improves reconstruction on hard samples.

where j is the selected sub-codebook internal index. Finally, the semantic quantized tokens and pixel quantized tokens are concatenated to $Z_q = concate_{dim}(Z_{q_{sem}}, Z_{q_{pix}})$ as the input of pixel decoder $\mathcal{D}_{pix}$ to reconstruct the raw pixel image:

$$\hat{X} = \mathcal{D}_{pix}(Z_q) \tag{6}$$

where $\hat{X}$ is reconstructed pixel image. The $\mathcal{E}_{pix}$, C_{pix} and $\mathcal{D}_{pix}$ are end-to-end trainable by minimizing reconstruction loss $L_{img} = \ell_1(\hat{X}, X)$, codebook loss Lc, perceptual loss L_{per} and represents adversarial loss L_{gan} (Yu et al., 2021a). The reconstruction loss is formulated as:

$$L_{rec} = L_{img} + \lambda_1 Lc + \lambda_2 L_{per} + \lambda_3 L_{gan} \tag{7}$$

where λ_1, λ_2 and λ_3 are loss weight of each item.

Our SGHC can be regarded as the refinement of a semantic discrete space to enable pixel reconstruction. We place a specific emphasis on two key advantages of SGHC: (1) ***Non-Conflicting Extension:*** Our method leverages a pre-trained semantic codebook as a foundation, with pixel reconstruction losses exclusively employed to optimize pixel branch modules during the PRE. This strategic approach effectively circumvents the suboptimal solutions that arise from joint optimization processes. Furthermore, SGHC's final output is generated by concatenating semantic-quantized features with pixel-quantized features, preserving the full expressive capacity of the original semantic features while integrating complementary texture information through this unified feature fusion paradigm. Subsequent tasks, such as reconstruction, multimodal understanding, and generation, all share the same discrete token representation; (2) ***Efficient Downstream Applications:*** SGHC effectively avoids two critical predicaments: token quantity inflation and codebook overexpansion. As defined before, the semantic codebook size is K, and each pixel sub-codebook size is m. Due to dimensional concatenation, the complete codebook flattens to $K \times m$, where m is much smaller than K. In our experimental default settings, we extend the complete codebook to a size comparable to existing LLMs' text vocabulary size, e.g., the size of Qwen2 vocabulary is 150k.

2.3 Unified MLLM

The framework diagram for unified MLLM is shown in Fig.2(c). We use SemHiTok to develop a unified multimodal model, which models discrete vision and text token sequences with a universal next-token prediction loss. Particularly, in image processing, SemHiTok is utilized to discretize images into token sequences. On the model side, we merely expand the text vocabulary and adjust the head layer to incorporate visual token IDs. To enable a unified head layer, we flatten SGHC by merging all sub-codebooks into a single flat representation as shown in Fig.2(b). Specifically, for the

Table 2: Comparative analysis of tokenizers on multimodal comprehension tasks. *: Both the tokenizer and LLM are reproduced in our setting. Our method achieves SOTA performance compared with other discrete tokenizers.

Model	LLM	Data	Res.	POPE	MME-P	SEED	GQA
SigLIP (Zhai et al., 2023)	Vicuna-7B	LLaVA-v1.5	256	83.76	1481.0	65.28	61.9
LlamaGen (Sun et al., 2024a)	Vicuna-7B	LLaVA-v1.5	256	65.6	716.8	35.0	39.8
VILA-U (Wu et al., 2024c)	Vicuna-7B	LLaVA-v1.5	256	81.6	1311.6	56.9	55.3
SDE(MUSE-VL)* (Xie et al., 2025)	Vicuna-7B	LLaVA-v1.5	256	77.3	1240.0	56.7	58.0
TokLIP (Lin et al., 2025)	Qwen2.5-7B-Ins	LLaVA-v1.5	256	81.2	1346.8	59.8	57.4
SemHiTok(Ours)	Vicuna-7B	LLaVA-v1.5	256	**82.5**	**1355.8**	**62.9**	**60.3**
TokenFlow-384 (Qu et al., 2024)	Vicuna-7B	LLaVA-v1.5	384	84.9	1416.4	62.7	61.2
TokLIP (Lin et al., 2025)	Qwen2.5-7B-Ins	LLaVA-v1.5	384	82.7	1410.2	**65.2**	59.3
SemHiTok-384(Ours)	Vicuna-7B	LLaVA-v1.5	384	**86.3**	**1465.6**	64.1	**62.3**

j_{th} semantic code in the i_{th} pixel sub-codebook, the discrete code index in the completed codebook is $h = i \times m + j$, where m is the sub-codebook size. It is also worth noting that the vocabulary expansion is merely for implementation convenience. We still use the features extracted from SGHC as input and align with LLM through a lightweight adapter layer. After training is completed, we replace the visual component in the embedding layer in order to achieve consistency between training and inference. To enable LLM to better handle features at two different levels, we introduce a Dual-MLP adapter layer, which projects semantic features and pixel features separately, and then concatenates them along the dimension before feeding them into the LLM. To enable classifier-free guidance (Ho & Salimans, 2022), we randomly replace the text condition with a probability of 0.1 to the unconditioned text during training. More deployment details are provided in the **Supplementary Material 8.2**.

3 Experiments

3.1 Experimental Setup

Tokenizer. For the semantic branch, we employ SigLIP (Zhai et al., 2023) as the semantic encoder and three self-attention layers as the semantic decoder to reconstruct semantic features. For the pixel branch, we employ ViT as both the pixel encoder and decoder, assigning 8 pixel sub-codes to each semantic code. More tokenizer detail, please refer to **Supplementary Material 8.3**.

Unified MLLM. We use Qwen2.5-7B-Instruct (Yang et al., 2024) as the base LLM, and expand its vocabulary and output head layer. We evaluate visual understanding on standard VQA benchmarks including SEEDB (Li et al., 2023a), POPE (Li et al., 2023c), GQA (Hudson & Manning, 2019), MMMU (Yue et al., 2024), MMB (Liu et al., 2024b) and MME (Fu et al., 2023). For visual generation evaluation, we report results on MJHQ-30K (Li et al., 2024c) and GenAI-Bench (Li et al., 2024a). More unified MLLM detail, please refer to **Supplementary Material 8.4**.

3.2 Unified Image Tokenizer

Image Reconstruction. We present the reconstruction performance on the ImageNet-50k validation set in Tab.1. Notably, SemHiTok excels in reconstruction quality compared to the unified tokenizer, recording an impressive 1.16 rFID with 16× downsampling ratio. While SemHiTok's vocabulary size(approx 196k) appears larger than baselines, it is crucial to consider the effective representational capacity defined by the code shape. VILA-U(using RQ) and FQGAN(using Product Quantization) operate within a combinatorial search space(N^D or $N_1 \times N_2$), resulting in a significantly larger effective capacity than SemHiTok's linearly constrained structure($K \times m$). Furthermore, these approaches typically employ a denser code shape, whereas SemHiTok relies on a single unified hierarchical index. This indicates that our superior performance derives from the structured efficiency of SGHC rather than brute-force capacity expansion. Increasing the training 384 resolution led to a significant improvement in the rFID score, reaching 0.66. The results validate the effectiveness of SGHC design in modeling pixel feature space of the semantic code.

Table 3: Quantitative results on multimodal understanding benchmarks. SemHiTok achieves SOTA performance on most benchmarks among *Und&Gen Discrete* MLLMs, and is comparable to or even surpasses some *Only Und* and *Und&Gen. Continuous* models. The performance on *Und&Gen Discrete* with top-1 and top-2 values is denoted in bold and underline, respectively.

