CATOK: TAMING MEAN FLOWS FOR ONE-DIMENSIONAL CAUSAL IMAGE TOKENIZATION

Anonymous authors
Paper under double-blind review

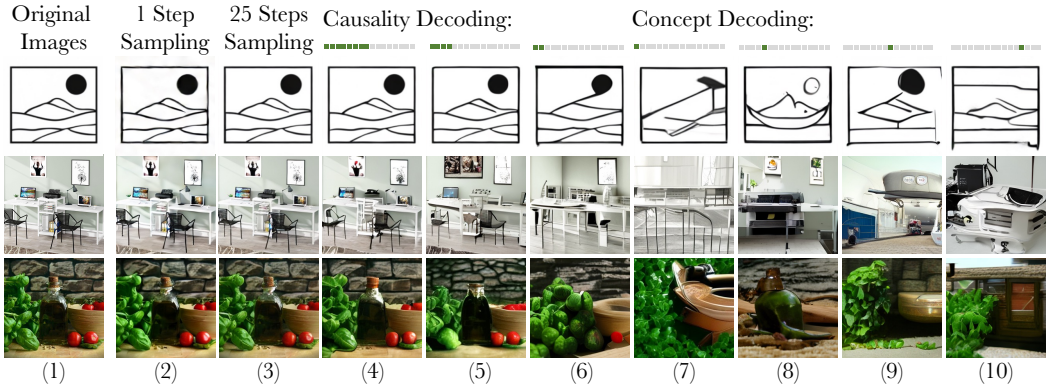


Figure 1: **Reconstruction samples.** CATOK with a MeanFlow decoder (Geng et al., 2025) supports fast one-step (col. 2) and high-quality multi-step (col. 3) sampling with 256 tokens. Reconstructions in cols. 3–7 show a fine-to-coarse trend as tokens are reduced from 256 to 16, highlighting the causality of the 1D tokens. Cols. 7–10 present reconstructions from different 16-token segments, demonstrating that CATOK naturally learns diverse visual concepts across token intervals.

ABSTRACT

Autoregressive (AR) language models rely on causal tokenization, but extending this paradigm to vision remains non-trivial. Current visual tokenizers either flatten 2D patches into non-causal sequences or enforce heuristic orderings that misalign with the "next-token prediction" pattern. Recent diffusion autoencoders similarly fall short: conditioning the decoder on all tokens lacks causality, while applying nested dropout mechanism introduces imbalance. To address these challenges, we present CATOK, a 1D causal image tokenizer with a MeanFlow decoder. By selecting tokens over time intervals and binding them to the MeanFlow objective, as illustrated in Fig. 1, CATOK learns causal 1D representations that support both fast one-step generation and high-fidelity multi-step sampling, while naturally capturing diverse visual concepts across token intervals. To further stabilize and accelerate training, we propose a straightforward regularization REPA-A, which aligns encoder features with Vision Foundation Models (VFMs). Experiments demonstrate that CATOK achieves state-of-the-art results on ImageNet reconstruction, reaching 22.72 PSNR and 0.681 SSIM with fewer training epochs, and the AR model attains performance comparable to leading approaches.

1 INTRODUCTION

The autoregressive (AR) paradigm enables generative large language models (LLMs) to achieve remarkable progress, exhibiting strong generalization and scalability (Achiam et al., 2023; OpenAI, 2025; Comanici et al., 2025; Grattafiori et al., 2024; Yang et al., 2025; Liu et al., 2024). Following the natural reading order of the text, LLMs tokenize a sentence into 1D causal tokens and perform generative modeling through next-token prediction. To emulate the capabilities and properties of LLMs in visual generation, the computer vision community has recently advanced large autoregressive

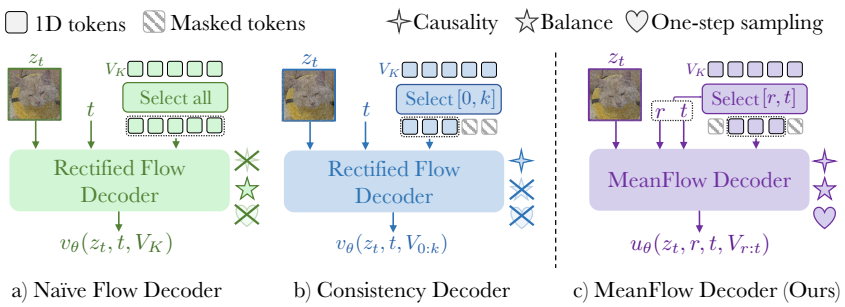


Figure 2: **Comparison among different decoders. a)** Naïve flow decoders (Sargent et al., 2025) condition on all 1D tokens from the encoder without dropout, leading the 1D tokens to lack causality; **b)** Consistency decoders obtain k by random sampling (Bachmann et al., 2025; Wen et al., 2025) or timestep binding (Pan et al., 2025; Wang et al., 2025a), and condition on the first k 1D tokens, which biases toward early tokens, introducing imbalance, leading to degraded performance of AR generation; **c)** Our MeanFlow decoder conditions on 1D tokens within the time interval $[r, t]$ to model the average velocity field along the subpath, which inherently maintains **causality** and **balance** of the 1D tokens, and supporting **one-step sampling** during reconstruction or generation.

vision models (Bai et al., 2024; Yu et al., 2024a; Wang et al., 2024; Sun et al., 2024; Han et al., 2025; Wang et al., 2025b). However, due to inferior performance, diffusion-based models (Ho et al., 2020; Song et al., 2021) like rectified flows (Liu et al., 2022; Lipman et al., 2022) remain the dominant approach in most scenarios (OpenAI, 2024; Wu et al., 2025).

