MoETok: Sparse Expert Routing for 1D Visual Tokenizers

Anonymous authors
Paper under double-blind review

Abstract

Modern visual tokenizers aim to preserve fine-grained visual details while learning high-level semantic representations for downstream generation and understanding. Query-based 1D visual tokenizers encode images into flexible latent sequences, yet how individual learnable queries gather and encode image information remains poorly understood. We reveal that these queries spontaneously develop complementary roles under semantic supervision: a small set of Content queries captures high-level semantic information, while a larger set of Detail queries preserves fine-grained visual details for reconstruction. To adapt token processing to these heterogeneous roles, we introduce MoETok, a 1D visual tokenizer that routes tokens through sparsely activated experts. Jointly optimized with reconstruction and semantic objectives, the router learns role-specialized processing paths adapted to each token's representation. Across model scales and token budgets, MoETok consistently improves reconstruction, semantic representation, and autoregressive generation over dense tokenizers under matched supervision and active computation. On ImageNet, it combines strong reconstruction and semantic representation with state-of-the-art class-conditional generation within the 1D tokenizer family. We further validate MoETok at scale in a unified autoregressive model jointly trained for understanding and generation with a next-token cross-entropy objective. The resulting model combines competitive visual understanding with state-of-the-art generation among unified discrete models, matching or surpassing diffusion-based models such as BAGEL and Qwen-Image on benchmarks including DPG-Bench (87.61), GenEval2 (40.04 GM), and WISE (0.49).

1 Introduction

Modern visual tokenizers aim to preserve fine-grained visual details while learning high-level semantic representations for downstream generation and understanding. Early tokenizers were primarily trained for image reconstruction (van den Oord et al., 2017; Esser et al., 2021), while recent methods introduce semantic supervision to improve the learned latent space and support both generation and understanding (Yao et al., 2025; Chen et al., 2025a; Ma et al., 2025; Zhao et al., 2025; Yao et al., 2026). TiTok (Yu et al., 2024) introduced 1D visual tokenization, using learnable queries to encode an image into a latent sequence. Subsequent methods extend this formulation through diverse forms of semantic supervision and demonstrate improved representation quality and scaling behavior (Xiong et al., 2025; Beyer et al., 2025; Wen et al., 2025; Jia et al., 2025). These advances are largely evaluated through the performance of the complete token sequence, leaving a fundamental question: *how do individual learnable queries gather and encode image information?*

We investigate this question through encoder attention analysis (Doshi et al., 2026) and per-query linear probe evaluation (Beyer et al., 2025). Across 1D tokenizers with diverse architectures and training objectives, semantic supervision promotes spontaneous query-role differentiation: a small subset of queries adapts its attention to image content and contributes strongly to classification, while most queries attend to stable spatial regions. Reconstruction and subset interventions further establish their functional significance: the small Content subset retains high-level semantic information, whereas Detail queries preserve fine-grained visual details for faithful reconstruction. Together, these analyses reveal complementary query roles within the 1D bottleneck.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

These findings motivate allocating parameter capacity according to the different information roles of visual tokens. Sparse expert routing provides token-dependent access to a larger parameter pool while activating only a subset for each token. We investigate this design in 1D tokenizers under a fixed token budget and comparable active computation.

Specifically, we introduce **MoETok**, a 1D visual tokenizer with sparse expert routing. MoETok retains shared self-attention and replaces each dense encoder feed-forward network (FFN) with a pool of experts. A learned router selects a small subset of experts for each visual token based on its current representation. Trained jointly with reconstruction and semantic objectives, the router learns expert preferences for different token roles without explicit role labels. The resulting role-specialized processing paths allow MoETok to adapt how visual information is processed as the 1D bottleneck forms.

We evaluate MoETok through controlled comparisons, ImageNet benchmarks, and large-scale multimodal modeling. Across Small, Base, and Large architectures and both 64- and 256-token bottlenecks, MoETok consistently improves reconstruction, semantic representation, and autoregressive generation over dense tokenizers. On ImageNet, it combines strong reconstruction and semantic representation with state-of-the-art class-conditional generation within the 1D tokenizer family. We further use MoETok in a unified autoregressive model that performs visual understanding and generation through next-token prediction (Wang et al., 2024; Cui et al., 2025). The resulting unified autoregressive model delivers competitive visual understanding and generation performance comparable to or exceeding BAGEL and Qwen-Image across the evaluated benchmarks (Deng et al., 2025; Wu et al., 2025a). These results demonstrate the effectiveness of sparse expert routing for learning visual representations that support both understanding and generation.

Our contributions are summarized as follows:

- We systematically study how learnable queries gather and encode image information in 1D visual tokenizers. Cross-tokenizer analysis reveals emergent Content and Detail queries, whose complementary roles in semantic representation and fine-grained reconstruction are validated through linear probe evaluation and controlled interventions.
- We introduce MoETok, a 1D visual tokenizer that learns role-specialized sparse processing paths for heterogeneous visual tokens. It consistently improves reconstruction, semantic representation, and autoregressive generation across model scales and token budgets, achieving state-of-the-art class-conditional generation within the 1D tokenizer family on ImageNet.
- We further extend MoETok to unified autoregressive modeling, achieving competitive visual understanding and state-of-the-art generation among unified autoregressive models across DPG-Bench, WISE, and GenEval2.

2 Related Work

1D visual tokenizers. TiTok represents images as compact sequences of learnable queries, establishing an alternative to spatially arranged latent grids (Yu et al., 2024). Subsequent work has advanced adaptive token allocation, continuous representations, and hierarchical token organization for efficient reconstruction and generation (Yan et al., 2025; Duggal et al., 2025; Chen et al., 2025b; Huang et al., 2025; Zhang et al., 2026a). FlexTok uses nested dropout to learn ordered token sequences whose prefixes capture progressively finer image content (Bachmann et al., 2025). TA-TiTok incorporates CLIP text features into the decoder, introducing textual conditioning during tokenizer training (Kim et al., 2025). Semanticist combines continuous tokens with a PCA-like hierarchy and a diffusion decoder to organize complementary semantic and fine-grained visual information (Wen et al., 2025). GigaTok demonstrates the importance of semantic regularization for scaling 1D tokenizers (Xiong et al., 2025), while ReTok improves sequence utilization through redundant token padding and hierarchical semantic regularization (Fu et al., 2026). We investigate emergent query roles and introduce adaptive processing paths for them.

Visual representations for understanding and generation. Semantic learning improves visual tokenization for reconstruction, representation, and generation (Yao et al., 2025; Chen et al., 2025a; Ma et al., 2025; Zhao et al., 2025; Yao et al., 2026; Du et al., 2026). UniTok combines image–

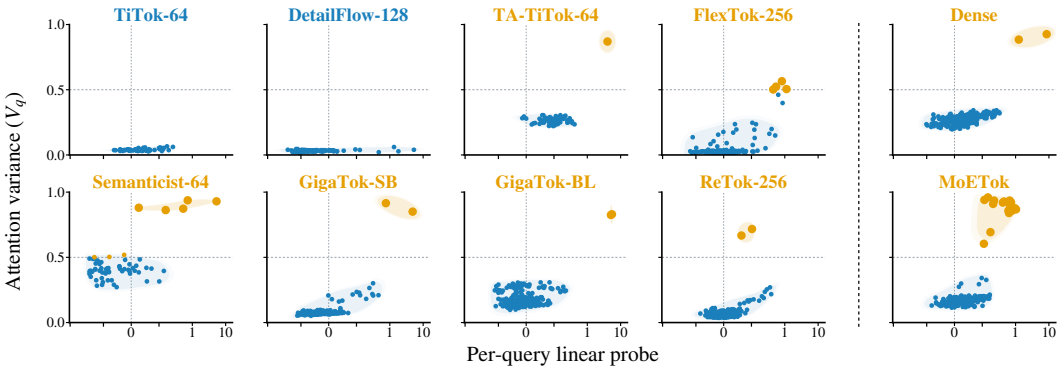

Figure 1: **Semantic supervision promotes query-role differentiation.** Each query is plotted by attention variance V_q and per-query linear probe score. **Blue** and **orange** indicate Detail and Content queries; model-name colors distinguish tokenizers without and with semantic supervision.

text learning and reconstruction with multi-codebook quantization (Ma et al., 2025); VTP studies joint representation learning and reconstruction for scalable pre-training (Yao et al., 2026); VQRAE quantizes pretrained visual representations for understanding and generation (Du et al., 2026). These representations also support unified autoregressive modeling (Wu et al., 2025b; Wang et al., 2026a). MoETok examines the tokenizer's representation space through token roles and conditional computation, linking heterogeneous information to role-specialized processing.

Sparse mixture of experts. Sparse mixture-of-experts models expand parameter capacity through input-dependent routing while activating only a subset of experts per token (Shazeer et al., 2017; Fedus et al., 2022), enabling efficient language-model scaling (Clark et al., 2022; Muennighoff et al., 2025). Vision research has also explored MoE for image classification (Riquelme et al., 2021; Puigcerver et al., 2024). Motivated by heterogeneous query roles, we introduce, to our knowledge, the first sparse mixture-of-experts architecture for unified visual tokenizers. It learns role-specialized processing paths for fine-grained visual details and high-level semantic information.