Method	**# Params**	**Res.**	**SEED**	**POPE**	**GQA**	**MMMU**	**MMB**	**MME**	**MME-P**	**MMV**
Only Und.										
LLaVA-Phi (Zhu et al., 2024b)	2.7B	256	-	85.0	-	-	59.8	-	1335.1	28.9
LLaVA-v1.5 (Liu et al., 2023b)	7B	336	58.6	85.9	62.0	35.4	64.3	-	1510.7	31.1
Qwen-Vl-Chat (Bai et al., 2023b)	7B	448	57.7	-	57.5	30.5	-	1848.3	1487.5	-
ShareGPT4V (Chen et al., 2024b)	7B	336	69.7	-	63.3	37.2	68.8	1943.8	1567.4	37.6
Und&Gen. Continuous										
LaVIT (Jin et al., 2023)	7B	224	-	-	46.8	-	58.0	-	-	-
Janus (Wu et al., 2024a)	1.5B	384	63.7	87.0	59.1	30.5	69.4	-	1338.0	34.3
Janus-Pro-1B (Chen et al., 2025)	1.5B	384	68.3	86.2	58.9	38.9	65.5	-	1444.0	-
MAR (Wu et al., 2025)	1.5B	512	67.1	87.6	58.9	38.9	65.5	-	1155.0	-
Und&Gen. Discrete										
LWM (Liu et al., 2024a)	7B	256	-	75.2	44.8	-	-	-	-	9.6
SEED-LLaMA (Li et al., 2024b)	13B	256	53.7	-	-	-	-	-	-	-
Show-o (Xie et al., 2024)	1.5B	256	-	80.0	-	26.7	-	-	1097.2	-
Liquid (Wu et al., 2024b)	7B	512	-	81.1	**71.3**	-	-	-	1119.3	-
EMU3 (Wang et al., 2024c)	8B	512	68.2	85.2	60.3	31.6	58.5	1509.9	1243.8	**37.2**
VILA-U (Wu et al., 2024c)	7B	256	56.3	83.9	58.3	-	-	-	1336.2	27.7
VILA-U (Wu et al., 2024c)	7B	384	59.0	**85.8**	60.8	-	-	-	1401.8	33.5
UniToken (Jiao et al., 2025)	7B	384	69.3	-	-	32.8	69.9	-	-	-
TokLIP (Lin et al., 2025)	7B	384	76.9	84.1	59.5	**43.1**	67.6	-	1488.4	29.8
TokenFlow-B (Qu et al., 2024)	13B	224	60.4	84.0	59.3	34.2	55.3	1660.4	1353.6	22.4
TokenFlow-L (Qu et al., 2024)	13B	256	62.6	85.0	60.3	34.4	60.3	1622.9	1365.4	27.7
SynerGen-VL (Li et al., 2025)	2.4B	512	62.0	85.3	59.7	34.2	53.7	1837.0	1381.0	34.5
SemHiTok(Ours)	7B	256	69.7	83.4	60.3	39.3	72.3	1775.9	1449.0	30.5
SemHiTok(Ours)	7B	384	**79.8**	85.5	61.7	41.0	**75.2**	**1993.8**	**1512.8**	36.6

Table 4: Comparison of generation quality on GenAI and MJHQ30K. SemHiTok achieves comparable results with specialist models and unified MLLMs.

Model	Params	Type	#Training Images	Res.	GenAI-Bench		MJHQ30K
					Basic ↑	Advanced ↑	gFID ↓
Only Gen.							
SD v2.1 (Rombach et al., 2022b)	–	Diff	2000M	1024	0.78	0.62	–
DALL-E 3 (Betker et al., 2023)	–	Diff	-	1024	0.90	0.70	–
PixArt-α (Chen et al., 2023)	0.6B	Diff	-	1024	–	–	6.14
SDXL (Podell et al., 2023)	2.6B	Diff	2000M	1024	0.83	0.63	9.55
Playgroundv2.5 (Li et al., 2024c)	–	Diff	-	1024	–	–	4.48
Und&Gen.							
LWM (Liu et al., 2024a)	7B	AR	-	256	0.63	0.53	17.77
Show-o (Xie et al., 2024)	1.5B	Diff	36M	256	0.70	0.60	15.18
Janus (Wu et al., 2024a)	1.3B	AR	65M	384	–	–	10.10
VILA-U (Wu et al., 2024c)	7B	AR	15M	256	0.76	0.64	12.81
VILA-U (Wu et al., 2024c)	7B	AR	15M	384	0.73	0.61	7.69
SynerGen-VL (Li et al., 2025)	2.4B	AR	25M	256	-	-	6.10
SDE(MUSE-VL) (Xie et al., 2025)	7B	AR	10M	256	-	-	7.73
ILLUME (Wang et al., 2024a)	7B	AR	15M	512	0.75	0.60	7.76
ILLUME+ (Huang et al., 2025a)	7B	AR	15M	512	0.75	0.60	7.76
TokenFlow (Qu et al., 2024)	7B	AR	60M	256	0.65	0.65	7.78
Liquid (Wu et al., 2024b)	7B	AR	30M	512	**0.83**	0.65	5.47
UniTok (Ma et al., 2025)	7B	AR	30M	256	0.85	0.67	7.46
SemHiTok(ours)	7B	AR	15M	256	**0.83**	0.64	**5.40**
SemHiTok(ours)	7B	AR	15M	384	**0.83**	**0.66**	5.70

LLaVa-v1.5 Multimodal Understanding. To ensure a fair comparison, we conduct experiments to evaluate the multimodal understanding performance of existing open-source tokenizers and SemHiTok under the standard LLaVA-v1.5 setting. The results are as shown in Tab.2. LlamaGen's performance is notably inferior, a consequence of its deficient pre-alignment with text. In contrast, unified image tokenizers show considerable advancements in understanding. However, due to their inherent hybrid structures or joint training strategies, a substantial disparity in understanding performance remains between prior discrete tokenizers and continuous representations. Notably, even though TokLIP uses a stronger base model (Qwen2.5-7B-Ins), its performance remains inferior to ours. SemHiTok achieves state-of-the-art results for discrete tokenizers, nearing the performance of continuous inputs such as SigLIP.

3.3 Unified MLLM

Multimodal Understanding. We evaluate the understanding performance of SemHiTok on diverse benchmarks in Table 3. *Und* means support multimodal understanding, *Gen* means support text-to-image generation. Among *Und&Gen. Discrete*, SemHiTok achieves state-of-the-art performance on most metrics, such as SEED, MMB, MME, and MME-P. Compared to other unified tokenizers (e.g., VILA-U, TokenFlow, UniToken), our models demonstrate significant advantages. Notably, our model surpasses expert-level models on key benchmarks, achieving 3.8 and 6.4 points higher than ShareGPT4V on MMMU and MMB, respectively. This bridges the gap between discrete visual tokens and continuous visual tokens in multimodal understanding tasks, strongly demonstrating the superiority and potential of our approach. Visualizations on understanding tasks are available in **Supplementary Material 8.6**.

