RecTok: Reconstruction Distillation along Rectified Flow

Qingyu Shi[1,3*], Size Wu[2†*], Jinbin Bai[1,4], Kaidong Yu[3], Yujing Wang[1],
Yunhai Tong[1‡], Xiangtai Li[2], Xuelong Li[3]

[1]Peking University [2]Nanyang Technological University [3]TeleAI [4]Collov Labs

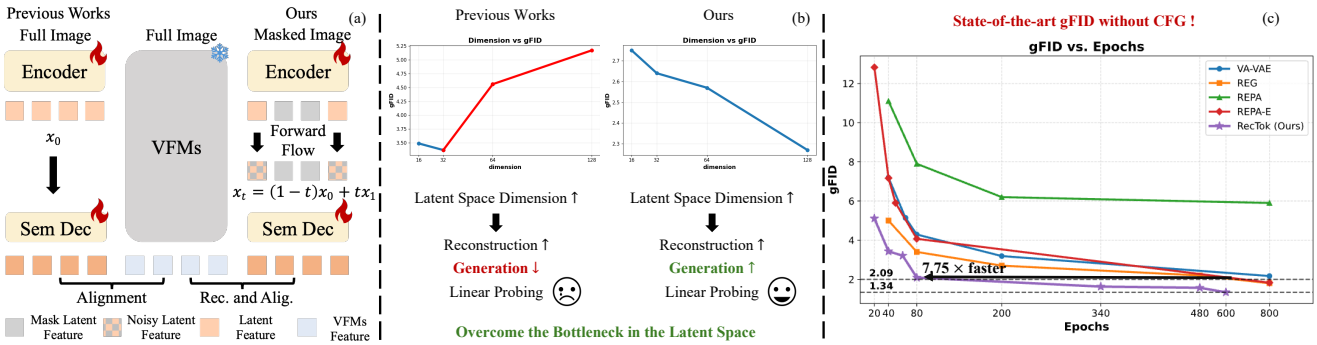


Figure 1. (a) presents the core insights of our approach. Unlike previous works, we enhance semantic information along the forward pass of the rectified flow via reconstruction distillation. Fig. (b) shows that increasing the latent space dimension consistently improves the generation performance of **RecTok**, indicating that the dimensional bottleneck no longer constrains the semantic information encoded in the latent features. (c) compares the gFID convergence across training epochs, where our method converges **7.75× faster** than prior works and achieves a final gFID of **1.34 without classifier-free guidance**, the state-of-the-art gFID performance to date.

Abstract

Visual tokenizers play a crucial role in diffusion models. The dimensionality of latent space governs both reconstruction fidelity and the semantic expressiveness of the latent feature. However, a fundamental trade-off is inherent between dimensionality and generation quality, constraining existing methods to low-dimensional latent spaces. Although recent works have leveraged vision foundation models (VFMs) to enrich the semantics of visual tokenizers and accelerate convergence, high-dimensional tokenizers still underperform their low-dimensional counterparts. In this work, we propose RecTok, which overcomes the limitations of high-dimensional visual tokenizers through two key innovations: flow semantic distillation and reconstruction–alignment distillation. Our key insight is to make the forward flow in flow matching semantically rich, which serves as the training space of diffusion transformers, rather than focusing on the latent space as in previous works. Specifically, our method distill the semantic information in VFMs into the forward flow trajectories in flow matching. And we further enhance the semantics by introducing a masked feature reconstruction loss. Our RecTok achieves superior image reconstruction, generation quality, and discriminative performance. It achieves state-of-the-art results on the gFID-50K under both with and without classifier-free guidance settings, while maintaining a semantically rich latent space structure. Furthermore, as the latent dimensionality increases, we observe consistent improvements. Code and model are available at https://shi-qingyu.github.io/rectok.github.io/.

1. Introduction

Diffusion modeling [13, 16, 21, 26, 35, 41] has become the dominant paradigm for image and video generation. As a crucial component, the visual tokenizer [24, 65] projects images from raw pixels to a compact latent space. Since the denoising network [5, 34, 38] is trained entirely in the latent space, computational cost is significantly reduced. However, the latent space is typically restricted to low feature dimensions to simplify diffusion training [13, 26, 41],

*Equal contribution. †Project lead. ‡Corresponding author.

which in turn limits both reconstruction fidelity and semantic expressiveness [52, 62]. Therefore, expanding the latent space while maintaining generative stability becomes a fundamental challenge in training visual tokenizers.

To address this limitation, previous methods [7, 60, 62] distill semantic information from vision foundation models (VFMs) [17, 25, 37, 55] into the latent space, aiming to enrich representation capacity and accelerate generative convergence. However, their generation quality in high dimensions still lags behind their low-dimensional counterparts. Thus, these approaches remain constrained to low-dimensional latent spaces (*e.g.*, dimension 32). Recently, RAE [69] increases DiT width to accommodate high-dimensional latents for diffusion training, achieving promising generative performance. However, its *reconstruction performance* lags behind previous methods — a limitation that is detrimental to tasks such as editing [1, 2, 14, 18, 50] and personalized generation [42, 49]. Furthermore, RAE does not systematically explore how dimensionality affects reconstruction, generation, and semantic representation. In this work, we revisit this question and present a principled framework for training high-dimensional visual tokenizers without compromising performance.

Unlike previous works [6, 7, 62] that directly inject semantics to the un-noised latent x_0, we take a more training-consistent perspective: Since DiT is trained on the forward flow $\{x_t \mid t \in [0, 1]\}$ rather than on x_0, we enhance the semantics of all flow states x_t. To understand the importance of semantic consistency along the flow, we first evaluate the discriminative capability of latent features across x_t. As shown in Fig. 2, the linear probing accuracy of several representative tokenizers [24, 61, 62] drops remarkably as the latent is propagated along the forward flow—the very representations that DiT receives during diffusion training. This degradation highlights the need for semantic enhancement throughout the entire flow, not just at x_0. In this work, we propose *RecTok* with two key innovations to enhance semantic consistency along the forward flow, simultaneously improving dimensionality and generative quality.