3 From Heterogeneous Visual Tokens to MoETok

3.1 Overview

Given an image x, a 1D visual tokenizer maps it to patch embeddings $P^0 = \{p_j^0\}_{j=1}^N$ and appends learnable queries $Q^0 = \{q_i^0\}_{i=1}^T$. A transformer encoder lets the queries gather information from the image patches. Its final query states are projected and quantized into a visual token sequence $z = \{z_i\}_{i=1}^T$, which supports image reconstruction and downstream autoregressive modeling. Each query can gather information from the full image, without being assigned to a fixed spatial cell.

MoETok introduces token-dependent sparse routing into this architecture. Both patch and query tokens draw from a shared expert pool within each encoder layer (Figure 3). We first analyze query roles, then introduce the resulting routing design.

3.2 Emergent Query Roles in Semantic 1D Tokenizers

Recent 1D tokenizers increasingly incorporate semantic supervision to improve their representations (Xiong et al., 2025; Wen et al., 2025; Fu et al., 2026). We examine how individual learnable queries gather and encode image information across representative tokenizers, and compare Dense and MoETok under matched reconstruction and feature regularization objectives.

Queries gather image information through attention to patch tokens. Inspired by recent analyses of attention patterns (Doshi et al., 2026; Marouani et al., 2026), we examine whether each query attends to stable spatial regions or changes its attention across images. Let $\mathbf{A}_{\ell,q}(x)$ denote the query-to-patch attention of query q at encoder layer ℓ on image x, normalized over patch positions,

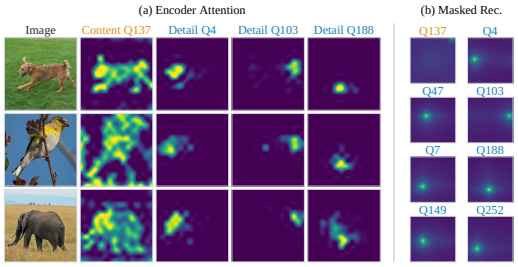

Figure 2: **Complementary query roles. (a)** **Content** attention follows objects; **Detail** attention stays localized. **(b)** Masking Detail queries affects reconstruction in matching regions.

Model	Subset	#	Linear probe ↑	rFID ↓
Dense	All	256	69.80	1.52
	Content	2	**66.97**$_{-2.82}$	**Failed**
	Detail	254	67.66	**1.56**$_{+0.04}$
MoETok	All	256	74.05	1.37
	Content	20	**76.01**$_{+1.96}$	**Failed**
	Detail	236	70.35	**2.08**$_{+0.71}$

Table 1: **Complementary subset functions.** **Content** subsets retain strong semantic information; **Detail** subsets preserve reconstruction. Content-only decoding fails (rFID > 300).

and let $\mathbf{U}$ denote uniform attention. We define the layer-by-query attention variance as

$$V_{\ell,q} = \frac{\mathrm{Var}_x[\mathbf{A}_{\ell,q}(x)]}{\mathrm{Var}_x[\mathbf{A}_{\ell,q}(x)] + \|\mathbb{E}_x[\mathbf{A}_{\ell,q}(x)] - \mathbf{U}\|_2^2}. \tag{1}$$

Here, $\mathrm{Var}_x[\mathbf{A}_{\ell,q}(x)]$ denotes the expected squared deviation from its cross-image mean. The entries form a layer-by-query matrix $\mathbf{V} = [V_{\ell,q}] \in [0,1]^{L\times T}$. Higher values indicate that cross-image variation dominates the query's stable spatial preference, whereas lower values indicate more consistent attention to similar image regions. For query-level analysis, we average across encoder depth, $V_q = L^{-1}\sum_{\ell=1}^{L} V_{\ell,q}$.

We relate these attention patterns to semantic information through each query's contribution to classification. Following recent work (Beyer et al., 2025), we train a linear probe on the complete bottleneck sequence, freeze it, and report the accuracy drop from masking each query as its *per-query linear probe* score.

Without semantic supervision, queries exhibit low attention variance, and most have small per-query classification effects (Figure 1). Across models incorporating semantic supervision, a small subset consistently exhibits both high attention variance and strong classification contributions, while most queries remain spatially stable. This recurring pattern indicates that semantic supervision promotes spontaneous query-role differentiation. We use the fixed midpoint $V_q > 0.5$ to identify Content queries and refer to the remaining group as Detail queries. We characterize their functional preferences through semantic probing and reconstruction interventions.

To assess the semantic information retained by each query group, we independently train a linear probe on its mean-pooled final encoder states. Small Content subsets retain strong semantic representations, while Detail subsets carry most of the information needed for reconstruction (Table 1). Encoder attention illustrates how Content queries follow foreground objects while Detail queries attend to stable spatial regions (Figure 2(a)). We then zero-mask each query and average the pixel-wise RGB L2 distance from the full-token reconstruction across 50,000 validation images. Detail queries affect reconstruction in regions matching their attention, whereas the illustrated Content query has little effect (Figure 2(b)). Together, these interventions validate the connection between attention patterns and complementary semantic and reconstructive functions. Cross-dataset validation and extended visualizations appear in Appendix B.1.

These findings motivate token-dependent parameter allocation: each token selects a subset of experts from a larger parameter pool, with active computation comparable to dense processing.

3.3 Sparse Expert Transformer Block

Let $X^\ell = [P^\ell; Q^\ell]$ denote the concatenated patch and query states entering a routed transformer block. The shared attention sublayer first produces

$$U^\ell = X^\ell + \mathrm{Attn}\big(\mathrm{LN}(X^\ell)\big)\,. \tag{2}$$

Each layer contains M experts, each implemented as a two-layer FFN with hidden width nD, where D is the encoder hidden dimension. For visual token u_i^ℓ, a lightweight router predicts a distribution

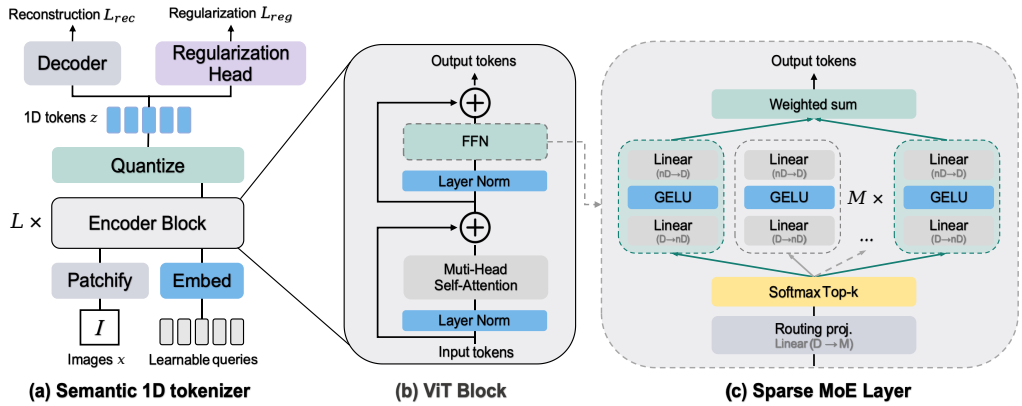

Figure 3: **MoETok adapts token processing through sparse expert routing.** **(a)** A 1D tokenizer learns with reconstruction and feature regularization. **(b)** A standard ViT block. **(c)** A sparse MoE layer replaces the shared FFN with token-dependent expert selection and weighted aggregation.

over experts,

$$\pi_i^\ell = \mathrm{softmax}\big(W_r^\ell \,\mathrm{LN}(u_i^\ell)\big)\,, \qquad \mathcal{S}_i^\ell = \mathrm{TopK}(\pi_i^\ell, k), \tag{3}$$

where $\mathcal{S}_i^\ell \subseteq \{1,\dots,M\}$ contains the k selected experts. After renormalizing the selected routing weights, the FFN update is

$$x_i^{\ell+1} = u_i^\ell + \sum_{e\in\mathcal{S}_i^\ell} \widetilde{\pi}_{i,e}^\ell \, E_e^\ell\big(\mathrm{LN}(u_i^\ell)\big)\,. \tag{4}$$

MoETok replaces the dense FFN in every encoder layer with a sparse MoE layer. Expert assignments are learned from the task objectives, with patch and query tokens sharing a layer-local expert pool. Only the final query states are projected and quantized.

3.4 Learning Objectives

MoETok is trained with reconstruction, vector-quantization, and visual feature regularization objectives (Xiong et al., 2025):

$$\mathcal{L} = \lambda_{\mathrm{rec}}\mathcal{L}_{\mathrm{rec}} + \lambda_{\mathrm{vq}}\mathcal{L}_{\mathrm{vq}} + \lambda_{\mathrm{reg}}\mathcal{L}_{\mathrm{reg}}. \tag{5}$$

We follow GigaTok for the reconstruction objective (Xiong et al., 2025). Feature regularization supervises features decoded from the quantized query sequence, with frozen SigLIP2-L features as the default target (Tschannen et al., 2025). We use StableVQ for quantization (Tang et al., 2026).