Text-to-Image Generation To evaluate the text-to-image generation, we use GenAI-bench (Li et al., 2024a) and MJHQ30K (Li et al., 2024c) benchmark, and the results are shown in Tab4. For GenAI-bench, we use clip-flant5-xxl as the VQA score model to reflect the consistency between text descriptions and generated images. On this challenging benchmark, our model achieves competitive performance, closely matching Liquid, even though Liquid employs a generation-focused tokenizer. Furthermore, our model even outperformed some diffusion-based expert models, such as SDXL and SD v2. The strong results underscore the superior capability of our unified MLLM in complex text-to-image generation tasks. For MJHQ30K, we use the generation FID metric on generated images and high-quality images. On this benchmark, SemHiTok-256 attains 5.40 gFID, setting a new state-of-the-art in autoregressive image generation. In addition, we conduct further quantitative comparisons on GenEval (Ghosh et al., 2023) and DPG (Hu et al., 2024), which are available in **Supplementary Material 8.7**. More quantitative analysis results and visualizations on generation tasks are available in **Supplementary Material 8.8**.

3.4 Quantitative analysis of pixel features in the semantic codebook

In Fig.4, we provide visualization results showing that image patches corresponding to the same semantic code share similar pixel features. To further support this analysis, we attempt to provide quantitative evidence using the following method:

1. Using the tokenizer to extract the code indices along with their corresponding image patches.
2. Applying the **DCT(Discrete Cosine Transform)** (Rao & Yip, 2014) to extract patch's pixel features.
3. For each code, we compute the variance of the DCT features extracted from its associated image patches, followed by **averaging these variances over all codes**, denoted as V_{mean}. The more similar the pixels of the tokenized results are, the smaller the V_{mean} becomes.
4. Computing the **global variance of all image patches**, denoted as V_{global}.
5. Computing **Variance Reduction Ratio (VRR)**, denoted as $VRR = 1 - V_{mean}/V_{global}$. The larger VRR indicates that the patches corresponding to the given code have more similar pixel features.

We randomly sampled 10K images from the ImageNet-Val (50K) dataset across 5 independent runs. The Random Baseline refers to randomly tokenized results, serving as the lower bound for the VRR metric. VQGAN (Yu et al., 2021a), as an expert model for pixel reconstruction, serves as the standard reference for the VRR metric.

As illustrated in Tab. 5, the results provide strong quantitative evidence for the effectiveness of our Semantic-Guided Hierarchical Codebook (SGHC) design. **1.)Validation of Semantic-Pixel Correlation.** The SemHiTok$_{\text{sem}}$ achieves a VRR of 5.245%, which is significantly higher than the Random Baseline (0.185%). This quantitatively verifies our observation in Subsec2.2 that image patches assigned to the same semantic code naturally exhibit intrinsic pixel-level similarities. This correlation forms the theoretical foundation for using semantic codes to guide pixel quantization. **2.)The Necessity of Pixel Branch.** However, compared to VQGAN (9.535%), the lower VRR

Table 5: Quantitative evaluation of pixel consistency using Variance Reduction Ratio (VRR). SemHiTok$_{\text{sem}}$ significantly outperforms the random baseline, verifying the intrinsic semantic-pixel correlation. Furthermore, our hierarchical refinement (SemHiTok$_{\text{pix}}$) achieves the best consistency, exceeding the standard VQGAN baseline.

Experiment	Random Baseline	SemHiTok$_{\text{sem}}$	VQGAN	SemHiTok$_{\text{pix}}$
VRR	0.185%±0.005%	5.245%±0.085%	9.535%±0.095%	13.705%±0.075%

Table 6: Impact of key design choices on reconstruction and multimodal understanding.

Semantic Codebook	SGHC	Dual MLP	Enhance Decoder	MME-P↑	MMB↑	SEED↑	MMU↑	rFID↓
✓				1387.5	61.3	62.3	35.6	3.17
✓	✓			1355.8	60.7	62.9	35.8	1.42
✓	✓	✓		1393.0	61.6	63.2	36.1	1.42
✓	✓	✓	✓	1393.0	61.6	63.2	36.1	1.16

of SemHiTok$_{\text{sem}}$ confirms that semantic codes alone are insufficient to capture fine-grained high-frequency details. This justifies the necessity of our pixel branch design to bridge the gap between semantic abstraction and pixel fidelity. **3.)Superiority of Hierarchical Design.** Most notably, SemHiTok$_{\text{pix}}$ achieves the highest VRR of **13.705%**, surpassing the expert pixel reconstruction model VQGAN by a substantial margin (+4.17%). This result demonstrates that decomposing the complex global pixel space into semantic-conditioned sub-spaces allows for significantly tighter feature clustering and lower variance than standard global quantization methods. This explains why SemHiTok achieves superior reconstruction performance despite utilizing a decoupled training strategy.

4 Ablation

4.1 Impact of Key Design.

In Tab.6, we validate the impact of our key design choices in SemHiTok: semantic codebook, semantic-guided hierarchical codebook(SGHC), Dual MLP, and Enhance Deocoder. For efficiency, we only test the understanding performance under LLaVA-v1.5, and train 40 epochs on ImageNet-1K for reconstruction tasks. We begin with the semantic codebook, which suffices for multimodal understanding but suffers from poor reconstruction. Incorporating SGHC enables pixel reconstruction and reduces rFID by 1.75, without noticeably affecting understanding performance. Introducing Dual-MLP further enhances multimodal understanding, even surpassing the semantic codebook alone, highlighting the effectiveness of multi-level feature modeling. Finally, a stronger pixel decoder brings additional improvements in reconstruction quality. figure 5 demonstrates the effects of various modules on reconstruction results. Compared to the semantic codebook, SGHC delivers more detailed reconstructions. Additionally, the enhanced decoder boosts performance on difficult samples. More reconstruction comparisons are available in **Supplementary Material 8.9**.

4.2 Impact of training strategy and structure.

As shown in Table 7 (**Exp 1**), we used the same architecture as SDE and adopted a joint training strategy. Its performance is observed to be the lowest across both multimodal understanding and reconstruction tasks. In **Exp 2**, using the same architecture as SemHiTok but with joint training, the performance showed an overall improvement compared to **Exp 1**. Furthermore, to examine the influence of architectural design, we conduct **Exp 3**, where two separately pretrained tokenizers are employed to independently extract semantic and pixel information. Specifically, we use the semantic branch of SemHiTok and LlamaGen. Compared with **Exp 3**, SemHiTok achieves superior performance on both multimodal understanding and reconstruction tasks. The above experiments demonstrate the advantages of structural decoupling and training decoupling in jointly modeling semantic and pixel-level information. **Exp 3** and **Exp 4** suggest that naively incorporating pixel

Table 7: Impact of training strategy and structure. DStruct: Using different models to extract information at various levels; DTrain: Adopt a phased optimization training strategy

Exp	DStruct	DTrain	Multimodal understanding				Reconstruction	
			GQA↑	MME-P↑	SEEDB↑	POPE↑	rFID↓	Usage↑
1	✗	✗	58.0	1240.0	56.7	77.3	3.78	92.9%
2	✓	✗	57.8	1357.4	55.3	80.4	3.22	45.9%
3	✓	✓	58.7	1210.9	56.1	80.1	2.19	97.0%
4	SemHiTok		60.3	1355.8	62.9	82.6	1.42	93.7%

features may negatively affect the alignment between image features and the LLM. This suggests that naively incorporating pixel features may negatively affect the alignment between image features and the LLM. In contrast, SemHiTok leverages pixel features as a complementary refinement to semantic features, thereby effectively bridging the gap between semantic and pixel representations.

4.3 More comparisons and ablation studies

In **Supplementary Material 8.5**, we compare our method with others across additional reconstruction metrics. Meanwhile, to further illustrate how different hyperparameters in our method affect model performance, we also conduct ablation studies on Concat Type and Sub-Codebook Size in **Supplementary Material 8.10**, Semantic VQ Type, Codebook dim and Codebook size in **Supplementary Material 8.11**, and different $K \times m$ and more comparison with other large codebook reconstruction methods in **Supplementary Material 8.12**.