In this paper, we argue that a crucial step toward bridging the gap between autoregressive language models and vision models lies in the causal tokenization of visual content. Autoregressive modeling relies on causal tokens and requires a predefined order of data. Unlike text, which inherently possesses a natural order, defining an appropriate order for images remains an open issue. VQGAN-like models (Esser et al., 2021) tokenize an image into grids of 2D tokens, and flatten them to a 1D sequence in raster (Razavi et al., 2019; Ramesh et al., 2021) or random (Yu et al., 2024b; Li et al., 2024b) order, which lacks causality between preceding and succeeding tokens (Pan et al., 2025; Wang et al., 2025a). VAR-like models (Tian et al., 2024; Han et al., 2025), on the other hand, tokenize images into multi-scale 2D tokens and establish a coarse-to-fine ordering via next-scale prediction. While this approach guarantees causality in visual tokens and yields promising results, it compromises the "next-token prediction" pattern of LLMs.

With the recent advances in 1D tokenizers (Cui et al., 2024; Yu et al., 2024c), the community has renewed its interest in diffusion autoencoders (Preechakul et al., 2022; Yang & Mandt, 2023) due to their demonstrated effectiveness in visual generation. Diffusion autoencoders extract 1D tokens from registers (Darcet et al., 2024) of encoders, and use them as conditions for the decoder to reconstruct images with denoising or rectified flow objective. However, as shown in Fig. 2 a), Naïve flow decoders, such as FlowMo (Sargent et al., 2025), condition on all 1D tokens from the encoder, causing the 1D tokens to lack causality and making AR learning difficult. To learn the causality for 1D tokens, as shown in Fig. 2 b), consistency decoders apply nested dropout (Rippel et al., 2014) by conditioning on the first k tokens, where k is determined either via random sampling, as in FlexTok (Bachmann et al., 2025) and Semanticist (Wen et al., 2025), or via timestep binding, as in DDT (Pan et al., 2025) and Selftok (Wang et al., 2025a). Since earlier tokens are more likely to be selected, this approach introduces imbalance and can be harmful to AR generation (see Tab. 3b).

Motivated by these observations, we propose CaTok, a 1D **Ca**usal image **Tok**enizer equipped with a MeanFlow decoder (Geng et al., 2025). As illustrated in Fig. 2 c), we address the imbalance problem by selecting 1D tokens within a sampled time interval $[r, t]$ and binding them with the corresponding time interval in the MeanFlow objective. This allows the 1D tokens to model the average velocity field along the subpath from r to t, capturing causality in the noise-to-image generation process while naturally supporting one-step sampling during generation. Moreover, inspired by REPA and REPA-E (Yu et al., 2024d; Leng et al., 2025), we align the image features from encoders with high-quality external visual representations, providing a regularization that effectively accelerates and stabilizes autoencoder training. We refer to this variant as REPA-A.

As shown in Fig. 1, CATOK supports both fast one-step sampling (col. 2) and high-quality multi-step sampling (col. 3) with 256 tokens, demonstrating its flexibility in balancing efficiency and fidelity. Reconstructions in cols. 3–7, obtained by progressively reducing the number of tokens from 256 to 16, exhibit a clear fine-to-coarse trend, providing evidence for the causality of the learned 1D tokens. Moreover, reconstructions in cols. 7–10 from different 16-token segments show that CATOK naturally learns diverse visual concepts across token intervals, underscoring its ability to disentangle semantic information and distribute it meaningfully among tokens. Our contributions can be summarized as:

1. We propose a novel architecture for 1D causal image tokenization based on diffusion autoencoders (Preechakul et al., 2022) with the MeanFlow (Geng et al., 2025) objective.
2. We seamlessly combine the training of a causal encoder and a one-step flow decoder, enabling one-step sampling in diffusion autoencoders ***for the first time***.
3. We propose REPA-A, an advanced technique that leverages existing vision foundation models to stabilize and accelerate diffusion autoencoder training.
4. On ImageNet, our CATOK-L achieves state-of-the-art results with 22.72 PSNR and 0.681 SSIM, while attains comparable performance o leading approaches with 0.84 rFID and 3.40 gFID.

2 BACKGROUND

In this section, we provide a concise introduction to rectified flows (Liu et al., 2022; Lipman et al., 2022) and MeanFlow models (Geng et al., 2025) as a preliminary to our CATOK.

2.1 RECTIFIED FLOWS

Given data $x \sim p_{data}(x)$ and prior $\epsilon \sim p_{prior}(\epsilon)$, rectified flows learn the conditional velocity fields $v_t = v_t(z_t|x)$ between these two distributions. Specifically, a flow path can be constructed as $z_t = (1-t)x + t\epsilon$ with time t, and the conditional velocity can be derived by:

$$v(z_t|x) = \frac{d}{dt} z_t = \epsilon - x. \tag{1}$$

A deep neural network $v_\theta(z_t, t)$ parameterized by θ is learned to model the marginal velocity field

$$v(z_t, t) \triangleq \mathbb{E}_{p_t(v_t|z_t)}[v_t], \tag{2}$$

which is equivalent to fitting the conditional velocity field in Eq. (1) (Lipman et al., 2022). In inference, starting from $z_1 = \epsilon \sim p_{prior}(\epsilon)$, samples can be generated by solving:

$$z_r = z_t - \int_r^t v_\theta(z_\tau, \tau) d\tau, \tag{3}$$

where r denotes another timestep and $r < t$. In practice, this integral is numerically approximated in discrete time steps. For instance, the Euler method updates each step as:

$$z_r = z_t - (t-r) v_\theta(z_t, t). \tag{4}$$

However, it estimates the average velocity over the interval $[r, t]$ using only the instantaneous velocity at time t, which introduces inaccuracies during sampling.