3.5 Analyzing Learned Conditional Routing

We analyze Top-1 expert assignments (Muennighoff et al., 2025) to characterize expert utilization and token-role specialization. The entropy-equivalent effective expert count $N_{\mathrm{eff},\ell} \in [1, M]$ ranges from complete concentration to uniform expert use. The normalized Jensen–Shannon divergence between role-conditioned routing distributions measures how distinctly Patch, Detail, and Content tokens select experts. Experts become broadly active from the middle encoder onward, while role specialization peaks at layer 5 and rises again at the output (Figure 4(a, b)). At the peak, distinct experts preferentially process Patch, Detail, and Content tokens (Figure 4(c)), revealing learned role-specialized routing. This depth-dependent utilization persists when the expert pool expands. In 1D tokenizers, loss-free balancing (DeepSeek-AI, 2024) makes expert use more uniform but degrades reconstruction, semantic representation, and generation, supporting task-driven expert allocation across layers. Further comparisons and metric details appear in Appendix B.2.

Compared with Dense, MoETok also expands the Content subset while preserving the reconstructive function of Detail queries (Table 1). Averaging the attention-variance matrix across queries gives the layer-wise statistic $V_\ell = T^{-1}\sum_{q=1}^{T} V_{\ell,q}$ (Figure 4(d)). Attention variance in the Dense

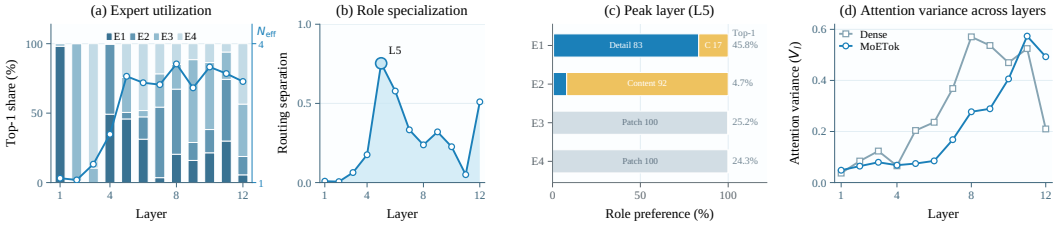

Figure 4: **Conditional routing learns role-specialized visual processing.** Expert use broadens mid-encoder **(a)**; role separation peaks mid-encoder and remains high at the output **(b)**. Experts show distinct role preferences **(c)**; MoETok retains higher attention variance at the bottleneck **(d)**.

model drops sharply at the final encoder layer, whereas MoETok preserves image-adaptive query attention near the discrete bottleneck. Together, these observations link role-specialized routing with image-adaptive attention near the bottleneck, providing an interpretation of the representation gains observed in our controlled comparisons.

4 Experiments

We first study tokenization structure, performance–compute scaling, and expert design, and then evaluate MoETok on ImageNet and in unified autoregressive modeling.

4.1 Experimental Setup

We train tokenizers on ImageNet (Deng et al., 2009) at 256×256 resolution. Unless otherwise specified, analyses and controlled studies use 256 queries and four experts with two activated per token. Visual feature regularization uses a frozen SigLIP2-L teacher (Tschannen et al., 2025). Dense baselines share the same backbone, supervision, and training settings. Tokenizer parameters are reported as active/total (A/T). Detailed architectures and training hyperparameters are provided in Appendix A.1.

We evaluate reconstruction, semantic representation, and downstream generation using rFID, linear probe, zero-shot Top-1 accuracy, and gFID, with a consistent evaluation protocol within each sweep. We train a linear classifier on flattened quantized tokens with the tokenizer frozen. Zero-shot classification compares the tokenizer's aligned visual features with class-name embeddings from the frozen SigLIP2 text encoder. AR models use LlamaGen with 1D RoPE and its B/XL configurations (Sun et al., 2024). Guided ImageNet results use the CFG scale minimizing gFID. Text-conditioned generation is evaluated on GenEval, DPG-Bench, WISE, and GenEval2 (Ghosh et al., 2023; Hu et al., 2024; Niu et al., 2026; Kamath et al., 2025). Additional protocols and controlled results appear in Appendices A and C.

4.2 Tokenization Structure and Expert Routing

Table 2: **1D tokenization amplifies the semantic benefits of expert routing.** Each pair matches active FFN capacity; gFID is evaluated without guidance.

Method	Routed tokens	rFID ↓	Linear probe ↑	Zero-shot ↑	gFID ↓
Dense	–	1.566	41.83	61.52	12.18
MoETok	Patch	**1.494**	**44.26**	**63.16**	**11.59**
Dense	–	1.522	48.72	59.17	10.35
MoETok	Patch + Query	**1.368**	**60.22**	**65.03**	**9.59**

We compare sparse routing in patch-based 2D tokenizers and query-based 1D tokenizers, with each dense–MoE pair matched in active FFN capacity, objectives, and training schedule (Table 2). Sparse routing improves reconstruction, linear separability, zero-shot Top-1 accuracy, and generation in

both structures. Its benefits are substantially greater for linear separability and zero-shot Top-1 accuracy in the query-based tokenizer, accompanied by a larger reconstruction gain.

Both Dense and MoETok yield lower gFID with query-based than patch-based tokenization, echoing GigaTok's findings on query-based generation (Xiong et al., 2025). The larger semantic gains in query-based tokenizers establish 1D tokenization as a particularly effective setting for sparse expert routing, consistent with the heterogeneous query roles identified in our analysis.

4.3 Performance–Compute Scaling

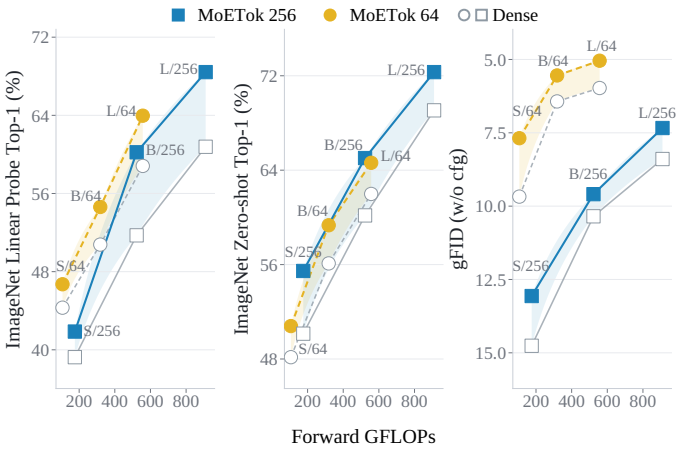

Figure 5: **Performance–compute scaling.** MoETok achieves stronger semantic representation and unguided generation than Dense at comparable forward computation across Small, Base, and Large models at both 64 and 256 tokens.

Across Small, Base, and Large architectures, MoETok achieves stronger linear probe and zero-shot Top-1 accuracy than Dense at both 64 and 256 tokens (Figure 5). Our scale variants jointly increase encoder and decoder capacity, while every Dense–MoETok pair shares the same supervision and quantizer. The advantage persists under the fourfold reduction in token count, showing that sparse routing benefits 1D representations at different compression levels. Unguided gFID also improves as model capacity increases, extending the scaling benefits from semantic representation to autoregressive generation.

4.4 Compact Design Studies

We examine routing depth and expert configuration in 1D tokenizers.

Full-depth routing provides the strongest semantic representation with comparable reconstruction quality and is used by default (Table 3). We vary expert width, expert count, and Top-k to examine per-expert capacity, the available expert pool, and active computation in 1D tokenization (Table 4).

Table 3: **Encoder depth.** Full-depth routing yields the strongest semantic representation while maintaining comparable reconstruction quality.

MoETok layers	rFID ↓	Linear probe ↑	Zero-shot ↑
Early 4	1.322	46.85	59.50
Late 4	1.378	51.21	61.46
Late 8	**1.309**	51.31	62.30
All	1.368	**60.22**	**65.03**

Table 4: **Expert design for 1D tokenizers.** We vary expert width, expert count, and Top-k. E8/T2 balances representation, generation, and computational cost.