5 Related Work

We present and analyze the Specialized Image Tokenizer and Unified Image Tokenizer. Please refer to **Supplementary Materia 8.13 and 8.14.**

6 Conclusion

Conclusion: In this work, we introduce SemHiTok, a unified image tokenizer that implements a better trade-off between semantic and pixel information, and is fully compatible with and readily deployable within existing next-token MLLMs architectures. SemHiTok innovatively utilizes a semantic-guided hierarchical codebook (SGHC) to realize the reconstruction capability of pixel features without affecting the understanding performance of the original semantic codebook, and achieves SOTA performance on multimodal understanding under LLaVA-v1.5 setting and on ImageNet-50k reconstruction in unified image tokenizers. We further develop a unified MLLM with SemHiTok, which demonstrates competitive performance on both understanding and generation tasks compared to existing unified MLLMs. This highlights the strong potential of SemHiTok and further bridges the gap between discrete and continuous tokenizers, providing the community with a powerful discrete tokenizer.

Limitation: We present two limitations: *(1). Low generation efficiency:* Owing to the use of standard image quantization methods and settings, each 256-resolution image is represented by 256 tokens. This results in relatively low efficiency and high computational cost. *(2). Unified large model potential:* In tasks involving natural language understanding and multimodal reasoning, advanced post-training techniques such as Chain-of-Thought (CoT) are not explored. This remains a promising direction for future research in the community.

Future Work: In the future, we can explore the potential of unified image tokenizers and test their performance on more difficult tasks, such as image editing and multiple rounds of conversations. In addition, improving the compression ratio of the model and designing a tokenizer that serializes the image in one dimension are also expected.

7 Statement

7.1 Ethics statement

This research does not involve potentially harmful insights, methodologies, or applications, and it raises no concerns regarding conflicts of interest, sponsorship, discrimination, bias, fairness, privacy, security, legal compliance, or research integrity.

7.2 Reproducibility statement

Data. The training datasets employed in this study are detailed in Appendix.8.3 and Appendix. 8.4, and all are publicly accessible open-source resources.

Method. To support reproducibility, we elaborate on the methodological details in Section.2, and the implementation will be open-sourced after acceptance.

Performance. All evaluations are carried out on open benchmarks, thereby ensuring the reproducibility of our results.

Code And Model Weight. We will open-source the code and model weight files after acceptance.

REFERENCES

Inclusion AI, Biao Gong, Cheng Zou, Dandan Zheng, Hu Yu, Jingdong Chen, Jianxin Sun, Junbo Zhao, Jun Zhou, Kaixiang Ji, et al. Ming-lite-uni: Advancements in unified architecture for natural multimodal interaction. *arXiv preprint arXiv:2505.02471*, 2025.

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*, abs/2308.12966, 2023a. URL `https://api.semanticscholar.org/CorpusID:263875678`.

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*, 1(2):3, 2023b.

Zechen Bai, Jianxiong Gao, Ziteng Gao, Pichao Wang, Zheng Zhang, Tong He, and Mike Zheng Shou. Factorized visual tokenization and generation. *arXiv preprint arXiv:2411.16681*, 2024.

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

Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset, 2022.

Shiyue Cao, Yueqin Yin, Lianghua Huang, Yu Liu, Xin Zhao, Deli Zhao, and Kaigi Huang. Efficient-vqgan: Towards high-resolution image generation with efficient vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7368–7377, 2023.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.

Kai Chen, Yunhao Gou, Runhui Huang, Zhili Liu, Daxin Tan, Jing Xu, Chunwei Wang, Yi Zhu, Yihan Zeng, Kuo Yang, et al. Emova: Empowering language models to see, hear and speak with vivid emotions. *arXiv preprint arXiv:2409.18042*, 2024a.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pp. 370–387. Springer, 2024b.

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.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, 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, 2024c.

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*, pp. 248–255. Ieee, 2009.

Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. Dreamllm: Synergistic multimodal comprehension and creation. *arXiv preprint arXiv:2309.11499*, 2023.

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.

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.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Jiaming Han, Hao Chen, Yang Zhao, Hanyu Wang, Qi Zhao, Ziyan Yang, Hao He, Xiangyu Yue, and Lu Jiang. Vision as a dialect: Unifying visual understanding and generation via text-aligned representations. *arXiv preprint arXiv:2506.18898*, 2025.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

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

Runhui Huang, Chunwei Wang, Junwei Yang, Guansong Lu, Yunlong Yuan, Jianhua Han, Lu Hou, Wei Zhang, Lanqing Hong, Hengshuang Zhao, et al. Illume+: Illuminating unified mllm with dual visual tokenization and diffusion refinement. *arXiv preprint arXiv:2504.01934*, 2025a.

Ziyuan Huang, DanDan Zheng, Cheng Zou, Rui Liu, Xiaolong Wang, Kaixiang Ji, Weilong Chai, Jianxin Sun, Libin Wang, Yongjie Lv, et al. Ming-univision: Joint image understanding and generation with a unified continuous tokenizer. *arXiv preprint arXiv:2510.06590*, 2025b.

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*, pp. 6700–6709, 2019.

J Stuart Hunter. The exponentially weighted moving average. *Journal of quality technology*, 18(4): 203–210, 1986.

Yang Jiao, Haibo Qiu, Zequn Jie, Shaoxiang Chen, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. Unitoken: Harmonizing multimodal understanding and generation through unified visual encoding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3600–3610, 2025.

Yang Jin, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, Quzhe Huang, Bin Chen, Chenyi Lei, An Liu, Chengru Song, et al. Unified language-vision pretraining in llm with dynamic discrete visual tokenization. *arXiv preprint arXiv:2309.04669*, 2023.

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*, pp. 11523–11532, 2022.

Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, et al. Genai-bench: Evaluating and improving compositional text-to-visual generation. *arXiv preprint arXiv:2406.13743*, 2024a.

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*, 2023a.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13299–13308, 2024b.

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

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. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29767–29779, 2025.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 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, 2023b.

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*, 2023c.

Wing Lian, Bleys Goodson, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknium". Openorca: An open dataset of gpt augmented flan reasoning traces, 2023.

Haokun Lin, Teng Wang, Yixiao Ge, Yuying Ge, Zhichao Lu, Ying Wei, Qingfu Zhang, Zhenan Sun, and Ying Shan. Toklip: Marry visual tokens to clip for multimodal comprehension and generation. *arXiv preprint arXiv:2505.05422*, 2025.

Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with ringattention. *arXiv e-prints*, pp. arXiv–2402, 2024a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023b.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024b.

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.

Chuofan Ma, Yi Jiang, Junfeng Wu, Jihan Yang, Xin Yu, Zehuan Yuan, Bingyue Peng, and Xiaojuan Qi. Unitok: A unified tokenizer for visual generation and understanding. *arXiv preprint arXiv:2502.20321*, 2025.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

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

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*, pp. 8748–8763. PmLR, 2021a.

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*, pp. 8748–8763. PmLR, 2021b.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

K Ramamohan Rao and Ping Yip. *Discrete cosine transform: algorithms, advantages, applications*. Academic press, 2014.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022a.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022b.

Fengyuan Shi, Zhuoyan Luo, Yixiao Ge, Yujiu Yang, Ying Shan, and Limin Wang. Taming scalable visual tokenizer for autoregressive image generation. *arXiv preprint arXiv:2412.02692*, 2024.