2.2 MEANFLOW MODELS

To mitigate the errors that arise with fewer sampling steps, MeanFlow models directly fit the average velocity u over the interval $[r, t]$. Formally, the average velocity u can be defined as:

$$u(z_t, r, t) \triangleq \frac{1}{t-r} \int_r^t v(z_\tau, \tau) d\tau. \tag{5}$$

Through derivations in Geng et al. (2025), the average velocity $u(z_t, r, t)$ can be obtained from the instantaneous velocity $v(z_t, t)$:

$$u(z_t, r, t) = v(z_t, t) - (t-r)(v(z_t, t)\partial_z u(z_t, r, t) + \partial_t u(z_t, r, t)), \tag{6}$$

and the MeanFlow objective is:

$$\mathcal{L}(\theta) = \mathbb{E}||u_\theta(z_t, r, t) - \text{sg}[v(z_t|x) - (t-r)(v(z_t|x)\partial_z u_\theta + \partial_t u_\theta)]||_2^2, \tag{7}$$

where $\text{sg}[\cdot]$ denotes the stop-gradient operation, avoiding double backpropagation through the Jacobian–vector product. Moreover, one-step sampling can be given by $z_0 = \epsilon - u_\theta(\epsilon, 0, 1)$.

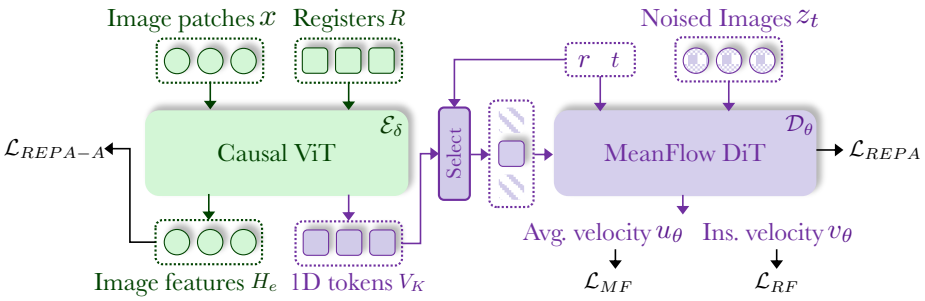


Figure 3: **Architecture of our CATOK.** CATOK is a diffusion autoencoder with a causal Vision Transformer (ViT) (Dosovitskiy et al., 2021) encoder and a MeanFlow Diffusion Transformer (DiT) (Peebles & Xie, 2023) decoder. The encoder leverages registers (Darcet et al., 2024) to extract rich visual information into 1D tokens, which are then conditioned to the decoder through time interval selecting. With two flow objectives and two representation alignment objectives, CATOK effectively learns causal 1D representations that support both one-step and multi-step sampling, while naturally capturing diverse visual concepts across different token intervals.

3 CATOK

We now introduce CATOK, a diffusion autoencoder (Preechakul et al., 2022; Yang & Mandt, 2023) with a causal Vision Transformer (ViT) (Dosovitskiy et al., 2021) encoder and a MeanFlow Diffusion Transformer (DiT) (Peebles & Xie, 2023) decoder, for 1D causal image tokenization. We begin in Sec. 3.1 by presenting the architecture of CATOK. Next, in Sec. 3.2, we describe how it is optimized through multiple objectives. Finally, in Sec. 3.3, we outline the autoregressive modeling procedure used for image generation with the trained CATOK.

3.1 ARCHITECTURE

As shown in Fig. 3, CATOK is a diffusion autoencoder with a causal ViT encoder $\mathcal{E}_\delta$ and a MeanFlow DiT decoder $\mathcal{D}_\theta$ parameterized by δ and θ respectively. Specifically, given an image x, we concatenate it with K registers R and send them into the encoder:

$$[H_e, V_K] = \mathcal{E}_\delta([x, R]), \tag{8}$$

where H_e denotes the image features and V_K represents the compressed 1D tokens. Furthermore, a causal attention mask is applied to enforce the dependency structure among 1D tokens (Cui et al., 2024; Bachmann et al., 2025; Wen et al., 2025). Specifically, image features can attend to each other but not to the 1D tokens; in contrast, 1D tokens are allowed to attend to all image features while being restricted to only their preceding 1D tokens.

In the MeanFlow DiT decoder phase, we first independently sample two timesteps r and t, ensuring that $r, t \in [0, 1]$ and $r < t$. Then, the flow path is constructed by linearly interpolating the image x with random noise $\epsilon \sim \mathcal{N}(0, 1)$:

$$z_t = (1 - t)x + t\epsilon. \tag{9}$$

By conditioning the noised image z_t with the 1D tokens from the interval $[r \cdot K, t \cdot K]$, denoted as $V_{r:t}$, and timesteps r, t, the DiT decoder predicts the average velocity u_θ over the time interval:

$$u_\theta = \mathcal{D}_\theta(z_t, r, t, V_{r:t}). \tag{10}$$

Since accurately modeling the instantaneous velocity field improves training stability when learning the average velocity field (Geng et al., 2025; Peng et al., 2025), we follow Eq. (5) and set $r = t$ to model the instantaneous velocity field v_θ:

$$v_\theta = \mathcal{D}_\theta(z_t, t, t, V_K), \tag{11}$$

and all the 1D tokens V_K are conditioned upon.