Experts / Top-k	Width	Total / Active	Linear probe ↑	Zero-shot ↑	gFID ↓
4 / 2	$2D$	$8D/4D$	54.01	61.82	10.47
4 / 2	$4D$	$16D/8D$	**60.22**	**65.03**	**9.59**
4 / 2	$4D$	$16D/8D$	60.22	65.03	9.59
8 / 2	$4D$	$32D/8D$	60.63	66.37	**8.99**
16 / 2	$4D$	$64D/8D$	**62.92**	**67.02**	9.14
8 / 1	$4D$	$32D/4D$	52.02	56.65	10.56
8 / 2	$4D$	$32D/8D$	60.63	66.37	8.99
8 / 8	$4D$	$32D/32D$	**63.96**	**67.62**	**8.66**

Routing the late four layers yields stronger semantic representation than routing the early four, consistent with mid-encoder and output-layer specialization (Figure 4); full-depth routing improves it further. In 1D tokenizers, wider experts improve all three metrics, and expanding the expert pool improves semantic representation at fixed active computation. Eight experts achieve the lowest gFID in this sweep. Increasing Top-k from one to two yields substantial gains; activating all eight provides smaller additional improvements at four times the active expert capacity. Top-2 thus captures most of the gains from full activation. We adopt E8/T2 to balance representation, generation, and computational cost. Further comparisons of routing in the encoder and decoder, along with studies of expert granularity, latent dimension, and codebook size, appear in Appendix C.

Table 5: **ImageNet reconstruction, recognition, and generation.** MoETok combines strong semantic representations with high-quality reconstruction and generation in a single 1D tokenizer.

Method	Tokens	Tok. Params A/T	Tokenizer		Generator		w/o CFG	w/ CFG
			rFID ↓	Zero-shot ↑	Params	Epochs	gFID ↓	gFID ↓
Perception and Generation Baselines								
SigLIP2-L (Tschannen et al., 2025)	256	882M	–	82.5	–	–	–	–
DiT (Peebles & Xie, 2023)	–	–	–	–	675M	1400	9.62	2.27
REPA (Yu et al., 2025)	–	–	0.61	–	675M	800	5.78	1.29
DDT (Wang et al., 2026b)	–	–	0.61	–	675M	400	6.27	1.26
RAE (SigLIP2-B) (Zheng et al., 2026)	256	509M	0.82	–	839M	80	4.93	–
FVQ-XL (Chang et al., 2026)	256	–	–	–	775M	–	–	2.07
VILA-U (Wu et al., 2025b)	256×4	–	1.80	73.3	–	–	–	–
UniTok (Ma et al., 2025)	256×8	–	0.38	78.4	1.4B	–	2.51	2.77
QLIP-B (Zhao et al., 2025)	256	175M	3.21	74.3	–	–	–	–
1D Visual Tokenizers								
TiTok-L (Yu et al., 2024)	32	641M	2.21	–	177M	–	3.15	2.77
Semanticist (Wen et al., 2025)	32	768M	2.63	–	493M	–	–	2.57
FlexTok d18–d18 (Bachmann et al., 2025)	32	950M	1.61	–	1.33B	–	–	2.00
DetailFlow-16 (Liu et al., 2025)	128	271M	1.22	–	326M	300	–	2.96
VFMToK (Zheng et al., 2025)	256	–	0.89	–	343M	300	2.11	2.75
ReTok-S-B (Fu et al., 2026)	256	232M	1.01	–	775M	300	–	2.27
GigaTok-B-L (Xiong et al., 2025)	256	622M	0.81	–	111M	300	–	3.26
GigaTok-XL-XXL (Xiong et al., 2025)	256	2.9B	0.79	–	1.4B	300	1.98	–
Dense (GigaTok)	256	505M	0.82	78.5	111M	300	5.06	2.37
MoETok	256	505M/845M	**0.76**	**80.4**	111M	300	4.44	1.99
					775M	300	**1.61**	**1.39**

4.5 ImageNet Reconstruction, Representation, and Generation

Under the same visual feature regularization, MoETok improves reconstruction, semantic representation, and generation over the matched Dense tokenizer (Table 5). These joint gains demonstrate that sparse expert routing strengthens the learned 1D representation across all three tasks. Stronger semantic representation accompanies improved reconstruction, consistent with the complementary Content and Detail roles identified by our token-level interventions.

Within the 1D tokenizer family, MoETok combines strong reconstruction with state-of-the-art class-conditional generation and the strongest reported zero-shot Top-1 accuracy among the evaluated methods. Its visual tokens support progressively better generation with B and XL LlamaGen priors trained for 300 epochs. With the tokenizer held fixed, this trend complements tokenizer scaling: a larger prior makes more effective use of the same learned visual representation. The following evaluation extends this setting from class-conditional generation to text-conditioned generation and multimodal understanding in a unified autoregressive model.

4.6 Unified Understanding and Generation

We initialize the unified autoregressive model from Qwen3.5-4B (Qwen Team, 2026) and extend its vocabulary with MoETok visual tokens. Following discrete multimodal modeling (Wang et al., 2024), we train the model with a next-token cross-entropy objective over text and visual tokens. Training details are provided in Appendix A.2. We evaluate seven standard understanding benchmarks (Li et al., 2023; Liu et al., 2024; Fu et al., 2025; Yue et al., 2024; Hudson & Manning, 2019; Goyal et al., 2017; Li et al., 2024) and four generation suites.

The 4B MoETok-based model is competitive with larger discrete models in understanding. It leads evaluated autoregressive baselines on overall GenEval and DPG-Bench and all six WISE domains, and evaluated open models on GenEval2 GM (Tables 6 to 9). These results demonstrate the versatility of MoETok's 1D representation for unified understanding and generation.

Table 6: **Multimodal understanding benchmarks.** Unified indicates joint training for understanding and generation; #Params reports the number of parameters in the language model backbone.

Model	Unified	#Params	POPE	MMB	MME_{Perc}	MMMU	GQA	VQAv2	SEED
Continuous visual representations									
Qwen3.5-4B (Qwen Team, 2026)	×	4B	–	89.4	–	77.6	–	–	–
Janus-Pro (Chen et al., 2025e)	✓	7B	87.4	79.2	1567	41.0	62.0	–	72.1
Show-o2 (Xie et al., 2025b)	✓	7B	–	79.3	1621	48.9	63.1	–	69.8
BAGEL (Deng et al., 2025)	✓	7B	–	85.0	1687	55.3	–	–	–
Discrete visual representations									
Show-o (Xie et al., 2025a)	✓	1.3B	80.0	–	1097	26.7	58.0	69.4	–
Liquid (Wu et al., 2026)	✓	7B	83.2	–	1448	–	61.1	76.8	–
VILA-U (Wu et al., 2025b)	✓	7B	85.8	–	1402	–	60.8	79.4	59.0
UniTok (Ma et al., 2025)	✓	7B	83.2	–	1448	–	61.1	76.8	–
Emu3 (Wang et al., 2024)	✓	8B	85.2	58.5	–	31.6	60.3	75.1	68.2
ARM (Wang et al., 2026a)	✓	7B	87.3	80.7	1463	40.2	59.8	76.1	73.1
MoETok	✓	4B	85.5	75.6	1525	39.7	60.0	76.7	74.2

Table 7: **Text-to-image generation on GenEval and DPG-Bench.**

Method	GenEval ↑				DPG-Bench ↑		
	Two Obj.	Position	Color Attri.	Overall	Global	Relation	Overall
Diffusion-based models							
FLUX.1-dev (Black Forest Labs, 2024)	0.81	0.22	0.45	0.66	74.35	90.87	83.84
Show-o (Xie et al., 2025a)	0.80	0.31	0.50	0.68	–	–	67.48
Lumina-Image 2.0 (Qin et al., 2025)	0.87	–	0.62	0.73	–	94.85	87.20
BAGEL (Deng et al., 2025)	0.94	0.64	0.63	0.82	–	–	–
Qwen-Image (Wu et al., 2025a)	0.92	0.76	0.77	0.87	91.32	94.31	88.32
Autoregressive models							
Emu3 (Wang et al., 2024)	0.71	0.17	0.21	0.54	85.21	90.22	80.60
Infinity (Han et al., 2025)	0.85	0.49	0.57	0.73	93.11	90.76	83.46
Janus-Pro-7B (Chen et al., 2025e)	0.89	0.79	0.66	0.80	86.90	89.32	84.19
ARM (Wang et al., 2026a)	0.91	0.75	0.60	0.79	89.85	92.00	84.48
MoETok	0.96	0.84	0.74	0.86	85.50	92.47	87.61

Table 8: **WISE.** World-knowledge-intensive generation across six knowledge domains.

Type	Method	Cult.	Time	Space	Bio.	Phys.	Chem.	Avg.
Diff.	Qwen-Image	.63	.53	.56	.34	.48	.25	.51
	Z-Image	.55	.47	.51	.33	.48	.18	.45
	BAGEL	.41	.35	.31	.20	.44	.26	.35
	FLUX.1-dev	.52	.40	.53	.18	.38	.24	.42
AR	Janus-Pro-7B	.37	.35	.28	.28	.40	.23	.33
	NextStep-1	.49	.38	.31	.23	.47	.25	.40
	MoETok	.57	.45	.47	.38	.53	.38	.49

Table 9: **GenEval2.** Compositional generation benchmark.