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*, 2024a.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

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*, pp. 14398–14409, 2024b.

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

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.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

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*, 2024a.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024b.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024c.

Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14668–14678, 2022.

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*, 2024a.

Junfeng Wu, Yi Jiang, Chuofan Ma, Yuliang Liu, Hengshuang Zhao, Zehuan Yuan, Song Bai, and Xiang Bai. Liquid: Language models are scalable multi-modal generators. *arXiv preprint arXiv:2412.04332*, 2024b.

Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Zhonghua Wu, Qingyi Tao, Wentao Liu, Wei Li, and Chen Change Loy. Harmonizing visual representations for unified multimodal understanding and generation. *arXiv preprint arXiv:2503.21979*, 2025.

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

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*, 2024d.

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.

Rongchang Xie, Chen Du, Ping Song, and Chang Liu. Muse-vl: Modeling unified vlm through semantic discrete encoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24135–24146, 2025.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. Wizardlm: Empowering large pre-trained language models to follow complex instructions. In *The Twelfth International Conference on Learning Representations*, 2024a.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. *arXiv preprint arXiv:2406.08464*, 2024b.

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.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. Mmada: Multimodal large diffusion language models. *arXiv preprint arXiv:2505.15809*, 2025.

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*, 2021a.

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*, 2021b.

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.

Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. An image is worth 32 tokens for reconstruction and generation. *Advances in Neural Information Processing Systems*, 37:128940–128966, 2025.

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*, pp. 9556–9567, 2024.

Kaiwen Zha, Lijun Yu, Alireza Fathi, David A Ross, Cordelia Schmid, Dina Katabi, and Xiuye Gu. Language-guided image tokenization for generation. *arXiv preprint arXiv:2412.05796*, 2024.

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

Jiahui Zhang, Fangneng Zhan, Christian Theobalt, and Shijian Lu. Regularized vector quantization for tokenized image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18467–18476, 2023.

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.

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.

Lei Zhu, Fangyun Wei, Yanye Lu, and Dong Chen. Scaling the codebook size of vqgan to 100,000 with a utilization rate of 99%. *arXiv preprint arXiv:2406.11837*, 2024a.

Yichen Zhu, Minjie Zhu, Ning Liu, Zhiyuan Xu, and Yaxin Peng. Llava-phi: Efficient multi-modal assistant with small language model. In *Proceedings of the 1st International Workshop on Efficient Multimedia Computing under Limited*, pp. 18–22, 2024b.

Yongxin Zhu, Bocheng Li, Yifei Xin, Zhihua Xia, and Linli Xu. Addressing representation collapse in vector quantized models with one linear layer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22968–22977, 2025.

8 Technical Appendices and Supplementary Material

8.1 LLM Usage Statement

We clarify that the use of LLMs in this study is restricted to writing assistance, specifically for grammar correction and enhancing readability. No LLM was involved in the research design, experimental execution, or data analysis.

8.2 Deployment Details of Unified MLLM.

Training Data Form. For multimodal understanding, we use <lim_startl> and <lim_endl> to delimit the image segment within the input sequence. To distinguish between modalities and enable visual content generation, we insert special tokens: <IMG_XXXXX>, which represent image codes in LLMs' vocabulary. Specifically, the i-th code in the unified tokenizer codebook corresponds to <IMG_i>. In addition, we add <start_of_image> and <end_of_image> to indicate the start and end of image generation.

Vocabulary Embedding Processing Flow. For understanding samples, we follow the LLaVA setting: discrete features are first extracted by the tokenizer and then fed into the LLM through an adapter layer. To ensure consistency between understanding and generation, we employ the same features for generation samples instead of using the embeddings of <IMG_XXXXX>. After training, we replace the visual code embedding in the vocabulary to align training with inference. The detailed procedure is illustrated in Algorithm 1.

Algorithm 1: Procedure: Training and Generation Pipeline

```
Before Training:;
1. Add special tokens <IMG_XXXXX> into LLM vocabulary 2. Preprocess training images, e.g.
 "generate a dog + <Image>" $\rightarrow$
 "generate a dog + <start_of_image><IMG_i>...<IMG_k><end_of_image>"
During Training:;
1. Feed preprocessed samples into LLM tokenizer $\rightarrow$ text_id;
2. Lookup embedding $V$ from text_id $\rightarrow E_{text}$;
3. Get image IDs from text_id $\rightarrow$ img_id;;
Lookup unified codebook embedding from img_id $\rightarrow E_{img}$;;
Apply adapter layer: $E_{img} \rightarrow E'_{img}$;
4. Replace image embeddings in $E_{text}$ with $E'_{img} \rightarrow E'_{text}$;
5. Feed $E'_{text}$ into LLM backbone for training;
Generation:;
1. Apply adapter layer on codebook embeddings $C_{img} \rightarrow C'_{img}$;
2. Replace image part of $V$ with $C'_{img} \rightarrow V'$;
3. Start autoregressive generation;
```

8.3 Tokenizer Experimental details

The training of SemHiTok is conducted in two stages. During semantic codebook training, we train the semantic tokenizer for one epoch on 50M subset of COYO-700M(Byeon et al., 2022). For the PRE stage, we first train the model(ViT-Base) on ImageNet(Deng et al., 2009), and then fine-tune on 50M COYO to improve its generalization, following LlamaGen(Sun et al., 2024a). To further improve the reconstruction and generation performance, we enlarge the size of the pixel decoder(ViT-Large) and fine-tune on the 20M COYO data and 20M MidJourney-style synthetic data. The full training takes about 3 days on 32 v100 GPUs.

8.4 Unified MLLM Experimental details

Following existing work(Wu et al., 2024c; Ma et al., 2025), we first pretrain the model and adapter layer on a mix of multimodal data, which is composed of 3.5M language data from Magpie(Xu et al.,

Table 8: More Comparison on more metrics for the reconstruction task.

Method	Res	Code Shape	Codebook Size	rFID↓	PSNR↑	SSIM↑
LlamaGen (Sun et al., 2024a)	256	16×16	16,384	2.19	20.65	0.54
Show-o (Xie et al., 2024)	256	16×16	8192	3.21	21.34	0.59
TokenFlow (Qu et al., 2024)	256	16×16	16,384	1.37	21.41	0.69
SDE(MUSE-VL) (Xie et al., 2025)	256	16×16	32,768	2.26	20.14	0.65
QLIP-B (Lin et al., 2025)	256	16×16	2^{28}	3.21	21.34	0.59
UniTok (Ma et al., 2025)	256	$16 \times 16 \times 8$	$16,384 \times 4^{\dagger}$	0.41	27.28	0.77
SemHiTok(ours)	256	16×16	196,608	1.16	21.38	0.69

2024b) and Openorca(Lian et al., 2023), 10M caption image-text pairs data, and 15M MidJourney-style synthetic data. Subsequently, we finetune the model on 1M language dataset from Magpie(Xu et al., 2024b) and Evol-Instruct(Xu et al., 2024a), 4M generation data and 4M understanding data from emova(Chen et al., 2024a) and LLaVA-SFT(Liu et al., 2023b). The full training takes about 7 days on 32 A800 GPUs. During image generation inference, we apply classifier-free guidance(Ho & Salimans, 2022) with a scale factor of 2.5.

8.5 Comparison with other methods on more metrics for the reconstruction task

We present the reconstruction metrics of several tokenizers in Tab. 8, including rFID, PSNR, and SSIM. Our method achieves competitive performance across multiple reconstruction task metrics.

8.6 Visualizations on understanding tasks

We present more visualization of multimodal understanding samples in figure 6.