3.2 TRAINING

As illustrated in Fig. 3, CATOK is jointly optimized with two flow objectives—MeanFlow (Geng et al., 2025) and Rectified Flow (Liu et al., 2022; Lipman et al., 2022)—and two representation alignment objectives—REPA (Yu et al., 2024d) and our proposed REPA-A.

MeanFlow objective. From Eq. (1), Eq. (7) and Eq. (10), we define our MeanFlow objective as:

$$\mathcal{L}_{MF} := \mathbb{E}||u_\theta - (\epsilon - x) - \text{sg}[(t-r)((\epsilon - x)\partial_z u_\theta + \partial_t u_\theta)]||_2^2, \tag{12}$$

where $\text{sg}[\cdot]$ denotes the stop-gradient operation, and $(\epsilon - x)\partial_z u_\theta + \partial_t u_\theta$ is computed using the Jacobian-vector product operation.

Recitified Flow objective. We also model the instantaneous velocity field to enhance training stability. Based on Eq. (1), we define our Rectified Flow objective as follows:

$$\mathcal{L}_{RF} := \mathbb{E}||v_\theta - (\epsilon - x)||_2^2. \tag{13}$$

Following Geng et al. (2025), we employ an adaptive L_2 loss in place of the standard L_2 loss to enhance performance, defined as $\mathcal{L}_{\text{adaptive}} = ||\Delta||_2^2 \,/\, \text{sg}[(||\Delta||_2^2 + c)^w]$, where Δ denotes the regression error, and integrate the two objectives in $\mathcal{L}_F$ by fixing a proportion q of samples with $r = t$. In our implementation, we set $c = 10^{-3}$, $w = 1.0$, and $q = 75\%$.

REPA objective. REPA (Yu et al., 2024d) is a regularization technique that leverages Vision Foundation Models (VFMs) to assist DiT training and accelerate convergence. Formally, given the hidden states H_d from a middle layer of the DiT decoder and pretrained representations H_{vfm} from a VFM, our REPA objective can be defined as:

$$\mathcal{L}_{REPA} := -\mathbb{E}[\frac{1}{N}\sum_{n=1}^{N} \text{sim}(H_{vfm}^{[n]}, \text{proj}(H_d^{[n]}))], \tag{14}$$

where n is a patch index, $\text{sim}(\cdot,\cdot)$ is the cosine similarity function and $\text{proj}(\cdot)$ is the projection layer.

Our proposed REPA-A objective. Unlike REPA-E (Leng et al., 2025), which backpropagates gradients to the VAE (Kingma & Welling, 2014), or VA-VAE (Yao et al., 2025), which directly regularizes the compressed features of VAE using VFMs, we propose REPA-A, a representation alignment method specifically tailored for conditional diffusion autoencoders such as our CATOK. Formally, given the image features H_e from the ViT encoder and the same VFM representations H_{vfm}, the REPA-A objective can be defined as:

$$\mathcal{L}_{REPA-A} := -\mathbb{E}[\frac{1}{N}\sum_{n=1}^{N} \text{sim}(H_{vfm}^{[n]}, H_e^{[n]})], \tag{15}$$

where n is a patch index and $\text{sim}(\cdot,\cdot)$ is the cosine similarity function. With REPA-A, the encoder produces higher quality semantic representations, allowing 1D tokens to extract more informative and discriminative visual content, thereby accelerating convergence and enhancing overall performance.

3.3 AUTOREGRESSIVE MODELING

Once the causal 1D tokens V_K are obtained from a well-trained CATOK encoder, we train a standard autoregressive model following the "next-token prediction" paradigm to generate images. Formally, the AR model defines the generation process as:

$$p(V_1, V_2, ..., V_k) = \prod_{k=1}^{K} p(V_k | V_1, ..., V_k). \tag{16}$$

When V_k is represented as discrete indices, this probabilistic model can be optimized via cross-entropy loss. In contrast, when V_k is continuous-valued, as in our setting, optimization is performed using a diffusion loss introduced in Li et al. (2024b). More details are provided in Appendix A.

For image generation, given a prior such as a class token, we first obtain a predicted sequence $\hat{V}_K$ via Eq. (16). By feeding it into MeanFlow decoder, we can directly perform one-step sampling to render an image through $\hat{x} = \epsilon - \mathcal{D}_\theta(\epsilon, 0, 1, \hat{V}_K)$, where ϵ denotes a random Gaussian noise.

4 RELATED WORKS

AR modeling requires compressing raw data into a sequence of tokens, which in turn has spurred a line of research on visual tokenizers. In this section, we categorize them into three types.