The *first* innovation is **Flow Semantic Distillation (FSD)**. Our key insight is to distill the semantics of VFMs into the forward flow trajectory $\{x_t \mid t \in [0, 1]\}$, which represents the interpolation of clean data x_0 and noise x_1. We utilize a lightweight semantic decoder to extract semantic features from points along the flow. These features are supervised by the corresponding representations from VFMs, as illustrated in Fig. 1 (a). FSD explicitly encourages the forward flow path $\{x_t \mid t \in [0, 1]\}$ to remain semantically discriminative. Consequently, our RecTok exhibits even better accuracy on the flow than the latent features, as shown in Fig. 2. The *second* innovation is **Reconstruction and Alignment Distillation (RAD)**. Inspired by masked image modeling methods [17, 58, 71], which obtain semantically rich features through pixel or feature reconstruction, we introduce a reconstructive target during FSD. Specifically, we apply random masks to the input image and reconstruct the missing regions based on the visible noisy latent features. We align the reconstructed latent features with full image features extracted from VFMs.

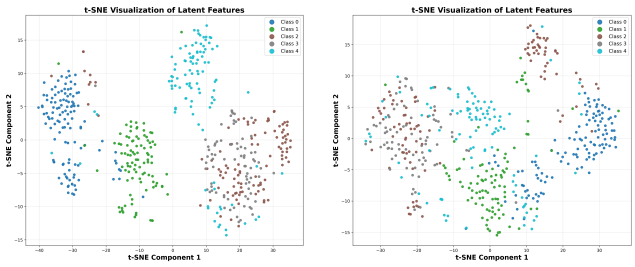
Following previous works [61, 62, 69], we train and evaluate our tokenizer and DiT [69] on the ImageNet-1K dataset [43]. As the latent dimensionality increases, we observe consistent improvements across reconstruction, generation, and linear probing tasks. Compared to other distillation or VFM-based visual tokenizers, our approach exhibits a clear advantage in convergence speed and generation quality, especially under without classifier-free guidance [20] setting. To summarize, our key contributions include:

- We identify the significance of enhancing semantics of forward flow trajectories, and introduce FSD and RAD that effectively expedite diffusion training.
- Our tokenizer achieves an effective balance among reconstruction, generation quality, and semantic representation.
- We demonstrate that all three aspects mentioned above can be consistently improved by increasing the dimensionality of the latent space.

2. Related Work

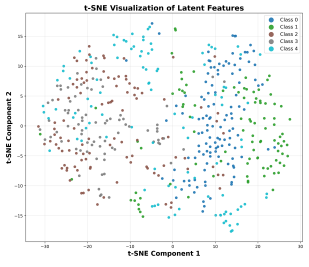
Visual Tokenizers for Image Generation. Broadly, visual tokenizers fall into two categories: discrete and continuous. Discrete tokenizers quantize image features with a learnable codebook [12, 36, 40, 63], and later works focus on enlarging the codebook and improving utilization [64, 65]. Despite these advances, their reconstruction quality remains inferior to continuous tokenizers, limiting downstream generative performance [3, 4, 48, 59]. Continuous tokenizers instead map images into a continuous latent space. Representative models such as VAE [24] regularize this latent space using a KL loss, while subsequent works [12] introduce perceptual and adversarial losses to improve reconstruction quality. Recent studies [7, 62] have also shown the advantage of aligning the latent space with the features of Vision Foundation Models (VFMs) [37, 55], which accelerates convergence and improves downstream generation quality. However, these approaches still restrict the latent representation to a low-dimensional space, constraining semantic expressiveness and reconstruction fidelity. In contrast, our work further expands the dimensionality of the latent space and observes continued improvements in generation quality.

High-dimensional Latent Space for Diffusion Models. A high-dimensional latent space is crucial for high-fidelity reconstruction and preserving rich semantic information. However, scaling latent dimensionality presents an inherent optimization challenge that often degrades generative per-

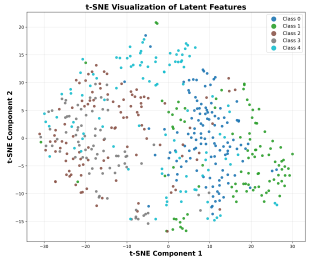


(a) RecTok, L.P. Acc.=55.61%

(b) VA-VAE, L.P. Acc.=18.36%



(c) DeTok, L.P. Acc.=10.2%



(d) VAE, L.P. Acc.=5.4%

Figure 2. **Linear probing results on x_t.** We evaluate the discriminative ability of representative tokenizers on the forward flow through linear probing on x_t. Specifically, we fix $t = 0.5$. As shown in Fig. 2b–2d, both the t-SNE visualization and the linear probing accuracy demonstrate that their latent features perform poorly during the training of DiT. In contrast, our RecTok exhibits a clear advantage even under noise interpolation.

formance. Although VA-VAE [62] alleviates part of this difficulty via a VFM loss, its convergence in high dimensions remains noticeably slower than that of low-dimensional variants. Another line of works [6, 68] initialize the visual encoder with VFMs. Yet, these methods still project high-dimensional features into low-dimensional latents (e.g., dimension 32), inevitably discarding rich semantics in the VFMs. More recently, RAE [69] makes the first attempt to perform diffusion directly in the high-dimensional feature space of VFMs. However, because the VFM is kept frozen, this approach inevitably loses fine-grained details, leading to reconstruction artifacts. In concurrent work, SVG [46, 47] employs a residual encoder to enhance reconstruction fidelity while preserving semantics from VFMs. Nevertheless, a performance gap remains between SVG and state-of-the-art generation methods. In this work, we develop a high-dimensional visual tokenizer that simultaneously excels in reconstruction fidelity, generative capability, and semantic representation.