8.7 More Comparison of generation on GenEval and DPG

To more fully demonstrate the superiority of SemHiTok, we conduct further comparisons with other generative models and unified models on GenEval Ghosh et al. (2023) and DPG Hu et al. (2024). As shown in Tab.9, our method still demonstrates competitive performance.

8.8 Visualizations on generation tasks

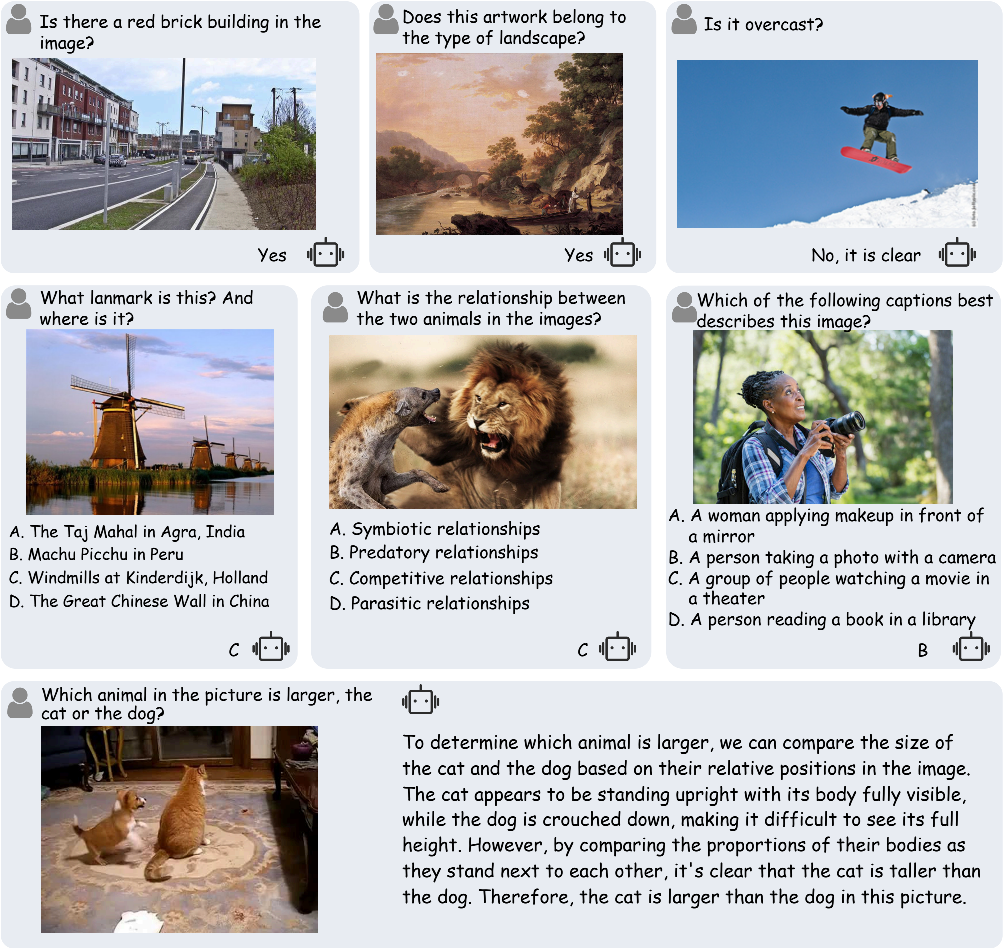
We present more visualization of generated images in figure 7.

8.9 More Visualized reconstruction results from the ablation of key module

We show more reconstruction effects on the ablation of key modules in figure 8.

8.10 Impact of Concat Type and Sub-Codebook size.

For efficiency, we only conduct training and evaluation on ImageNet-1K. We investigate the impact of the concat type of semantics between pixel and the sub-codebook size, on reconstruction performance as shown in Tab. 10. For the Concat type, w/o semantic tokens leads to a significant drop in reconstruction quality, with s core increase of 0.73 compared to the default setting. Furthermore, concatenation along the sequence length performs worse than along the dimension, as the two token sets are spatially aligned, making dimensional concatenation more appropriate. For sub-codebook size, increasing the size can improve the model's reconstruction performance, but it exhibits marginal utility. In addition, the codebook usage significantly decreases when the sub-codebook size is set to 16, which indicates that too large a sub-codebook size is not cost-effective.



Is there a red brick building in the image?

Yes

Does this artwork belong to the type of landscape?

Yes

Is it overcast?

No, it is clear

What lanmark is this? And where is it?

A. The Taj Mahal in Agra, India
B. Machu Picchu in Peru
C. Windmills at Kinderdijk, Holland
D. The Great Chinese Wall in China

C

What is the relationship between the two animals in the images?

A. Symbiotic relationships
B. Predatory relationships
C. Competitive relationships
D. Parasitic relationships

C

Which of the following captions best describes this image?

A. A woman applying makeup in front of a mirror
B. A person taking a photo with a camera
C. A group of people watching a movie in a theater
D. A person reading a book in a library

B

Which animal in the picture is larger, the cat or the dog?

To determine which animal is larger, we can compare the size of the cat and the dog based on their relative positions in the image. The cat appears to be standing upright with its body fully visible, while the dog is crouched down, making it difficult to see its full height. However, by comparing the proportions of their bodies as they stand next to each other, it's clear that the cat is taller than the dog. Therefore, the cat is larger than the dog in this picture.

Figure 6: Visualizations on understanding tasks.

8.11 Impact of Semanitc VQ Type, Codebook dim and Codebook size.

In Tab.11, we ablate VQ type and codebook dimension on multimodal understanding. Specifically, Norm VQ is not suitable for semantic discretization and shows the worst performance in understanding tasks. This indicates that it is difficult to discretely model complex and rich semantic information by Norm VQ. Replacing it with vanilla VQ brings a clear improvement. To further stabilize the semantic codebook, we use EMA VQ, which achieves the best results. For the semantic codebook dimension, higher values offer better representation but with a marginal effect. We empirically set the dimension to 48 as the default.

Furthermore, we investigate the impact of semantic codebook size K on multimodal understanding in Table 12. The results indicate that while larger codebooks generally enhance performance, increasing K beyond 16,384 yields only marginal gains and leads to a noticeable decline in codebook usage. Therefore, we select $K = 16,384$ as the optimal balance between representational capacity and utilization efficiency.

8.12 Impact of different $K \times m$ and more comparison with other large codebook reconstruction methods.

We investigate the impact of semantic codebook size K and sub-codebook size m in Table 13. The results indicate that increasing either K or m generally enhances reconstruction fidelity (lower rFID). However, excessive expansion leads to a significant decline in codebook usage. Furthermore, when

Table 9: Comparison with other methods on GenEval and DPG Bench. SemHiTok still achieves competitive performance even with a smaller amount of data.