Method	AM ↑	GM ↑
Gemini (closed)	82.80	44.60
SDXL	50.10	9.10
SD3.5-large	69.70	22.80
FLUX.1-dev	67.10	21.10
Qwen-Image	80.80	33.80
BAGEL + CoT	70.90	23.10
MoETok	77.69	40.04

5 Conclusion

Learnable queries in 1D visual tokenizers develop complementary roles under semantic supervision: Content queries capture high-level semantics, while Detail queries preserve fine-grained visual details. MoETok adapts processing to these roles through sparse expert routing, improving reconstruction, semantic representation, and autoregressive generation across model scales and token budgets. It achieves state-of-the-art class-conditional generation within the 1D tokenizer family and supports competitive understanding and leading generation in unified autoregressive modeling. These results establish token-dependent processing as an effective approach to unified visual representation learning, matching query processing to the needs of reconstruction, understanding, and generation.

AI USE STATEMENT

We used AI tools for translation, language polishing, and grammar correction. The authors reviewed and take responsibility for all content.

REFERENCES

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, Chunsheng Wu, Huajie Tan, Chunyuan Li, Jing Yang, Jie Yu, Xiyao Wang, Bin Qin, Yumeng Wang, Zizhen Yan, Ziyong Feng, Ziwei Liu, Bo Li, and Jiankang Deng. LLaVA-OneVision-1.5: Fully open framework for democratized multimodal training, 2025. URL https://arxiv.org/abs/2509.23661.

Roman Bachmann, Jesse Allardice, David Mizrahi, Enrico Fini, Oguzhan Fatih Kar, Elmira Amirloo, Alaaeldin El-Nouby, Amir Zamir, and Afshin Dehghan. Flextok: Resampling images into 1d token sequences of flexible length. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267, 2025.

Lukas Lao Beyer, Tianhong Li, Xinlei Chen, Sertac Karaman, and Kaiming He. Highly compressed tokenizer can generate without training. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267, 2025.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

Yifan Chang, Jie Qin, Limeng Qiao, Xiaofeng Wang, Zheng Zhu, Lin Ma, and Xingang Wang. Scalable training for vector-quantized networks with 100% codebook utilization. In *International Conference on Learning Representations*, 2026.

Hao Chen, Yujin Han, Fangyi Chen, Xiang Li, Yidong Wang, Jindong Wang, Ze Wang, Zicheng Liu, Difan Zou, and Bhiksha Raj. Masked autoencoders are effective tokenizers for diffusion models. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267, 2025a.

Hao Chen, Ze Wang, Xiang Li, Ximeng Sun, Fangyi Chen, Jiang Liu, Jindong Wang, Bhiksha Raj, Zicheng Liu, and Emad Barsoum. Softvq-vae: Efficient 1-dimensional continuous tokenizer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 28358–28370. Computer Vision Foundation / IEEE, 2025b.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. BLIP3-o: A family of fully open unified multimodal models—architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025c.

Junying Chen, Zhenyang Cai, Pengcheng Chen, Shunian Chen, Ke Ji, Xidong Wang, Yunjin Yang, and Benyou Wang. ShareGPT-4o-Image: Aligning multimodal models with gpt-4o-level image generation, 2025d. URL https://arxiv.org/abs/2506.18095.

Sixiang Chen, Jianyu Lai, Jialin Gao, Tian Ye, Haoyu Chen, Hengyu Shi, Shitong Shao, Yunlong Lin, Song Fei, Zhaohu Xing, Yeying Jin, Junfeng Luo, Xiaoming Wei, and Lei Zhu. Postercraft: Rethinking high-quality aesthetic poster generation in a unified framework. In *International Conference on Learning Representations*, volume 2026, pp. 64080–64122, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/68d88dcd1e1917c74993902073f08e40-Paper-Conference.pdf.

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*, 2025e.

Aidan Clark, Diego de Las Casas, Aurelia Guy, Arthur Mensch, Michela Paganini, Jordan Hoffmann, Bogdan Damoc, Blake A. Hechtman, Trevor Cai, Sebastian Borgeaud, George van den Driessche, Eliza Rutherford, Tom Hennigan, Matthew J. Johnson, Albin Cassirer, Chris Jones, Elena Buchatskaya, David Budden, Laurent Sifre, Simon Osindero, Oriol Vinyals, Marc'Aurelio Ranzato, Jack W. Rae, Erich Elsen, Koray Kavukcuoglu, and Karen Simonyan. Unified scaling

laws for routed language models. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 4057–4086. PMLR, 2022.

Yufeng Cui, Honghao Chen, Haoge Deng, Xu Huang, Xinghang Li, Jirong Liu, Yang Liu, Zhuoyan Luo, Jinsheng Wang, Wenxuan Wang, Yueze Wang, Chengyuan Wang, Fan Zhang, Yingli Zhao, Ting Pan, Xianduo Li, Zecheng Hao, Wenxuan Ma, Zhuo Chen, Yulong Ao, Tiejun Huang, Zhongyuan Wang, and Xinlong Wang. Emu3.5: Native multimodal models are world learners. *arXiv preprint arXiv:2510.26583*, 2025.

DeepSeek-AI. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

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 Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pp. 248–255. IEEE Computer Society, 2009.

Fenil R. Doshi, Thomas Fel, Talia Konkle, and George Alvarez. Bi-orthogonal factor decomposition for vision transformers. *arXiv preprint arXiv:2601.05328*, 2026.

Sinan Du, Jiahao Guo, Bo Li, Shuhao Cui, Zhengzhuo Xu, Yifu Luo, Yongxian Wei, Kun Gai, Xinggang Wang, Kai Wu, and Chun Yuan. VQRAE: representation quantization autoencoders for multimodal understanding, generation and reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2026, Denver, CO, USA, June 3-7, 2026*, pp. 30322–30334. Computer Vision Foundation, 2026.

Shivam Duggal, Phillip Isola, Antonio Torralba, and William T. Freeman. Adaptive length image tokenization via recurrent allocation. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 12873–12883, 2021.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *J. Mach. Learn. Res.*, 23:120:1–120:39, 2022.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. MME: A comprehensive evaluation benchmark for multimodal large language models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025.

Zixuan Fu, Lanqing Guo, Chong Wang, Binbin Song, Ding Liu, and Bihan Wen. Improving flexible image tokenizers for autoregressive image generation. *arXiv preprint arXiv:2601.01535*, 2026.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah M. Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander J. Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. Datacomp: In search of the next generation of multimodal datasets. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 6325–6334. IEEE Computer Society, 2017.

Jiawei Guo, Tianyu Zheng, Yizhi Li, Yuelin Bai, Bo Li, Yubo Wang, King Zhu, Graham Neubig, Wenhu Chen, and Xiang Yue. Mammoth-vl: Eliciting multimodal reasoning with instruction tuning at scale. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 13869–13920. Association for Computational Linguistics, 2025.

Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 15733–15744. Computer Vision Foundation / IEEE, 2025.

M. O. Hill. Diversity and evenness: A unifying notation and its consequences. *Ecology*, 54(2): 427–432, 1973. doi: 10.2307/1934352.

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.

Yuanhui Huang, Weiliang Chen, Wenzhao Zheng, Yueqi Duan, Jie Zhou, and Jiwen Lu. Spectralar: Spectral autoregressive visual generation. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pp. 15842–15852. IEEE, 2025.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 6700–6709. Computer Vision Foundation / IEEE, 2019.

Mingkai Jia, Mingxiao Li, Liaoyuan Fan, Tianxing Shi, Jiaxin Guo, Zeming Li, Xiaoyang Guo, Xiao-Xiao Long, Qian Zhang, Ping Tan, and Wei Yin. Dino-tok: Adapting DINO for visual tokenizers. *arXiv preprint arXiv:2511.20565*, 2025. URL https://arxiv.org/abs/2511.20565v1.

Amita Kamath, Kai-Wei Chang, Ranjay Krishna, Luke Zettlemoyer, Yushi Hu, and Marjan Ghazvininejad. GenEval 2: Addressing benchmark drift in text-to-image evaluation. *arXiv preprint arXiv:2512.16853*, 2025.

Dongwon Kim, Ju He, Qihang Yu, Chenglin Yang, Xiaohui Shen, Suha Kwak, and Liang-Chieh Chen. Democratizing text-to-image masked generative models with compact text-aware one-dimensional tokens. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pp. 18442–18452. IEEE, 2025.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 13299–13308. IEEE, 2024.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 292–305. Association for Computational Linguistics, 2023.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Trans. Inf. Theory*, 37(1): 145–151, 1991. doi: 10.1109/18.61115. URL https://doi.org/10.1109/18.61115.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pp. 740–755. Springer, 2014.

Yiheng Liu, Liao Qu, Huichao Zhang, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Xian Li, Shuai Wang, Daniel K. Du, et al. Detailflow: 1d coarse-to-fine autoregressive image generation via next-detail prediction. *arXiv preprint arXiv:2505.21473*, 2025.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part VI*, volume 15064 of *Lecture Notes in Computer Science*, pp. 216–233. Springer, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

Chuofan Ma, Yi Jiang, Junfeng Wu, Jihan Yang, Xin Yu, Zehuan Yuan, Bingyue Peng, and Xiaojuan Qi. Unitok: a unified tokenizer for visual generation and understanding. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025.