3. Method

3.1. Rectified Flow in Image Generation

Flow Matching. Flow matching methods [31] construct a distribution transformation between data x_0 and noise x_1 through forward and reverse flows. As a representative approach, Rectified Flow [32] adopts a forward flow defined by linear interpolation, which simplifies the formulation of the velocity field:

$$x_t = (1-t)\,x_0 + t\,x_1, \quad v_t = \frac{dx_t}{dt} = x_1 - x_0. \tag{1}$$

During training, a neural network $v_\theta(x, t)$ is optimized on the forward flow $\{x_t \mid t \in (0, 1]\}$ to approximate the velocity field as:

$$\mathcal{L}_{\text{RF}} = \mathbb{E}_{t,\,x_0,\,x_1}\left[\left\| v_\theta(x_t, t) - (x_1 - x_0)\right\|_2^2\right], \tag{2}$$

During generation, $v_\theta(x, t)$ predicts the velocity that gradually transforms noisy data into clean data x_0 through the ODE solver [32, 33].

Visual Tokenizers. To reduce the training cost of generation models, prior works [13, 26, 41] project images into a compact latent space via an encoder-decoder image tokenizer [24]. Given an input image $I \in \mathbb{R}^{H\times W\times 3}$, the encoder E_θ produces a latent $x_0 \in \mathbb{R}^{h\times w\times c}$, and the decoder D_ϕ reconstructs $\hat{I}$:

$$x_0 = E_\theta(I), \qquad \hat{I} = D_\phi(x_0). \tag{3}$$

Both the encoder and decoder are typically based on CNN [27] or ViT [11] architectures. Considering computational efficiency and scalability [69], we adopt a ViT-based encoder–decoder design in this work.

The training of visual tokenizers typically involves multiple objectives, including reconstruction loss, perceptual loss, GAN loss [36], and KL loss [24]. Moreover, recent studies [62] have shown that distilling semantic information from Vision Foundation Models (VFMs) [17, 37] into the latent space can accelerate the convergence of downstream generative models and further improve image quality. Overall, the general loss function of the visual tokenizer can be formulated as follows:

$$\mathcal{L} = \lambda_{\text{rec}}\,\mathcal{L}_{\text{rec}} + \lambda_{\text{per}}\,\mathcal{L}_{\text{per}} + \lambda_{\text{GAN}}\,\mathcal{L}_{\text{GAN}} + \lambda_{\text{KL}}\,\mathcal{L}_{\text{KL}} + \lambda_{\text{sem}}\,\mathcal{L}_{\text{sem}}. \tag{4}$$

3.2. Our Method: RecTok

Our motivation is illustrated in Fig. 2. Although previous methods improve the semantic information in x_0, the discriminative ability of x_t deteriorates significantly when training the diffusion transformers (DiTs). Our key insight is to enhance the semantic information not only in x_0, but also the forward flow $\{x_t \mid t \in [0, 1]\}$, where the DiTs are trained. In the following section, we present two key innovations that enhance semantic representation throughout the forward flow.

Flow Semantic Distillation (FSD). Our goal is to make every point x_t along the forward flow discriminative and semantically rich. Fortunately, the forward flow from data x_0

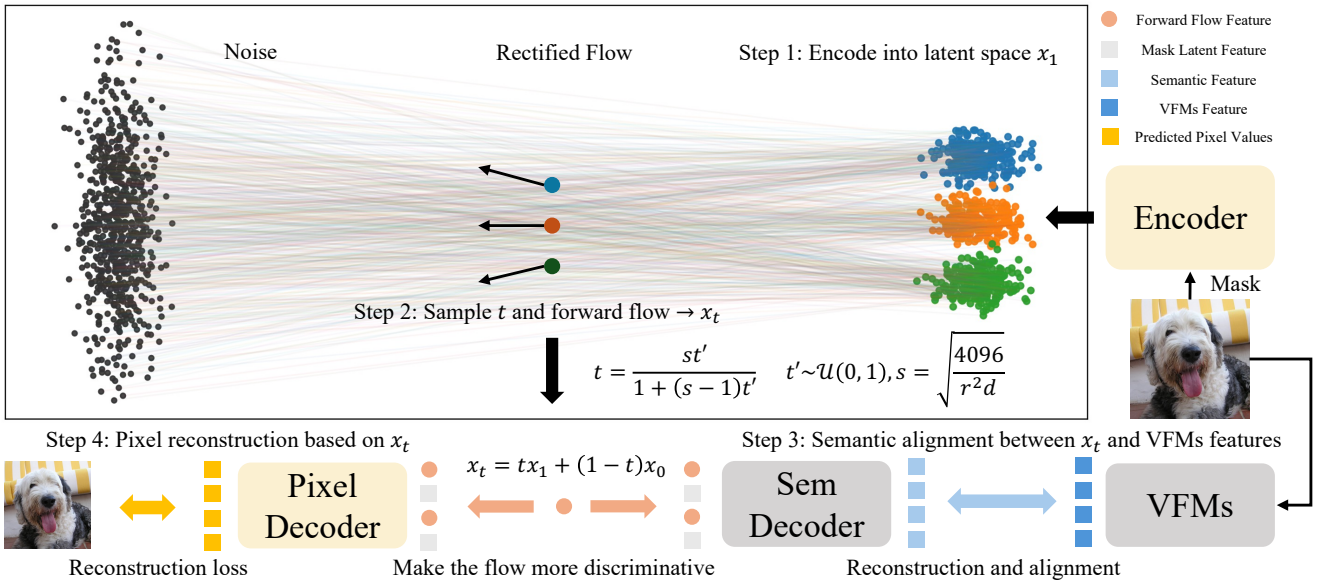


Figure 3. **Pipeline of RecTok.** During the training of RecTok, we apply a random mask to the input image and encode the visible regions using the encoder to obtain x_1. We then sample a time step t and use the forward flow to generate the corresponding x_t. Subsequently, x_t is fed into two decoders: the Semantic Decoder reconstructs the features of VFMs, while the Pixel Decoder reconstructs the pixel space. After training, both the Semantic Decoder and VFMs are discarded, ensuring the efficiency of RecTok during inference.