2D visual tokenizers. VQ-VAE (Van Den Oord et al., 2017; Razavi et al., 2019) is one of the most widely adopted 2D visual tokenizers, integrating Vector Quantization (VQ) into the VAE (Kingma & Welling, 2014) to produce discrete tokens from image patches. Subsequent works improve upon this design: VQGAN (Esser et al., 2021) introduces an adversarial loss to enhance reconstruction quality, while RQ-VAE (Lee et al., 2022) employs multiple quantization stages. MAGVIT-v2 Yu et al. (2024a) further alleviates quantization bottlenecks with Look-up Free Quantization (LFQ), and MaskBit modernizes the VQGAN framework with binary quantized tokens. Most recently, VAR models (Tian et al., 2024; Han et al., 2025) tokenize images into multi-scale 2D tokens and establish a coarse-to-fine ordering via "next-scale prediction". However, these 2D tokenizers either lack causality across tokens or compromise the "next-token prediction" paradigm.

1D visual tokenizers. SEED (Cui et al., 2024) employs a causal Q-Former (Li et al., 2023) to extract 1D tokens from a ViT Dosovitskiy et al. (2021) encoder and performs semantic reconstruction with a pre-trained text encoder. TiTok (Yu et al., 2024c) derives 1D tokens using learnable registers and conditions a ViT decoder for mask-to-patch reconstruction. Building on these designs, a line of work explores 1D causal visual tokenizers. TexTok (Zha et al., 2025) and TA-TiTok (Kim et al., 2025) leverage textual conditioning to enhance performance, ALIT (Duggal et al., 2025) introduces adaptive-length tokenization via recurrent encoding, One-D-Piece (Miwa et al., 2025) applies nested dropout (Rippel et al., 2014) on tokens to introduce causality, and SpectralAR (Huang et al., 2025) adopts a similar architecture but imposes explicit spectral interpretations to supervise different tokens. In contrast, CATOK adopts a diffusion-based decoder, which we introduce next.

Diffusion autoencoders as 1D tokenizers. Diffusion autoencoders (Preechakul et al., 2022; Yang & Mandt, 2023; Li et al., 2024a; Wang et al., 2025c) compress image features into 1D tokens, which serve as conditioning inputs for diffusion models trained with denoising or rectified flow objectives. However, naïve flow decoders such as FlowMo (Sargent et al., 2025) and DiTo (Chen et al., 2025) condition on all tokens simultaneously, eliminating causal structure and thereby hindering AR learning. To address this, consistency decoders introduce causality through nested dropout, conditioning only on early tokens. The early-token set is determined either stochastically, as in FlexTok (Bachmann et al., 2025) and Semanticist (Wen et al., 2025), or deterministically via timestep binding, as in DDT (Pan et al., 2025) and Selftok (Wang et al., 2025a). However, because earlier tokens are disproportionately favored, these methods induce imbalance, which can degrade AR generation quality. In contrast, our CATOK leverages an additional MeanFlow (Geng et al., 2025) objective to capture causality in a balanced manner while naturally supporting one-step sampling.

5 EXPERIMENTS

For fair comparison, we follow common practice (Li et al., 2024b) and conduct experiments on ImageNet-1K (Deng et al., 2009) at 256 $\times$ 256 resolution.

5.1 IMPLEMENTATION DETAILS.

CATOK. The CATOK encoder is a ViT-B/8 (Dosovitskiy et al., 2021) with registers (Darcet et al., 2024) and causal attention masks (Cui et al., 2024; Bachmann et al., 2025). For fair comparison, the extracted 1D tokens are 16-dimensional and normalized before being passed to the decoder following (Li et al., 2024a; Wen et al., 2025). The decoder is either a DiT-B/4 or DiT-L/2 (Peebles & Xie, 2023), which are denoted as CATOK-B and CATOK-L respectively, operating on the latent space of a frozen, publicly available KL-16 MAR-VAE (Li et al., 2024b) to reduce computation. Both the encoder and decoder are trained from scratch on ImageNet-1K (Deng et al., 2009) training split. Besides, we utilize DINOv2-B/16 (Oquab et al., 2024) as the VFM of REPA and REPA-A, and the loss weights for $\mathcal{L}_R$, $\mathcal{L}_{REPA}$, and $\mathcal{L}_{REPA-A}$ are set to 1.0, 1.0, and 0.8, respectively.

Autoregressive modeling. Following (Wen et al., 2025), we evaluate frozen CATOK by training autoregressive generators LlamaGen (Sun et al., 2024) with a diffusion loss (Li et al., 2024b). The input sequence is conditioned with a learnable class token, which is randomly dropped with

Table 1: **Reconstruction results on ImageNet 256×256 benchmark.** "Token" denotes the number of tokens used for reconstruction, and "Dim." denotes the dimension of these tokens. "Param." indicates the model size, and "VQ" specifies whether the tokens are vector-quantized. "↓" or "↑" denote lower or higher values are better. †: enabling one-step sampling. ∗: without one-step render.