Xu Ma, Yitian Zhang, Qihua Dong, and Yun Fu. Fine-T2I: An open, large-scale, and diverse dataset for high-quality t2i fine-tuning, 2026. URL https://arxiv.org/abs/2602.09439.

Alexis Marouani, Oriane Siméoni, Hervé Jégou, Piotr Bojanowski, and Huy V. Vo. Revisiting [CLS] and patch token interaction in vision transformers. In *International Conference on Learning Representations*, 2026.

Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Evan Pete Walsh, Oyvind Tafjord, Nathan Lambert, Yuling Gu, Shane Arora, Akshita Bhagia, Dustin Schwenk, David Wadden, Alexander Wettig, Binyuan Hui, Tim Dettmers, Douwe Kiela, Ali Farhadi, and et al. Olmoe: Open mixture-of-experts language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Chaoran Feng, Fanqing Meng, Kun-Peng Ning, Bin Zhu, and Li Yuan. WISE: World knowledge-informed semantic evaluation for text-to-image generation. In *Forty-third International Conference on Machine Learning*, 2026.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 4172–4182. IEEE, 2023.

Joan Puigcerver, Carlos Riquelme Ruiz, Basil Mustafa, and Neil Houlsby. From sparse to soft mixtures of experts. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

Qi Qin, Le Zhuo, Yi Xin, Ruoyi Du, Zhen Li, Bin Fu, Yiting Lu, Xinyue Li, Dongyang Liu, Xiangyang Zhu, Will Beddow, Erwann Millon, Victor Perez, Wenhai Wang, Yu Qiao, Bo Zhang, Xiaohong Liu, Hongsheng Li, Chang Xu, and Peng Gao. Lumina-image 2.0: a unified and efficient image generative framework. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pp. 20031–20042. IEEE, 2025.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 8583–8595, 2021.

Juan A. Rodríguez, Xiangru Jian, Siba Smarak Panigrahi, Tianyu Zhang, Aarash Feizi, Abhay Puri, Akshay Kalkunte Suresh, François Savard, Ahmed Masry, Shravan Nayak, Rabiul Awal, Mahsa Massoud, Amirhossein Abaskohi, Zichao Li, Suyuchen Wang, Pierre-André Noël, Mats Leon Richter, Saverio Vadacchino, Shubham Agarwal, Sanket Biswas, and et al. Bigdocs: An open dataset for training multimodal models on document and code tasks. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: an open large-scale dataset for training next generation image-text models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, Jifeng Dai, Yu Qiao, Limin Wang, and Hongsheng Li. Journeydb: A benchmark for generative image understanding. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

Bao Tang, Jiahao Guo, Haoxiang Cao, Wenyu Liu, Changqian Yu, Kun Gai, and Xinggang Wang. StableVQ: Practical guidelines for stable vector-quantized tokenizer training, 2026. URL https://arxiv.org/abs/2609.26774.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 6306–6315, 2017.

Alex Jinpeng Wang, Dongxing Mao, Jiawei Zhang, Weiming Han, Zhuobai Dong, Linjie Li, Yiqi Lin, Zhengyuan Yang, Libo Qin, Fuwei Zhang, Lijuan Wang, and Min Li. TextAtlas5M: A large-scale dataset for dense text image generation. *arXiv preprint arXiv:2502.07870*, 2025. URL https://arxiv.org/abs/2502.07870.

Junke Wang, Xiao Wang, Jiacheng Pan, Xuefeng Hu, Feng Li, Jingxiang Sun, Chaorui Deng, Zilong Chen, Yunpeng Chen, Kaibin Tian, Matthew Gwilliam, Hao Chen, Danhui Guan, Kun Xu, Weilin Huang, Zuxuan Wu, Haoqi Fan, Yu-Gang Jiang, and Zhenheng Yang. ARM: An autoregressive large multimodal model with unified discrete representations. *arXiv preprint arXiv:2606.11188*, 2026a.

Shuai Wang, Zhi Tian, Weilin Huang, and Limin Wang. DDT: decoupled diffusion transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2026, Denver, CO, USA, June 3-7, 2026*, pp. 40633–40642. Computer Vision Foundation, 2026b.

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

Xin Wen, Bingchen Zhao, Ismail Elezi, Jiankang Deng, and Xiaojuan Qi. "principal components" enable a new language of images. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pp. 16641–16651. IEEE, 2025.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025a.

Junfeng Wu, Yi Jiang, Chuofan Ma, Yuliang Liu, Hengshuang Zhao, Zehuan Yuan, Song Bai, and Xiang Bai. Liquid: Language models are scalable and unified multi-modal generators. *Int. J. Comput. Vis.*, 134(1):39, 2026.

Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, Song Han, and Yao Lu. VILA-U: a unified foundation model integrating visual understanding and generation. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025b.

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. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025a.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025b.

Tianwei Xiong, Jun Hao Liew, Zilong Huang, Jiashi Feng, and Xihui Liu. Gigatok: Scaling visual tokenizers to 3 billion parameters for autoregressive image generation. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pp. 18770–18780. IEEE, 2025.

Wilson Yan, Volodymyr Mnih, Aleksandra Faust, Matei Zaharia, Pieter Abbeel, and Hao Liu. Elastictok: Adaptive tokenization for image and video. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 15703–15712. Computer Vision Foundation / IEEE, 2025.

Jingfeng Yao, Yuda Song, Yucong Zhou, and Xinggang Wang. Towards scalable pre-training of visual tokenizers for generation. In *Computer Vision - ECCV 2026 - 19th European Conference, Malmö, Sweden, September 8-12, 2026, Proceedings, Part XXIX*, volume 17029 of *Lecture Notes in Computer Science*, pp. 661–678. Springer, 2026.

Junyan Ye, Dongzhi Jiang, Zihao Wang, Leqi Zhu, Zhenghao Hu, Zilong Huang, Jun He, Zhiyuan Yan, Jinghua Yu, Hongsheng Li, Conghui He, and Weijia Li. Echo-4o: Harnessing the power of gpt-4o synthetic images for improved image generation, 2025. URL https://arxiv.org/abs/2508.09987.

Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. An image is worth 32 tokens for reconstruction and generation. In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 9556–9567. IEEE, 2024.

Xu Zhang, Cheng Da, Huan Yang, Kun Gai, Ming Lu, and Zhan Ma. ResTok: Learning hierarchical residuals in 1d visual tokenizers for autoregressive image generation. *arXiv preprint arXiv:2601.03955*, 2026a.

Yi Zhang, Bolin Ni, Xin-Sheng Chen, Hengrui Zhang, Yongming Rao, Houwen Peng, Qinglin Lu, Winston Hu, Menghao Guo, and Shi-Min Hu. Bee: A high-quality corpus and full-stack suite to unlock advanced fully open mllms. In *The Fourteenth International Conference on Learning Representations, ICLR 2026, Rio de Janeiro, Brazil, April 23-27, 2026*. proceedings.iclr.cc, 2026b.

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.

Anlin Zheng, Xin Wen, Xuanyang Zhang, Chuofan Ma, Tiancai Wang, Gang Yu, Xiangyu Zhang, and Xiaojuan Qi. Vision foundation models as effective visual tokenizers for autoregressive generation. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 35791–35820, 2026.

A Additional Experimental Details

A.1 Tokenizer Architecture and Training

Tables A.1 and A.2 specify the tokenizer architectures and optimization settings for the main results. For controlled ablations, we train the tokenizer for 20 epochs. We adopt the adversarial training configuration of prior work (Xiong et al., 2025). Each controlled dense–MoETok comparison matches the backbone, active FFN capacity, supervision, and training schedule.

Table A.1: **Tokenizer model scales.** Model scaling jointly varies encoder and decoder capacity.

Configuration	Small	Base	Large
Encoder depth	8	12	24
Encoder heads	8	12	16
Encoder width	512	768	1024
Decoder depth	12	24	24
Decoder heads	12	16	16
Decoder width	768	1024	1024
Latent tokens	64 / 256	64 / 256	64 / 256
MoE layers	All encoder blocks		

Table A.2: **Tokenizer optimization settings for the main results.**

Configuration	Setting
Experts / Top-k	8 / 2
Latent dimension / codebook size	256 / 8K
Training epochs	100
Global batch size	512
Learning-rate schedule	Cosine decay
Base learning rate	2×10^{-4}
Minimum learning rate	1×10^{-5}
Warm-up iterations	3,000
Optimizer	AdamW (Loshchilov & Hutter, 2019)
Optimizer (β_1, β_2)	$(0.9, 0.95)$
Reconstruction ℓ_2 weight	1.0
Perceptual loss weight	1.1
GAN loss weight	0.1
GAN start ratio	0.3
SigLIP weights (patch / global)	1.0 / 0.1

Linear probe evaluation. For standard tokenizer evaluation, we train a linear classifier (Yao et al., 2026) on flattened quantized tokens with the tokenizer frozen.