to noise ϵ is independent of the velocity network $v_\theta(x,t)$, allowing us to obtain $x_t = (1-t)x_0 + t\epsilon,\ t \in [0,1]$ easily through interpolation between the encoded $x_0 = E_\theta(I)$ and Gaussian noise ϵ. Each x_t is then decoded by a lightweight semantic decoder D_{sem} to obtain semantic features, which are supervised by Vision Foundation Models (VFMs) E_{VFM}:

$$\mathcal{L}_{\text{sem}} = 1 - \cos(D_{\text{sem}}(x_t), E_{\text{VFM}}(I)) \tag{5}$$

Specifically, the lightweight semantic decoder D_{sem} adopts a transformer architecture with only 1.5M parameters. A lightweight design enforces the encoder to capture richer semantic representations, as an overly powerful semantic decoder would otherwise draw away the semantic information from the encoder. We remove the normalization on $D_{\text{sem}}(x_t)$ and $E_{\text{VFM}}(I)$ for simplicity.

During FSD, we need to sample the timestep t. Considering the redundancy in high-dimensional latent spaces, we apply a dimension-dependent shift to the distribution of t, following RAE [69], and sample it as follows:

$$t = \frac{st'}{1+(s-1)t'}, \quad t' \sim \mathcal{U}(0,1), \quad s = \sqrt{\frac{4096}{r^2 d}} \tag{6}$$

where r, d is the resolution and dimension of the latent feature, respectively.

Reconstruction and Alignment Distillation (RAD). Inspired by masked image modeling methods [7, 17, 61], which enforce the model to learn robust representations by predicting unseen image patches. To further enhance the semantics along the flow. We introduce a reconstruction target in the FSD. Specifically, we apply random masks to the input image. We use a random mask ratio between -0.1 and 0.4. A negative ratio means that no mask is applied. After encoding the visible image into latent feature x_0^{vis}, we utilize a semantic decoder to reconstruct VFM features based on the $x_t^{\text{vis}} = (1-t)x_0^{\text{vis}} + t\epsilon$. To ensure compatibility with the reconstruction task, we utilize a transformer-based semantic decoder D_{sem}. The semantic loss $\mathcal{L}_{\text{sem}}$ is applied to both masked and unmasked regions. Our ablation study demonstrates that jointly performing semantic alignment and reconstruction yields the best overall performance.

Dimension of Latent Space. A fundamental limitation of previous tokenizers in generative models is their confinement to low-dimensional latent spaces. Although semantic distillation [62] and channel regularization [8] partially alleviate this issue, the best practice remains restricted to 32 dimensions. We progressively increase the dimensionality of the latent space. As shown in Tab. 2. Interestingly, this leads to consistent improvements in reconstruction (rFID, PSNR), generation (gFID, IS), and semantics (linear probing). This finding suggests the emergence of a shared latent space in higher dimensions that effectively supports low-level and high-level tasks.

Decoder Finetuning. After joint pixel and VFM-feature training, we freeze the encoder to preserve the learned latent semantics and finetune only the pixel decoder for image reconstruction. We disable the FSD and RAD and remove the

Method	Epochs	Params	Generation@256 w/o guidance				Generation@256 w/ guidance			
			gFID↓	IS↑	Prec.↑	Rec.↑	gFID↓	IS↑	Prec.↑	Rec.↑
Autoregressive										
VAR [54]	350	2.0B	1.92	**323.1**	**0.82**	0.59	1.73	**350.2**	0.82	0.60
MAR [30]	800	943M	2.35	227.8	0.79	0.62	1.55	303.7	0.81	0.62
l-DeTok [61]	800	479M	1.86	238.6	**0.82**	0.61	1.35	304.1	0.81	0.62
Pixel Diffusion										
ADM [10]	400	554M	10.94	101.0	0.69	0.63	3.94	215.8	**0.83**	0.53
RIN [22]	480	410M	3.42	182.0	-	-	-	-	-	-
PixelFlow [9]	320	677M	-	-	-	-	1.98	282.1	0.81	0.60
PixNerd [56]	160	700M	-	-	-	-	2.15	297.0	0.79	0.59
JiT [29]	600	2.0B	-	-	-	-	1.82	292.6	-	-
Latent Diffusion										
DiT [38]	1400	675M	9.62	121.5	0.67	0.67	2.27	278.2	**0.83**	0.57
MaskDiT [70]	1600	675M	5.69	177.9	0.74	0.60	2.28	276.6	0.80	0.61
SiT [34]	1400	675M	8.61	131.7	0.68	0.67	2.06	270.3	0.82	0.59
MDTv2 [15]	1080	675M	-	-	-	-	1.58	314.7	0.79	0.65
VA-VAE [62]	80	675M	4.29	-	-	-	-	-	-	-
	800		2.17	205.6	0.77	0.65	1.35	295.3	0.79	0.65
AFM [6]	800	675M	2.04	206.2	0.76	0.67	1.37	293.6	0.79	0.65
REPA [66]	80	675M	7.94	121.3	0.69	0.64	-	-	-	-
	800		5.90	157.8	0.70	**0.69**	1.42	305.7	0.80	0.64
DDT [57]	80	675M	6.62	135.2	0.69	0.67	1.52	263.7	0.78	0.63
	400		6.27	154.7	0.68	**0.69**	1.26	310.6	0.79	0.65
REPA-E [28]	80	675M	3.46	159.8	0.77	0.63	1.67	266.3	0.80	0.63
	800		1.83	217.3	0.77	0.66	1.26	**314.9**	0.79	0.66
SVGTok [47]	1400	675M	3.36	181.2	-	-	1.92	264.9	-	-
RAE [69]	80	839M	2.16	214.8	**0.82**	0.59	–	–	–	–
	800		1.51	242.9	0.79	0.63	**1.13**	262.6	0.78	**0.67**
RecTok (Ours)	80	839M	2.09	198.6	0.79	0.62	1.48	223.8	0.79	0.65
	600		**1.34**	**254.6**	0.78	0.65	**1.13**	289.2	0.79	**0.67**