Method	Token	#Dim.	#Param.	Epochs	VQ	rFID↓	PSNR↑	SSIM↑
One-step 2D tokenizers								
VQGAN	16x16	16	307M	-	✓	7.94	-	-
LlamaGen	16x16	8	72M	40	✓	2.19	20.67	0.589
MaskBit	16x16	12	54M	270	✓	1.37	21.50	0.560
MAR-VAE	16x16	16	-	-	×	1.22	-	-
OpenMagViT-V2	16x16	-	116M	270	✓	1.17	21.63	0.640
One-step 1D tokenizers								
SpectralAR-64	64	16	172M	300	✓	4.03	-	-
TiTok-S-128	128	16	44M	300	✓	<u>1.71</u>	17.52	0.437
TiTok-L-32	32	8	614M	300	✓	2.21	15.60	0.359
One-D-Piece-B-256	256	16	172M	300	✓	**1.11**	18.77	-
CATOK-B-256†	256	16	224M	80	×	4.89	<u>20.77</u>	<u>0.617</u>
CATOK-L-32†	32	16	552M	160	×	4.48	17.25	0.441
CATOK-L-256†	256	16	552M	160	×	4.63	**20.99**	**0.630**
Diffusion tokenizers								
FlexTok d12-d12	256	6	170M	640	✓	4.20	-	-
FlexTok d18-d18	256	6	573M	640	✓	1.61	-	-
FlexTok d18-d28	256	6	1.4B	640	✓	1.45	18.53	0.465
Semanticist-L-256	256	16	552M	400	×	**0.78**	21.61	0.626
SelfTok-512*	512	16	-	-	✓	-	21.86	0.600
FlowMo-Lo-256	256	-	945M	130	✓	0.95	22.07	0.649
CATOK-B-256	256	16	224M	80	×	1.17	<u>22.10</u>	<u>0.666</u>
CATOK-L-32	32	16	552M	160	×	2.03	17.85	0.465
CATOK-L-256	256	16	552M	160	×	<u>0.84</u>	**22.72**	**0.681**

probability 0.1 during training to enable Classifier-Free Guidance (CFG). At inference, we adopt a CFG schedule following (Chang et al., 2023; Li et al., 2024b; Wen et al., 2025) without temperature sampling. Additional details are provided in Appendix B.

5.2 RECONSTRUCTION

We report reconstruction FID (Heusel et al., 2017) (distributional dissimilarity), PSNR (pixel-wise MSE), and SSIM (Wang et al., 2004) (perceptual similarity) on the ImageNet-1K validation set at 256 × 256 resolution. We evaluate three variants: CATOK-B with 256 1D tokens, CATOK-L with 32 tokens and CATOK-L with 256 tokens. The results are compared against state-of-the-art variants with comparable latent spaces and model sizes. As shown in Tab. 1, among diffusion autoencoders, CATOK-L-256 achieves superior PSNR and SSIM, with SSIM significantly outperforming the 945M FlowMo-Lo-256, while also reaching competitive rFID with less than half the training epochs copared with Semanticist-L-256. Remarkably, CATOK-B-256 attains comparable results with only 80 epochs, demonstrating the high training efficiency of CATOK.

Notably, CATOK-L-256† achieves the best PSNR and SSIM among one-step 1D tokenizers, demonstrating its flexibility in sampling: it supports fast one-step sampling while also benefiting from multi-step sampling for improved reconstruction. However, although its rFID surpasses VQGAN (Esser et al., 2021)—a classic 2D tokenizer—by three points, it still lags behind modern tokenizers. This gap arises because those methods rely on more challenging objectives (*e.g.*, GAN (Goodfellow et al., 2014) loss) and complex training recipes (Yu et al., 2024c; Miwa et al., 2025), whereas CATOK is not specifically optimized for one-step sampling but attains comparable results as a byproduct.

Table 2: **Class-conditional generation results on ImageNet-1K 256 × 256 benchmark.** "#Param." denotes the parameters of generator, "Token" and "Step" indicates the number of tokens and steps used for generation, respectively. "↓" or "↑" denote lower or higher values are better.

Method	Generator	#Param.	Token	Step	gFID↓	IS↑
2D autoregressive models						
VQGAN	Tam. Trans.	1.4B	256	256	15.78	74.3
RQ-VAE	RQ-Trans.	3.8B	256	68	7.55	134.0
Causal MAR	MAR-L	481M	256	256	4.07	232.4
LlamaGen	LlamaGen-L	343M	256	256	3.80	248.3
VAR	VAR-d16	310M	680	10	3.30	274.4
1D masked-prediction models						
FlowMo-Lo-256	MaskGiT-L	227M	256	-	4.30	274.0
TiTok-L-32	MaskGiT-L	227M	32	8	2.77	194.0
TiTok-S-128	MaskGiT-L	227M	128	64	1.97	281.8
1D autoregressive models						
FlexTok d12-d12	AR Trans.	1.3B	32	32	3.83	-
SpectralAR-64	VAR	310M	64	64	3.02	282.2
Semanticist-L-256	ϵLlamaGen-L	343M	32	32	2.57	260.9
CaTok-L-32	ϵLlamaGen-L	343M	32	32	3.40	288.6

Table 3: **Ablation on technique designs.**

(a) **Ablation on training recipe.** FID@n: n-step sampling.

Method	rFID@1	rFID@25
Naïve Decoder ($\mathcal{L}_{RF}$ in Eq. (13))	183.69	1.81
+ $\mathcal{L}_{MF}$ in Eq. (12)	4.71	1.90
+ $\mathcal{L}_{REPA}$ in Eq. (14)	4.31	1.71
+ $\mathcal{L}_{REPA-A}$ in Eq. (15)	3.92	1.15
+ Selecting tokens in $[r, t]$	4.89	1.17

(b) **Ablation on the causality and balance of 1D visual tokens.**

Select	Token	rFID	gFID
$[r, t]$	256	1.17	4.91
All	256	1.15	13.54
First k	256	1.37	9.21
First k	128	5.32	7.49