Autoregressive evaluation. To probe each controlled tokenizer's downstream capabilities, we train an autoregressive model with the LlamaGen GPT-B architecture (Sun et al., 2024) for 50,000 iterations, using the same data and optimization settings across tokenizers. We evaluate generation using gFID at CFG scale 1. For the full ImageNet evaluation, we sweep the CFG scale from 1 to 16 in increments of 0.5 and report the lowest gFID for each model alongside its unguided result.

A.2 Unified Autoregressive Model Training

We initialize the language model from Qwen3.5-4B (Qwen Team, 2026), extend its vocabulary with MoETok visual tokens, and jointly train understanding and generation through next-token cross-entropy (Wang et al., 2024). Pretraining is followed by supervised fine-tuning (SFT). Tables A.3 and A.4 report the two stages; batch size is measured in packed text and visual tokens per training iteration.

Table A.3: **Unified-model pretraining.** Understanding and generation share one autoregressive model and visual vocabulary.

Configuration	Pretraining
Language-model initialization	Qwen3.5-4B
Visual tokenizer	MoETok E8/T2; dimension 256; codebook 8K
Visual token count	256
Generation / understanding loss weights	$1.5 : 1.0$
Generation / understanding token ratio	$60\% : 40\%$
Peak learning rate	10^{-4}
Warm-up iterations	5,000
Learning-rate schedule	Constant after warm-up
Optimizer	AdamW (Loshchilov & Hutter, 2019)
Optimizer $(\beta_1, \beta_2, \epsilon)$	$(0.9, 0.95, 10^{-15})$
Weight decay	0
Gradient clipping	1.0
Input image resolution	256×256
Iterations	100,000
Global batch size	1,048,576 packed tokens / iteration
Data mixture	Listed below

Pretraining data. Generation data comprise LAION-5B (Schuhmann et al., 2022), DataComp (Gadre et al., 2023), JourneyDB (Sun et al., 2023), TextAtlas5M (Wang et al., 2025), and PosterCraft (Chen et al., 2026). Understanding data comprise LLaVA-OneVision-1.5-Mid-Training (An et al., 2025), BigDocs (Rodríguez et al., 2025), Honey-Data-15M (Zhang et al., 2026b), and MAmmoTH-VL (Guo et al., 2025).

Table A.4: **Unified-model supervised fine-tuning.** SFT increases the relative weight and token share of understanding tasks.

Configuration	Supervised fine-tuning
Language-model initialization	Pretraining checkpoint at 100,000 iterations
Visual tokenizer	MoETok E8/T2; dimension 256; codebook 8K
Visual token count	256
Generation / understanding loss weights	$1.0 : 1.5$
Generation / understanding token ratio	$40\% : 60\%$
Peak learning rate	5×10^{-5}
Warm-up iterations	100
Learning-rate schedule	Cosine decay to 10^{-6}
Optimizer	AdamW (Loshchilov & Hutter, 2019)
Optimizer $(\beta_1, \beta_2, \epsilon)$	$(0.9, 0.95, 10^{-15})$
Weight decay	0
Gradient clipping	1.0
Input image resolution	256×256
Iterations	19,073
Global batch size	1,048,576 packed tokens / iteration
Data mixture	Listed below

SFT data. Generation data comprise Fine-T2I (Ma et al., 2026), JourneyDB (Sun et al., 2023), BLIP3o-60k (Chen et al., 2025c), ShareGPT-4o-Image (Chen et al., 2025d), Echo-4o-Image (Ye et al., 2025), TextAtlas5M (Wang et al., 2025), and PosterCraft (Chen et al., 2026). Understanding data comprise LLaVA-OneVision-1.5-Instruct (An et al., 2025) and MAmmoTH-VL (Guo et al., 2025).

WISE evaluation. All results in Table 8 use the WISE_Verified protocol with the Qwen3.5-35B-A3B judge (Niu et al., 2026).

B TOKEN-ROLE AND ROUTING ANALYSES

B.1 ATTENTION VARIANCE

We extend the main-text analysis across image distributions, encoder layers, and the complete query sequence. Attention maps are averaged across heads and normalized over patch positions before computing $V_{\ell,q}$ as defined in Eq. equation 1. For the degenerate case of uniform attention on every image, we set $V_{\ell,q} = 0$.

Cross-dataset stability. We compute query-level attention variance V_q independently on 50,000 ImageNet validation images and 30,000 COCO images (Lin et al., 2014). Across the ten tokenizers in Figure B.1, Pearson correlations range from 0.93 to 1.00 and Spearman correlations from 0.83 to 0.99. The strong linear and rank correlations show that both the values and relative ordering of query attention variance remain consistent across these image distributions.

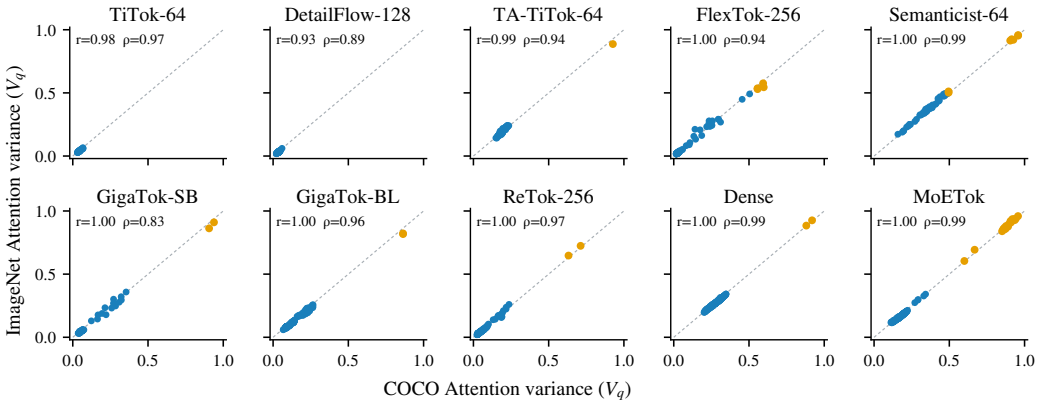

Figure B.1: **Query-level attention variance is consistent across datasets.** Each point compares the same query on COCO and ImageNet; r and ρ denote Pearson and Spearman correlations.

Attention patterns across depth. Figure B.2 follows selected queries through the final six encoder layers. The first two rows for each model show Content queries, and the remaining rows show Detail queries. Both models are evaluated on the same images; each head-averaged map is peak-normalized to highlight its spatial pattern. The layer-wise comparison reveals how the two roles evolve toward the bottleneck: their attention patterns become less distinct at the Dense output, while MoETok maintains clearer differentiation.

Attention and reconstruction across queries. Figure B.3 extends the single-query masking analysis (Beyer et al., 2025) to all 256 Dense queries. Matched grid cells show each query's encoder attention and the pixel-wise RGB L2 change induced by zero-masking its quantized token. Detail queries have strongly aligned attention and reconstruction footprints, whereas Content queries attend adaptively across images and induce negligible reconstruction changes. This full-sequence view establishes that the correspondence observed in the main-text examples extends across the query set.

B.2 LAYER-WISE ROUTING ANALYSIS

Following category-conditioned routing analyses (Muennighoff et al., 2025), we measure $p_\ell(e \mid c)$, the fraction of tokens of role c assigned to expert e by Top-1 routing. The overall assignment distribution $p_\ell(e)$ pools all tokens with their observed frequencies. Panel (a) of Figure 4 reports its entropy-equivalent effective expert count, $N_{\text{eff},\ell} = \exp H(p_\ell)$ (Hill, 1973), where H is Shannon entropy. Panel (b) reports the generalized Jensen–Shannon divergence among the three role-conditioned distributions, using equal role weights and dividing by $\log 3$ (Lin, 1991). All logarithms are natural; the normalized score lies in $[0, 1]$, with larger values indicating stronger routing separation.

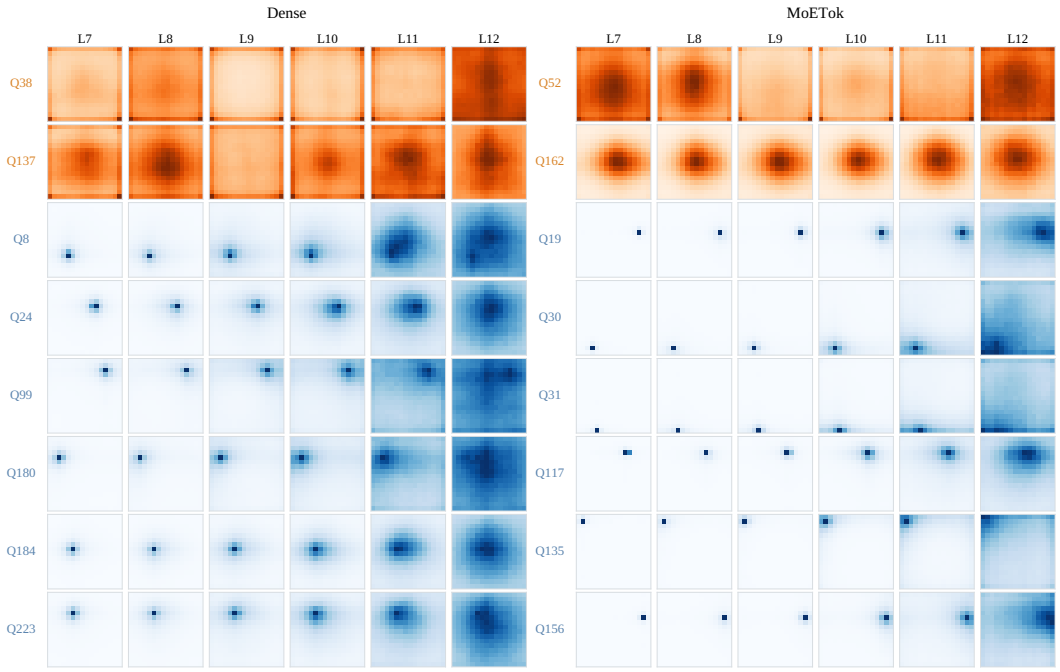

Figure B.2: **Content and Detail attention patterns across encoder depth.** MoETok preserves clearer role differentiation near the bottleneck than Dense.