Table 1. **Class-conditional performance on ImageNet 256×256.** RecTok reaches an FID of 1.34 and an IS of 254.6 without guidance, outperforming previous methods by a large margin. With AutoGuidance [23], it achieves an FID of 1.13 and an IS of 289.2 using only 600 epochs, representing the best overall performance.

losses $\mathcal{L}_{\text{KL}}$ and $\mathcal{L}_{\text{sem}}$. While we do not claim this as our primary contribution, it is a crucial step to improve reliability and quality of the reconstruction. We show the performance improvement in Tab. 10.

4. Experiment

4.1. Implementation Details

Visual Tokenizer. We adopt an architecture and training strategy largely following *l*-DeTok [61]. Specifically, we employ a ViT-B [11] backbone equipped with ROPE [53],

Table 2. **Results across different feature dimensions.** L.P. Acc. (L) denotes linear probing accuracy on latent features, while L.P. Acc. (SL) refers to linear probing accuracy on second-last layer features. As the feature dimension increases, discriminative ability, reconstruction, and generation show consistent gains.

Dim	L.P. Acc. (L)	L.P. Acc. (SL)	rFID	PSNR	gFID
16	24.1	62.9	0.74	22.75	2.75
32	38.8	63.7	0.71	24.08	2.64
64	47.2	65.0	0.66	24.93	2.57
128	**55.4**	**68.1**	**0.65**	**25.28**	**2.27**

SwiGLU [45], and RMSNorm [67] for both encoder and decoder. To investigate the impact of dimension on semantics, generation, and reconstruction, we train models with latent dimensions of 16, 32, 64, and 128. Note that this only affects the dimensionality of the ViT's linear head, so the resulting changes in parameter count and computational cost are negligible. Reparameterization and KL divergence are used to regularize the latent space. We train our tokenizer on the ImageNet-1K training set for 200 epochs. We set $\lambda_{rec} = 1.0$, $\lambda_{per} = 1.0$, $\lambda_{adv} = 0.5$, $\lambda_{kl} = 1 \times 10^{-6}$, and $\lambda_{sem} = 1$. The learning rate is set to 4×10^{-4}, with a linear warmup during the first 50 epochs followed by a cosine decay schedule for the remaining 150 epochs. We use a global batch size of 1024 and an EMA rate of 0.999. We evaluate our tokenizer through rFID [19] and PSNR on the ImageNet-1K validation set.

Diffusion Model. For the diffusion model, inspired by the advanced architecture of DiT^{DH} [69], we utilize DiT^{DH}-XL as our diffusion transformers. We train DiT^{DH}-XL on ImageNet-1K using rectified flow with a timestep shift strategy. The model is trained for 800 epochs with an initial learning rate of 2×10^{-4} and global batch size 1024, followed by a linear decay to 2×10^{-5} after 40 epochs. During training, we apply gradient clipping with a value of 1.0 and no weight decay. We use an EMA rate of 0.995, and all evaluations are conducted using the EMA-weighted model. For the ablation studies, the model trains for 80 epochs using the same training strategy. We evaluate the diffusion models on the ImageNet-1K validation set, measuring gFID, Inception Score (IS) [44], Precision, and Recall. We utilize AutoGuidance [23] as the classifier-free guidance method. The bad version in the AutoGuidance is a DiT^{DH}-S trained on ImageNet-1K for 30 epochs. During inference, we sample 150 steps using the Euler solver with a timestep shift; In the ablation studies, we sample only 50 steps and skip decoder finetuning to reduce computational cost.

All experiments are conducted on 32 H100 GPUs. Training the RecTok requires roughly 19 hours, while DiT^{DH} requires 10 hours for 80 epochs and 3 days for 600 epochs.

Table 3. **Tokenizer comparison on ImageNet-1K.** We compare RecTok with representative tokenizers in terms of parameters, GFLOPs, reconstruction, and generation. RecTok achieves the best performance among ViT-based tokenizers.

Tokenizer	Params	GFlops	ImageNet		
			rFID	PSNR	gFID
SD-VAE†	84M	445	0.62	26.04	8.30
VA-VAE	**70M**	310	**0.28**	**26.30**	2.17
MAETok	176M	54.2	0.48	23.61	2.21
DeTok	176M	**44.4**	0.52	23.53	1.86
RAE	395M	128.9	0.57	18.98	1.51
RecTok	176M	**44.4**	0.48	26.16	**1.34**

†Numbers reported from the original papers.

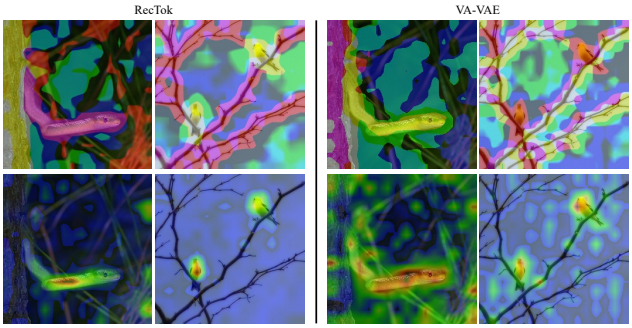


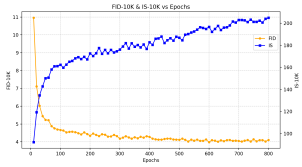
Figure 4. **Visualization of latent features.** We present the PCA projection and cosine similarity heatmap of the latent features from RecTok and VA-VAE. RecTok exhibits a more semantically rich latent space.