5.3 Autoregressive generation

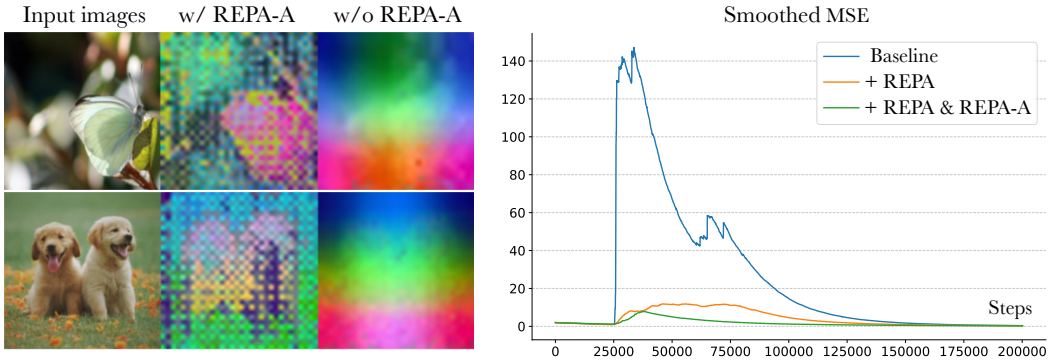
Following common prective (Li et al., 2024b), we report generation FID and IS (Salimans et al., 2016) (image quality and class diversity) with evaluation suite provided by Dhariwal & Nichol (2021). For fair comparison and efficient training, we train ϵLlamaGen-L, *i.e.*, standard LlamaGen (Sun et al., 2024) with the diffusion loss (Li et al., 2024b) modified by Wen et al. (2025), for 400 epochs. As shown in Tab. 2, CaTok attains comparable gFID and IS scores to the state-of-the-art tokenizers, with far fewer tokenization training epochs (160 *vs.* 300+), demonstrating its capability to learn 1D causal tokens well-suited for standard autoregressive modeling. We present qualitative visualizations in Fig. 4. It is worth noting that training a state-of-the-art visual generative model is computationally expensive and beyond the scope of this work. Instead, we focus on building a 1D tokenizer that captures visual causality and validating its advantages on AR modeling under fair comparison.

5.4 Ablation study

We conduct ablation studies on the smaller CaTok-B-256 models, training for 80 epochs on both reconstruction and generation tasks.

Improved training recipe. We present a roadmap from the conventional diffusion autoencoder with naïve decoder to our CaTok step by step in Tab. 3a. Traditional DiT decoders lack one-step sampling capability, but equipping them with the MeanFlow objective enables reasonable one-step results. Both REPA and REPA-A accelerate convergence and enhance performance. Moreover, optimizing MeanFlow objective on 1D tokens selected from a time interval $[r, t]$ allows the model to learn visual causality, at the cost of a slight performance drop.

Figure 4: **Qualitative Results.** 256×256 generated images on ImageNet-1K with CATOK-L-32.



a) PCA visualization of encoder features

b) Training with REPA-A improves convergence

Figure 5: **Effectiveness of our REPA-A.**a) We apply principal component analysis (PCA) to visualize image features from the CATOK encoder. b) Training curves of the smoothed MSE between prediction and target, with the MeanFlow loss ($\mathcal{L}_{MF}$) added at 25K steps.

Causality and balance matter in AR modeling. We evaluate three variants of 1D token selection: (1) selecting tokens within an interval $[r, t]$ (our default setting); (2) selecting all tokens; and (3) selecting the first k tokens. For the third variant, we train two AR models using either all 256 tokens or only the first 128 tokens. As shown in Tab. 3b, CATOK achieves the best gFID. Non-causal tokens hinder AR modeling, and, consistent with Bachmann et al. (2025); Wen et al. (2025), imbalance reduces the contribution of later tokens—an issue that CATOK fundamentally addresses without requiring additional re-weighting mechanism (Wang et al., 2025a).

REPA-A stabilizes training and improves performance. As shown in Fig. 5 a), REPA-A makes encoder features more informative and discriminative, helping the registers capture richer content. In Fig. 5 b), REPA-A mitigates the loss spike at 25K steps when the MeanFlow loss is introduced, stabilizing decoder training and improving overall performance.

6 CONCLUSION

We presented CATOK, a novel 1D causal image tokenizer to bridge the gap between autoregressive language models and vision models. By binding the average velocity field in the MeanFlow objective to the corresponding 1D token segments, we enabled the diffusion autoencoder to learn visual causality along the flow path while supporting one-step sampling. Furthermore, we proposed an advanced regularization method REPA-A, which effectively stabilized and accelerated the training of the autoencoder. Experiments demonstrated that we achieved state-of-the-art PSNR and SSIM on ImageNet reconstruction, and obtaining comparable results on the class-conditional generation.

REPRODUCIBILITY STATEMENT

We provide hyperparameter details in Sec. 5 and Appendix B. We will also release the codebase and model checkpoints to reproduce the results in the paper.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1

Roman Bachmann, Jesse Allardice, David Mizrahi, Enrico Fini, Oğuzhan Fatih Kar, Elmira Amirloo, Alaaeldin El-Nouby, Amir Zamir, and Afshin Dehghan. Flextok: Resampling images into 1d token sequences of flexible length. In *ICML*, 2025. 2, 4, 6, 9

Yutong Bai, Xinyang Geng, Karttikeya Mangalam, Amir Bar, Alan L Yuille, Trevor Darrell, Jitendra Malik, and Alexei A Efros. Sequential modeling enables scalable learning for large vision models. In *CVPR*, 2024. 2

Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. In *ICML*, 2023. 7, 14

Yinbo Chen, Rohit Girdhar, Xiaolong Wang, Sai Saketh Rambhatla, and Ishan Misra. Diffusion autoencoders are scalable image tokenizers. *arXiv preprint arXiv:2501.18593*, 2025. 6