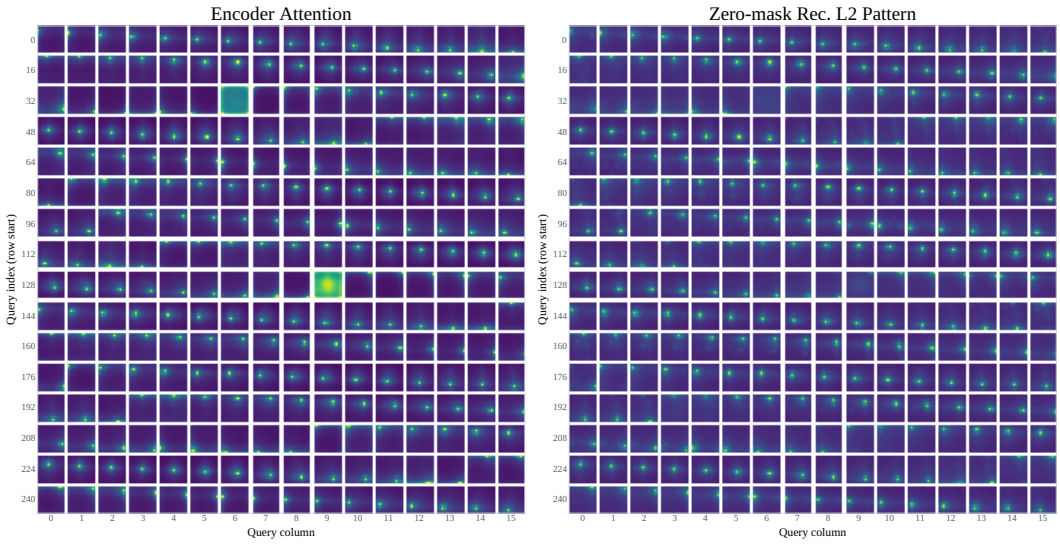

Figure B.3: **Encoder attention aligns with reconstruction contributions.** Detail queries exhibit highly consistent spatial patterns in encoder attention (left) and reconstruction changes under zero masking (right). Content queries with image-adaptive attention have negligible reconstruction effects.

Panel (c) normalizes $p_\ell(e \mid c)$ across roles within each expert to show role-balanced preferences. The percentages at right give overall Top-1 assignment shares; expert indices are layer-local.

Depth-dependent expert utilization. Both E4/T2 and E8/T2 concentrate expert use in early layers and engage a broader expert pool in middle and late layers (Figure B.4). We additionally evaluate loss-free balancing (DeepSeek-AI, 2024) with E4/T2. Balancing moves N_{eff} closer to its maximum of four across most layers, but worsens reconstruction, semantic representation, and generation (Table B.1). These results support allowing tokenizer objectives to determine expert allocation across depth, rather than encouraging uniform utilization in every layer.

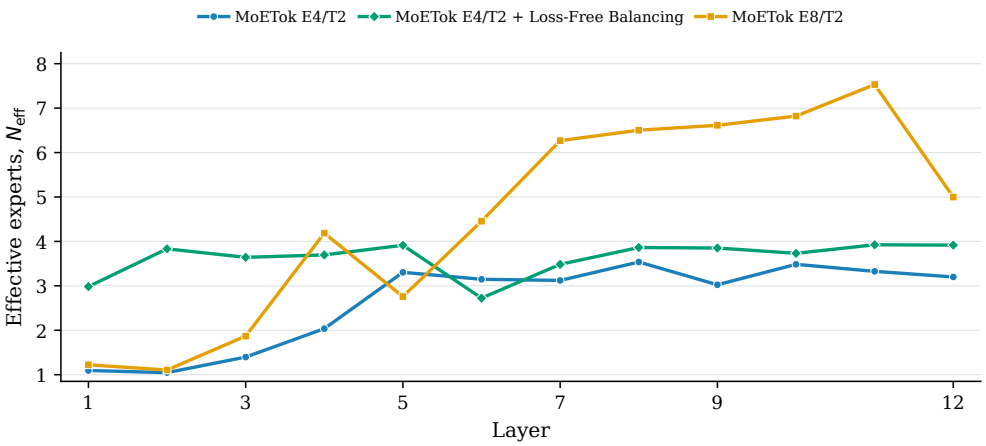

Figure B.4: **Expert utilization across depth in 1D tokenizers.** Both expert-pool sizes exhibit concentrated early-layer routing and broader utilization in deeper layers. Loss-free balancing makes utilization more uniform across most layers.

Table B.1: **Loss-free balancing in a 1D tokenizer.** For E4/T2, task-driven routing achieves better reconstruction, semantic representation, and generation than the balancing variant.

Loss-free balancing	rFID ↓	Linear probe ↑	Zero-shot ↑	gFID ↓
Disabled	**1.368**	**60.22**	**65.03**	**9.59**
Enabled	1.491	55.39	63.29	12.65

C Additional Controlled Studies

Encoder and decoder routing. In 1D tokenizers, routing in the encoder substantially improves semantic representation and achieves the best reconstruction and generation among the four configurations (Table C.1). Routing in the decoder alone provides modest semantic gains but worsens reconstruction. Routing in both further improves semantic representation, while yielding higher rFID and gFID than routing in the encoder alone. These results support our design of applying expert routing throughout the encoder while retaining a dense decoder.

Routing granularity in 1D tokenizers. We jointly vary expert count, width, and Top-k while keeping total and active expert capacities fixed (Table C.2). E4/T2 outperforms both fewer, wider experts and more numerous, narrower experts across all three metrics. The allocation of a fixed expert capacity budget thus affects representation and generation in 1D tokenizers.

Latent dimension and codebook size. Table C.3 separates the effects of codebook size and latent dimension. At dimension 256, increasing the codebook from 8K to 16K improves all four metrics. Expanding it further to 64K slightly improves rFID while reducing semantic and generation performance. With an 8K codebook, dimension 256 gives the strongest linear probe and lowest

Table C.1: **Encoder and decoder routing in 1D tokenizers.** Routing in the encoder gives the best reconstruction and generation; extending routing to the decoder further improves semantic representation.

Encoder	Decoder	rFID ↓	Linear probe ↑	Zero-shot ↑	gFID ↓
Dense	Dense	1.522	48.72	59.17	10.35
MoE	Dense	**1.368**	60.22	65.03	**9.59**
Dense	MoE	1.813	49.41	60.39	9.95
MoE	MoE	1.458	**62.75**	**66.61**	11.23

Table C.2: **Routing granularity in 1D tokenizers.** Total and active expert capacities are fixed at $16D$ and $8D$. E4/T2 performs best across all three metrics.

Experts / Top-k	Width	Total / Active	Linear probe ↑	Zero-shot ↑	gFID ↓
2 / 1	$8D$	$16D/8D$	55.08	61.69	10.80
4 / 2	$4D$	$16D/8D$	**60.22**	**65.03**	**9.59**
8 / 4	$2D$	$16D/8D$	57.21	64.02	9.85

gFID, whereas dimension 1024 favors zero-shot accuracy and reconstruction. These results show that codebook size and latent width affect representation, reconstruction, and generation differently, with intermediate settings offering a favorable overall balance.

Table C.3: **Latent dimension and codebook size.** We vary codebook size at dimension 256 (top) and dimension with an 8K codebook (bottom). gFID uses CFG scale 1.

Dim.	Codebook	Linear probe ↑	Zero-shot ↑	rFID ↓	gFID ↓
256	8K	60.63	66.37	1.53	8.99
256	16K	**62.77**	**67.48**	1.40	**8.74**
256	64K	62.01	67.28	**1.39**	10.32
64	8K	58.84	64.61	1.56	9.70
256	8K	**60.63**	66.37	1.53	**8.99**
1024	8K	59.38	**66.69**	**1.49**	9.42