4.2. Main Results

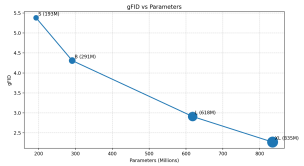
Tokenizer Performance. In Tab. 3, we compare RecTok with representative tokenizers [7, 41, 61, 62, 69] in parameters, computation, reconstruction, and generation. Although RecTok has more parameters than CNN-based methods, it achieves the lowest computational cost due to its efficient ViT architecture, which also benefits from modern acceleration. For reconstruction, RecTok outperforms other ViT-based methods and achieves state-of-the-art generation performance. It also learns a semantically richer latent space, surpassing prior tokenizers in linear probing (Fig. 2). As shown in Fig. 4, PCA and similarity heatmaps further demonstrate its more structured latent space compared to VA-VAE. Overall, RecTok provides the best trade-off among reconstruction fidelity, generation quality, and semantic representation.

In terms of latent dimension, we gradually expand the dimensionality of the latent space. As shown in Tab. 2, we observe a clear trend that the reconstruction, generation, and discriminative performances consistently improve as the dimension increases. To the best of our knowledge, RecTok is the first work to demonstrate such improvement across all

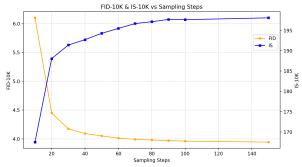
Figure 5. **Qualitative results on ImageNet-1K 256 × 256.** We show selected examples of class-conditional generation using DiT^{DH}-XL with AutoGuidance.



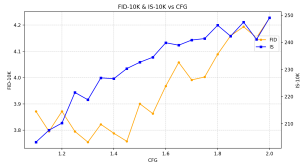
(a) **FID-10K and IS-10K vs Training Epochs.** FID converges around 600 epochs (between 4.0–4.1), while IS keeps increasing.



(b) **FID-50K vs Model Parameters.** Larger models yield lower gFID; evaluated with the 80 epochs checkpoint.



(c) **FID-10K and IS-10K vs Sampling Steps.** Strong performance is achieved by ∼60 steps (trained for 600 epochs).



(d) **FID-10K and IS-10K vs CFG.** Performance across classifier-free guidance scales.

Figure 6. **FID and IS under different settings.**

three aspects.

Generation Comparison. As shown in Tab. 1, RecTok with DiT^{DH}-XL achieves the best gFID=1.34 without classifier-free guidance [20]. When employing classifier-free guidance, we adopt the AutoGuidance strategy and achieve a gFID of 1.13, matching the gFID of RAE [69] while showing a clear advantage in Inception Score (IS) [44]. In Fig. 6a, we plot the gFID-10K and IS-10K curves over training epochs. We observe that gFID-10K stabilizes around 600 epochs, and therefore use the checkpoint at epoch 600 for reporting final results. The scaling results are presented in Fig. 6b, where the latent space of RecTok demonstrates strong scaling capability. Figs. 6c–6d illustrate how different inference settings affect gFID and IS. Considering the overall performance, we sample 150 steps with a guidance scale of 1.29.

Table 4. **Ablations on flow semantic distillation.** We compare our FSD with the x_0 semantic distillation using cosine similarity and VF loss. The experimental results demonstrate that FSD yields a significant improvement.

Setting	Sem Loss	L.P. Acc.	rFID	gFID	IS
w/o FSD	Cos Sim	44.35	0.69	3.35	157.3
	VF Loss [62]	37.52	0.72	3.91	142.1
w FSD	Cos Sim	**55.40**	**0.65**	**2.27**	**196.4**

4.3. Ablation Studies

Ablations on FSD. In Tab. 4, we study the effectiveness of FSD (i.e., applying distillation on x_t). When we only align the latent features x_0 to VFM features, a notable degradation of performance in generation and linear probing is observed. This suggests that distilling semantic information along the flow matching path benefits both generation performance and semantic representation. In Tab 5, we compare different λ_{sem}, considering overall performance, we set $\lambda_{sem} = 1$.

Ablations on Noise Schedule. As shown in Tab. 8, we compare different noise sampling strategies during RecTok training, including the Dimension-dependent Shift (referred to as Shift), Uniform, and Logit-Normal (Lognorm) methods. We observe that uniform sampling achieves the best reconstruction performance but performs the worst in generation quality. In contrast, the shift strategy yields slightly lower reconstruction scores but delivers the best generative results. Considering that the reconstruction quality can be further improved through decoder finetuning, as shown in Tab. 10, we adopt the Shift noise schedule as our default configuration.

Table 5. **Ablations on semantic loss weight λ_{sem}.** $\lambda_{sem} = 1$ achieves the best generation performance.

λ_{sem}	L.P. Acc.	rFID	gFID	IS
0.5	54.8	**0.59**	2.78	179.5
1	55.4	0.65	**2.27**	196.4
2	**56.1**	0.87	2.43	**199.7**

Table 6. **Ablations on vision foundation models (VFMs).** DINOv2 excels in low-dimensional latents (e.g., 16), while DINOv3 performs best in higher dimensions (e.g., 128).

VFM	Dim=16			Dim=128		
	rFID	gFID	IS	rFID	gFID	IS
DINOv3 [51]	0.81	2.86	**195.2**	0.65	**2.27**	**196.4**
DINOv2 [37]	0.74	**2.75**	183.3	0.53	2.38	184.7
SigLIP 2 [55]	0.71	3.59	172.3	0.51	3.14	178.6
RADIOv2 [39]	0.79	2.97	193.1	0.64	2.59	193.4
SAM [25]	0.83	4.96	141.1	0.69	4.47	157.2
Two VFMs	**0.69**	3.26	164.7	**0.49**	2.51	181.3

†Using two VFMs simultaneously (DINOv3 and SigLIP 2)

Table 7. **Ablations on reconstruction and alignment distillation.** RAD improves the generation performance, and the gain does not originate from the transformer architecture.