Model	Params	Type	#Training Images	Res.	GenEval	DPG
Only Gen.						
SD v2.1 (Rombach et al., 2022b)	–	Diff	2000M	1024	0.50	–
DALL-E 2 (Ramesh et al., 2022)	–	Diff	650M	1024	0.52	–
DALL-E 3 (Betker et al., 2023)	–	Diff	–	1024	0.67	83.50
PixArt-α (Chen et al., 2023)	0.6B	Diff	–	1024	0.48	–
SDXL (Podell et al., 2023)	2.6B	Diff	2000M	1024	0.55	74.65
Playgroundv2.5 (Li et al., 2024c)	–	Diff	–	1024	–	75.47
Und&Gen.						
Transfusion (Zhou et al., 2024)	7.3B	AR	3.5B	256	0.63	–
Ming-Lite-Uni (AI et al., 2025)	8B	AR-Scale	5M	512	0.62	–
Show-o (Xie et al., 2024)	1.5B	Diff	36M	256	0.53	67.27
Emu3 (Wang et al., 2024c)	8B	AR	–	256	0.66	80.60
Janus (Wu et al., 2024a)	1.3B	AR	65M	384	0.61	–
SynerGen-VL (Li et al., 2025)	2.4B	AR	25M	256	0.61	–
UniFork (Li et al., 2025)	0.76B	AR	–	384	0.46	–
ILLUME (Wang et al., 2024a)	7B	AR	15M	512	0.61	–
ILLUME+ (Huang et al., 2025a)	3B	AR	46M	512	0.72	–
MUSE-VL (Xie et al., 2025)	7B	AR	10M	256	0.57	–
QLIP-B (Zhao et al., 2025)	1.5B	AR	18M	256	0.48	78.17
TokenFlow (Qu et al., 2024)	7B	AR	60M	256	0.63	73.38
MMaDA (Yang et al., 2025)	8B	Diff	–	512	0.63	69.97
Liquid (Wu et al., 2024b)	7B	AR	30M	512	0.55	83.45
UniTok (Ma et al., 2025)	7B	AR	30M	256	0.59	83.45
Tar-1.5B (Han et al., 2025)	3B	AR	46M	256	0.76	82.96
Tar-1.5B W/Self-Reflect (Han et al., 2025)	3B	AR	46M	256	0.78	84.10
Tar-7B (Han et al., 2025)	7B	AR	46M	256	0.84	84.19
Tar-7B W/Self-Reflect (Han et al., 2025)	7B	AR	46M	256	0.85	84.65
Ming-UniVision (Huang et al., 2025b)	16B-A3B	AR-Continuous	–	512	0.85	82.12
SemHiTok(ours)	7B	AR	15M	256	0.71	83.59

controlling for the total codebook size ($K \times m$), we observe that prioritizing a larger semantic codebook K yields slightly better performance than increasing the sub-codebook size m.

To further demonstrate the effectiveness of our method, we conduct a more detailed comparison with other approaches that use large codebooks. As show in Tab.14, we conduct comparisons on the reconstruction task under two codebook sizes (65K and 262K). Under the 65K setting, our model surpasses the expert reconstruction model while maintaining good usage. Under the 262K setting, our method still achieves competitive performance compared to the expert model.

8.13 Specialized Image Tokenizer

Tokenization for Generation. Image tokenizers are crucial for autoregressive image generation(Van Den Oord et al., 2017; Rombach et al., 2022a; Tian et al., 2025). VQVAE(Van Den Oord et al., 2017) learns a discrete representation using a learnable codebook in auto-encoder architectures. VQGAN

Table 10: Impact of Concat type and sub-codebook size. w/o sem: not use the semantic discrete token. The gray bar represents the default setting in our experiments.

Concat	Subc Size	rFID ↓	Usage↑
w/o	12	1.99	95.4%
Len	12	1.45	94.1%
Dim	8	1.42	96.4%
Dim	12	1.26	93.7%
Dim	16	1.19	79.3%

Table 11: Impact of VQ Type and dim of the semantic codebook. We evaluate multimodal understanding performance under the LLaVA-v1.5 setting. The gray bar represents the default setting in our experiments.

VQ Type	Codebok Dim	MME-P↑	MMB↑	SEED↑	MMU↑
Norm	48	1249.6	52.2	52.8	34.8
Vanilla	48	1319.9	56.3	57.2	33.1
EMA	32	1385.9	58.3	61.2	35.1
EMA	48	1387.5	61.3	62.3	35.6
EMA	64	1428.7	60.9	62.5	35.5

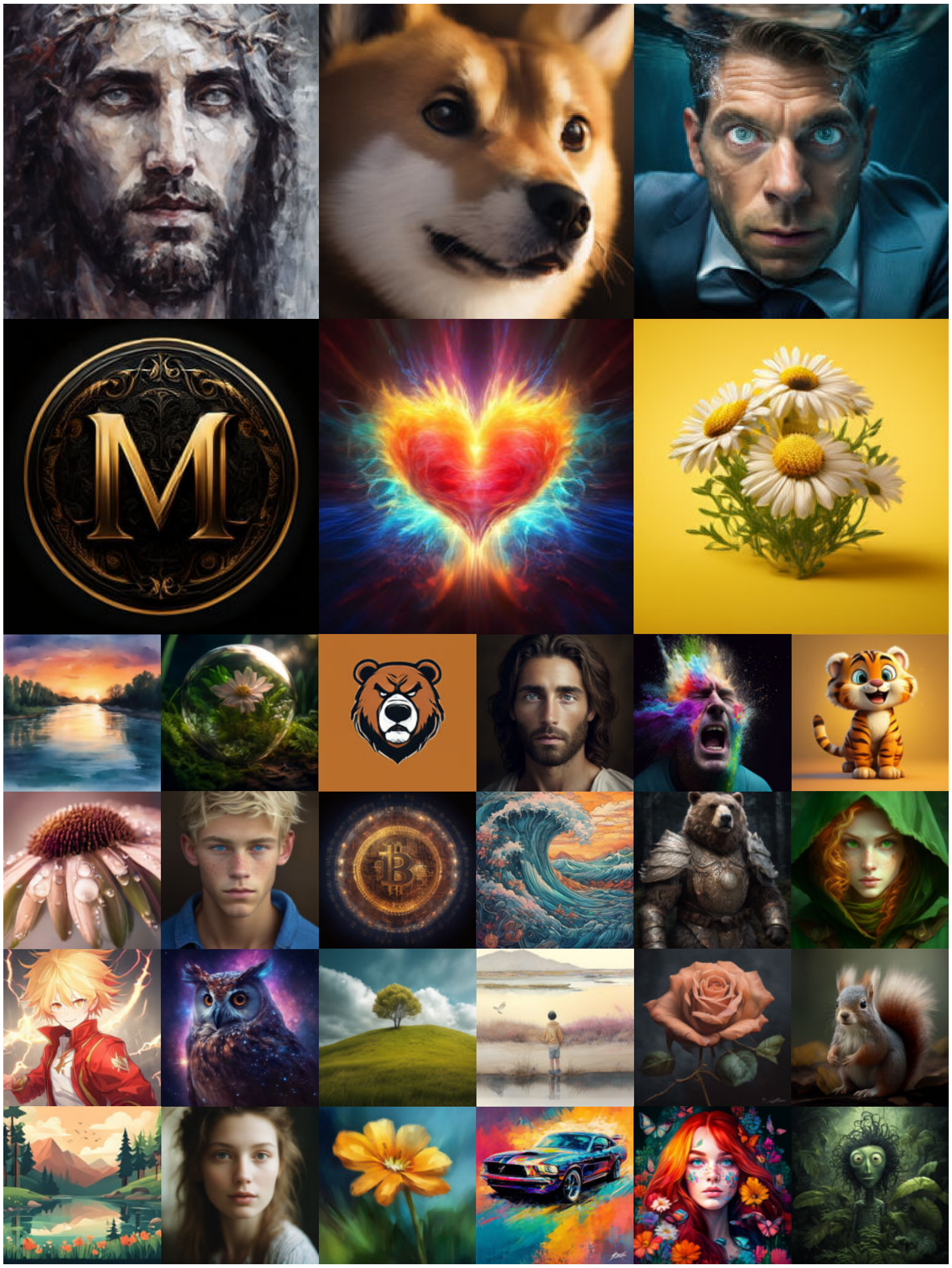


Figure 7: More generated images presentation.