Setting	Sem Dec	rFID	gFID	IS
Align. only	MLP	0.76	3.02	175.3
Align. only	Transformer	**0.57**	2.52	184.5
Rec. only	Transformer	0.75	2.97	174.2
Rec. + Align.	Transformer	0.65	**2.27**	**196.4**

Ablations on VFMs. We study the impact of different VFMs in FSD, including DINOv3 [51], DINOv2 [37], SigLIP2 [55], RADIOv2 [39], and SAM [25], using large variants for consistency with VA-VAE. DINOv3 performs best in high-dimensional latent spaces, while DINOv2 excels in low-dimensional settings. Using multiple VFMs degrades generation performance. Thus, we adopt DINOv2 for 16/32 dimensions and DINOv3 for 64+ dimensions.

Ablations on RAD. We conduct four groups of ablation studies, including (1) alignment only, (2) reconstruction only, and (3) joint reconstruction and alignment. For the alignment only setting, we experiment with two types of semantic decoders: an MLP (4M parameters) and a lightweight Transformer (1.5M parameters). As shown in Table 7, the joint reconstruction and alignment strategy achieves the best overall performance, obtaining the lowest gFID, and the highest IS score.

Ablations on Semantic Decoder. In Tab. 9, we compare different architectural designs of the Semantic Decoder. We observe that using a transformer architecture consistently outperforms an MLP across all metrics. However, increasing the transformer's capacity leads to degraded performance in both linear probing accuracy and generation quality. Therefore, a lightweight transformer design provides the best overall trade-off.

Table 8. **Comparison of different sampling distributions.** Since the reconstruction quality can be improved through decoder finetuning, we adopt the Shift schedule.

Noise Schedule	L.P. Acc.	rFID	PSNR	gFID	IS
Uniform	55.1	**0.53**	**26.71**	2.50	191.6
Lognorm	53.5	0.57	25.03	2.37	**199.7**
Shift	**55.4**	0.65	25.28	**2.27**	196.4

Table 9. **Comparison of the performance on different semantic decoders.** A lightweight transformer achieves the best generation performance.

Sem Dec	Params	L.P. Acc.	rFID	gFID	IS
MLP	4M	51.2	0.76	3.02	175.3
Transformer	10M	47.3	**0.63**	3.13	170.5
Transfromer	1.5M	**55.4**	0.65	**2.27**	**196.4**

Table 10. **Overall ablation study of FSD, RAD, and Decoder Finetuning.** Each method provides a clear improvement.

Method	L.P. Acc.	rFID	PSNR	gFID	IS
Baseline	7.1	**0.22**	**29.76**	12.07	57.5
+ FSD	52.7	0.57	25.62	2.52	184.5
+ RAD	**55.4**	0.65	25.28	2.27	196.4
+ Dec FT	**55.4**	0.48	26.16	**2.23**	**198.2**

Overall Ablation Study. In Tab. 10, we present an ablation study on the two key innovations and the decoder finetuning. Each component brings a clear performance gain.

5. Conclusion

In this work, we address the fundamental challenge posed by the latent dimensionality of visual tokenizers through RecTok. Building on our core insight—enhancing semantic consistency along the forward flow rather than only at the un-noised latents. We introduce two key innovations: Flow Semantic Distillation (FSD) and Reconstruction and Alignment Distillation (RAD). Together, FSD and RAD effectively enrich the semantics of RecTok's latent space, and we observe consistent improvements as the latent dimension increases. Experiments on ImageNet-1K demonstrate that RecTok achieves state-of-the-art generation performance while maintaining strong reconstruction quality and semantic representation. We hope this work inspires future research on high-dimensional visual tokenizers.

Acknowledgement. This work is supported by the National Key Research and Development Program of China

(No. 2023YFC3807600)

References

[1] Jinbin Bai, Zhen Dong, Aosong Feng, Xiao Zhang, Tian Ye, and Kaicheng Zhou. Integrating view conditions for image synthesis. *arXiv preprint arXiv:2310.16002*, 2023. 2

[2] Jinbin Bai, Wei Chow, Ling Yang, Xiangtai Li, Juncheng Li, Hanwang Zhang, and Shuicheng Yan. Humanedit: A high-quality human-rewarded dataset for instruction-based image editing. *arXiv preprint arXiv:2412.04280*, 2024. 2

[3] Jinbin Bai, Tian Ye, Wei Chow, Enxin Song, Qing-Guo Chen, Xiangtai Li, Zhen Dong, Lei Zhu, and Shuicheng Yan. Meissonic: Revitalizing masked generative transformers for efficient high-resolution text-to-image synthesis. In *The Thirteenth International Conference on Learning Representations*, 2024. 2

[4] Jinbin Bai, Yu Lei, Hecong Wu, Yuchen Zhu, Shufan Li, Yi Xin, Xiangtai Li, Molei Tao, Aditya Grover, and Ming-Hsuan Yang. From masks to worlds: A hitchhiker's guide to world models. *arXiv preprint arXiv:2510.20668*, 2025. 2

[5] Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A ViT backbone for diffusion models. 2023. 1

[6] Bowei Chen, Sai Bi, Hao Tan, He Zhang, Tianyuan Zhang, Zhengqi Li, Yuanjun Xiong, Jianming Zhang, and Kai Zhang. Aligning visual foundation encoders to tokenizers for diffusion models. *arXiv preprint arXiv:2509.25162*, 2025. 2, 3, 5

[7] 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 *ICML*, 2025. 2, 4, 6

[8] Junyu Chen, Dongyun Zou, Wenkun He, Junsong Chen, Enze Xie, Song Han, and Han Cai. Dc-ae 1.5: Accelerating diffusion model convergence with structured latent space. *ICCV*, 2025. 4

[9] Shoufa Chen, Chongjian Ge, Shilong Zhang, Peize Sun, and Ping Luo. Pixelflow: Pixel-space generative models with flow. *arXiv preprint arXiv:2504.07963*, 2025. 5