further improves better perceptual quality by using adversarial training(Goodfellow et al., 2014). advanced the existing architecture by integrating perceptual loss and discriminator loss, alongside adversarial loss, to enhance reconstruction quality. This approach yields more precise and detailed image representations, significantly improving upon previous methodologies in image generation and processing. Subsequently, ViT-VQGAN(Yu et al., 2021b) and Efficient-VQGAN(Cao et al., 2023) advance the framework with the transformer design. In recent literature, researchers are turning to efficient codebook structures(Shi et al., 2024; Zhang et al., 2023; Yu et al., 2021b; Bai et al., 2024) and better quantization methods(Zha et al., 2024; Yu et al., 2025) to improve generation performance and compression rates. IBQ(Shi et al., 2024) proposes the Index Backpropagation Quantization

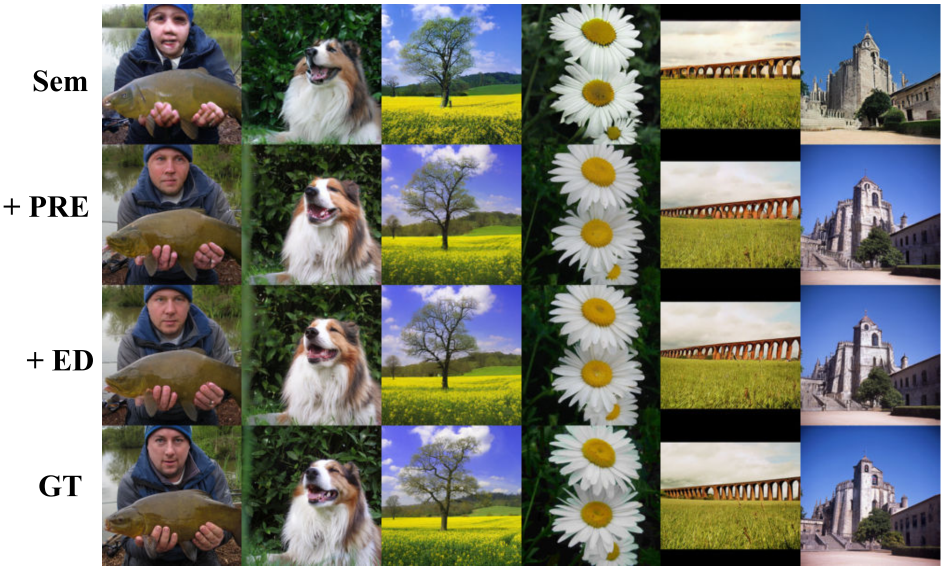


Figure 8: More visualized reconstruction results from the ablation of key module.

Table 12: Ablation of Semantic Codebook K. We evaluate on multimodal understanding tasks.

K	POPE	MME-P	SEED	GQA	Usage
4096	79.7	1297.7	60.4	56.6	100
8192	81.9	1343.0	61.7	59.7	99.7
16384	82.5	1355.8	62.9	60.3	99.0
32768	82.8	1364.2	62.5	60.6	93.0

Table 13: Comparative experiments with Different $K \times m$ on reconstruction task.

K	m	rFID ↓	Usage ↑
8192	12	1.58	95.5
	16	1.45	92.7
	24	1.31	89.8
	32	1.22	83.7
16384	8	1.42	96.4
	12	1.26	93.7
	16	1.19	79.3

codebook update method, achieving stable training of large-scale codebooks. Although these methods efficiently retain low-level texture information, they frequently fail to capture high-level semantic information, which limits their application to multimodal understanding tasks. FQGAN(Bai et al., 2024) uses multiple codebooks and product quantization, where each codebook encodes a different type of feature (Pixel, CLIP(Radford et al., 2021a), DINO(Caron et al., 2021), etc.). However, this work focuses only on image reconstruction and generation tasks, performing poorly on multimodal understanding tasks(Wu et al., 2024b; Liu et al., 2024a).

Tokenization for Understanding. In multimodal large language models (MLLMs)(Li et al., 2023b; 2022; Radford et al., 2021b; Liu et al., 2023a; Bai et al., 2023a; Chen et al., 2024c), researchers leverage CLIP(Radford et al., 2021b) and BLIP(Li et al., 2022) to extract visual characteristics that align with the language during its pre-training phase. Building upon many works (Liu et al., 2023a; Bai et al., 2023a; Chen et al., 2024c) have been collected and trained on high-quality datasets to achieve remarkable performance. LLaVA(Liu et al., 2023a) utilizes a vision encoder to align the vision inputs before LLMs. QwenVL(Bai et al., 2023a) and InterVL(Chen et al., 2024c) achieve better results through increased resolution, higher-quality datasets, etc. However, these text-aligned image encoders tend to focus on semantic information and ignore texture information, which is important for the generation task.

8.14 Unified Image Tokenizer

Numerous efforts have emerged to develop unified visual generation and understanding within one MLLM(Wang et al., 2024a;c; Xie et al., 2024; Dong et al., 2023; Ge et al., 2024; Sun et al., 2024b; Team, 2024; Wu et al., 2024d). There are two main lines to bridge the gap. Many workers (Dong et al., 2023; Ge et al., 2024; Sun et al., 2024b) combine diffusion models with LLMs for image

Table 14: Quantitative comparison of reconstruction quality under consistent total codebook sizes. SemHiTok outperforms existing quantization methods at the 65k scale and achieves comparable rFID to expert models at the 262k scale.

Method	Codebook Size	rFID↓	Usage↑
VQGAN-LC (Zhu et al., 2024a)	65546	2.63	100.0
VQGAN-LC(CLIP) (Zhu et al., 2024a)	65546	2.40	100.0
FSQ (Bai et al., 2024)	64000	2.80	100.0
LFQ (Yu et al., 2023)	65536	2.88	100.0
SimVQ (Zhu et al., 2025)	65536	2.24	100.0
SemHiTok	8192×8(65536)	1.93	98.1
	16384×4(65536)	1.84	99.0
SimVQ (Zhu et al., 2025)	262144	1.99	100.0
Open-MAGVIT2 (Luo et al., 2024)	262144	1.17	100.0
IBQ (Shi et al., 2024)	262144	1.00	79.3
SemHiTok	8192×32(262144)	1.21	83.7
	16384×8(262144)	1.19	79.3

generation. DREAMLLM(Dong et al., 2023) presents a unified framework that not only provides multimodal understanding but also creates multimodal content via diffusion models. Emu2(Sun et al., 2024b) trains a unified generative model using a diffusion-based decoder. These approaches inevitably increase model complexity and are not simple enough. Other workers (Team, 2024; Wu et al., 2024d; Xie et al., 2024; Wang et al., 2024c) adopt VQVAE-based encoders to convert images into discrete tokens. Chameleon(Team, 2024) and EMU3(Wang et al., 2024c) directly use VQGAN(Yu et al., 2021a), which is optimized by pixel reconstruction as the image tokenizer, while this method increases resource consumption during the pre-training stage and degrades multimodal understanding performance. VILA-U(Wu et al., 2024d) introduces a unified image tokenizer that incorporates a text-aligned branch within the VQGAN training paradigm. However, due to the gap between the semantic feature and texture feature, the joint optimization approach may lead to suboptimal solutions In contrast, our proposed SemHiTok can add the ability of extracting texture features without changing the discrete semantic features, and avoids the challenges brought by joint optimization.