[10] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *NeurIPS*, 2021. 5

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3, 5

[12] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 2

[13] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024. 1, 3

[14] Aosong Feng, Weikang Qiu, Jinbin Bai, Zhen Dong, Kaicheng Zhou, Xiao Zhang, Rex Ying, and Leandros Tassiulas. An item is worth a prompt: Versatile image editing with disentangled control. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 16559–16567, 2025. 2

[15] Shanghua Gao, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. Mdtv2: Masked diffusion transformer is a strong image synthesizer. *arXiv preprint arXiv:2303.14389*, 2023. 5

[16] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis, 2022. 1

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2021. 2, 3, 4

[18] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. 2022. 2

[19] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 6

[20] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. 2, 7

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 1

[22] Allan Jabri, David Fleet, and Ting Chen. Scalable adaptive computation for iterative generation. In *ICML*, 2023. 5

[23] Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. In *NeurIPS*, 2025. 5, 6

[24] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 1, 2, 3

[25] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023. 2, 8

[26] Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024. 1, 3

[27] Yann LeCun and Yoshua Bengio. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 1998. 3

[28] Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. Repa-e: Unlocking vae for end-to-end tuning with latent diffusion transformers. In *ICCV*, 2025. 5

[29] Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. *arXiv preprint arXiv:2511.13720*, 2025. 5

[30] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. In *NeurIPS*, 2024. 5

[31] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *ICLR*, 2023. 3

[32] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2023. 3

[33] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. In *NeurIPS*, 2022. 3

[34] Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *ECCV*, 2024. 1, 5

[35] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, 2021. 1

[36] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *NeurIPS*, 2017. 2, 3

[37] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *TMLR*, 2023. 2, 3, 8

[38] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 1, 5

[39] Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Am-radio: Agglomerative vision foundation model reduce all domains into one. In *CVPR*, 2024. 8

[40] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In *NeurIPS*, 2019. 2

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 3, 6

[42] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 2

[43] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 2

[44] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *NeurIPS*, 2016. 6, 7

[45] Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020. 6

[46] Minglei Shi, Haolin Wang, Borui Zhang, Wenzhao Zheng, Bohan Zeng, Ziyang Yuan, Xiaoshi Wu, Yuanxing Zhang, Huan Yang, Xintao Wang, Pengfei Wan, Kun Gai, Jie Zhou, and Jiwen Lu. Svg-t2i: Scaling up text-to-image latent diffusion model without variational autoencoder, 2025. 3

[47] Minglei Shi, Haolin Wang, Wenzhao Zheng, Ziyang Yuan, Xiaoshi Wu, Xintao Wang, Pengfei Wan, Jie Zhou, and Jiwen Lu. Latent diffusion model without variational autoencoder, 2025. 3, 5

[48] Qingyu Shi, Jinbin Bai, Zhuoran Zhao, Wenhao Chai, Kaidong Yu, Jianzong Wu, Shuangyong Song, Yunhai Tong, Xiangtai Li, Xuelong Li, et al. Muddit: Liberating generation beyond text-to-image with a unified discrete diffusion model. *arXiv preprint arXiv:2505.23606*, 2025. 2

[49] Qingyu Shi, Lu Qi, Jianzong Wu, Jinbin Bai, Jingbo Wang, Yunhai Tong, and Xiangtai Li. Dreamrelation: Bridging customization and relation generaion. In *CVPR*, 2025. 2

[50] Qingyu Shi, Jianzong Wu, Jinbin Bai, Jiangning Zhang, Lu Qi, Yunhai Tong, and Xiangtai Li. Decouple and track: Benchmarking and improving video diffusion transformers for motion transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10995–11005, 2025. 2

[51] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. 8

[52] Ivan Skorokhodov, Sharath Girish, Benran Hu, Willi Menapace, Yanyu Li, Rameen Abdal, Sergey Tulyakov, and Aliaksandr Siarohin. Improving the diffusability of autoencoders. In *ICML*, 2025. 2

[53] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 2024. 5

[54] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. In *NeurIPS*, 2024. 5

[55] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. 2, 8

[56] Shuai Wang, Ziteng Gao, Chenhui Zhu, Weilin Huang, and Limin Wang. Pixnerd: Pixel neural field diffusion. *arXiv preprint arXiv:2507.23268*, 2025. 5

[57] Shuai Wang, Zhi Tian, Weilin Huang, and Limin Wang. Ddt: Decoupled diffusion transformer, 2025. 5

[58] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *CVPR*, 2022. 2

[59] Yi Xin, Qi Qin, Siqi Luo, Kaiwen Zhu, Juncheng Yan, Yan Tai, Jiayi Lei, Yuewen Cao, Keqi Wang, Yibin Wang, et al. Lumina-dimoo: An omni diffusion large language model for multi-modal generation and understanding. *arXiv preprint arXiv:2510.06308*, 2025. 2

[60] Tianwei Xiong, Jun Hao Liew, Zilong Huang, Jiashi Feng, and Xihui Liu. Gigatok: Scaling visual tokenizers to 3 billion parameters for autoregressive image generation, 2025. 2

[61] Jiawei Yang, Tianhong Li, Lijie Fan, Yonglong Tian, and Yue Wang. Latent denoising makes good visual tokenizers, 2025. 2, 4, 5, 6

[62] Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *CVPR*, 2025. 2, 3, 4, 5, 6, 7

[63] 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. In *ICLR*, 2022. 2

[64] 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. In *ICLR*, 2024. 2

[65] 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. In *ICLR*, 2024. 1, 2

[66] 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 *ICLR*, 2025. 5

[67] Biao Zhang and Rico Sennrich. Root mean square layer normalization. *NeurIPS*, 2019. 6

[68] 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 image generation, 2025. 3

[69] Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025. 2, 3, 4, 5, 6, 7

[70] Hongkai Zheng, Weili Nie, Arash Vahdat, and Anima Anandkumar. Fast training of diffusion models with masked transformers. *TMLR*, 2023. 5

[71] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021. 2