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 1

Jiaxi Cui, Munan Ning, Zongjian Li, Bohua Chen, Yang Yan, Hao Li, Bin Ling, Yonghong Tian, and Li Yuan. Chatlaw: A multi-agent collaborative legal assistant with knowledge graph enhanced mixture-of-experts large language model. In *ICLR*, 2024. 2, 4, 6

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *ICLR*, 2024. 2, 4, 6

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 6

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. 8

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. 4, 6

Shivam Duggal, Phillip Isola, Antonio Torralba, and William T Freeman. Adaptive length image tokenization via recurrent allocation. In *ICLR*, 2025. 6

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 2, 6, 7

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. *arXiv preprint arXiv:2505.13447*, 2025. 1, 2, 3, 4, 5, 6

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. In *NeurIPS*, 2014. 7

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 1

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 *CVPR*, 2025. 2, 6

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. 7

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2

Yuanhui Huang, Weiliang Chen, Wenzhao Zheng, Yueqi Duan, Jie Zhou, and Jiwen Lu. Spectralar: Spectral autoregressive visual generation. In *ICCV*, 2025. 6

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. *arXiv preprint arXiv:2501.07730*, 2025. 6

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 5, 6

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *CVPR*, 2022. 6

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. 2, 5

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 6

Tianhong Li, Dina Katabi, and Kaiming He. Return of unconditional generation: A self-supervised representation generation method. In *NeurIPS*, 2024a. 6

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. In *NeurIPS*, 2024b. 2, 5, 6, 7, 8, 14

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *ICLR*, 2022. 2, 3, 5

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. 1

Xingchao Liu, Chengyue Gong, et al. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2022. 2, 3, 5

Keita Miwa, Kento Sasaki, Hidehisa Arai, Tsubasa Takahashi, and Yu Yamaguchi. One-d-piece: Image tokenizer meets quality-controllable compression. *arXiv preprint arXiv:2501.10064*, 2025. 6, 7

OpenAI. Sora system card. `https://openai.com/index/sora-system-card/`, 2024. 2

OpenAI. Gpt-5 system card. `https://openai.com/index/gpt-5-system-card/`, 2025. 1

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*, 2024. 6

Kaihang Pan, Wang Lin, Zhongqi Yue, Tenglong Ao, Liyu Jia, Wei Zhao, Juncheng Li, Siliang Tang, and Hanwang Zhang. Generative multimodal pretraining with discrete diffusion timestep tokens. In *CVPR*, 2025. 2, 6

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 4, 6

Yansong Peng, Kai Zhu, Yu Liu, Pingyu Wu, Hebei Li, Xiaoyan Sun, and Feng Wu. Flow-anchored consistency models. *arXiv preprint arXiv:2507.03738*, 2025. 4

Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *CVPR*, 2022. 2, 3, 4, 6

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021. 2

Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In *NeurIPS*, 2019. 2, 6

Oren Rippel, Michael Gelbart, and Ryan Adams. Learning ordered representations with nested dropout. In *ICML*, 2014. 2, 6

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NeurIPS*, 2016. 8

Kyle Sargent, Kyle Hsu, Justin Johnson, Li Fei-Fei, and Jiajun Wu. Flow to the mode: Mode-seeking diffusion autoencoders for state-of-the-art image tokenization. In *ICCV*, 2025. 2, 6

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021. 2

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. 2, 6, 8

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. In *NeurIPS*, 2024. 2, 6

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *NeurIPS*, 2017. 6

Bohan Wang, Zhongqi Yue, Fengda Zhang, Shuo Chen, Li'an Bi, Junzhe Zhang, Xue Song, Kennard Yanting Chan, Jiachun Pan, Weijia Wu, et al. Selftok: Discrete visual tokens of autoregression, by diffusion, and for reasoning. *arXiv preprint arXiv:2505.07538*, 2025a. 2, 6, 9

Junke Wang, Zhi Tian, Xun Wang, Xinyu Zhang, Weilin Huang, Zuxuan Wu, and Yu-Gang Jiang. Simplear: Pushing the frontier of autoregressive visual generation through pretraining, sft, and rl. *arXiv preprint arXiv:2504.11455*, 2025b. 2

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. 2

XuDong Wang, Xingyi Zhou, Alireza Fathi, Trevor Darrell, and Cordelia Schmid. Visual lexicon: Rich image features in language space. In *CVPR*, 2025c. 6

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 2004. 7

Xin Wen, Bingchen Zhao, Ismail Elezi, Jiankang Deng, and Xiaojuan Qi. "principal components" enable a new language of images. In *ICCV*, 2025. 2, 4, 6, 7, 8, 9, 14

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*, 2025. 2

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 1

Ruihan Yang and Stephan Mandt. Lossy image compression with conditional diffusion models. In *NeurIPS*, 2023. 2, 4, 6

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *CVPR*, 2025. 5

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*, 2024a. 2, 6

Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Randomized autoregressive visual generation. *arXiv preprint arXiv:2411.00776*, 2024b. 2

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 *NeurIPS*, 2024c. 2, 6, 7

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*, 2024d. 2, 5

Kaiwen Zha, Lijun Yu, Alireza Fathi, David A Ross, Cordelia Schmid, Dina Katabi, and Xiuye Gu. Language-guided image tokenization for generation. In *CVPR*, 2025. 6

A Architecture of AR modeling

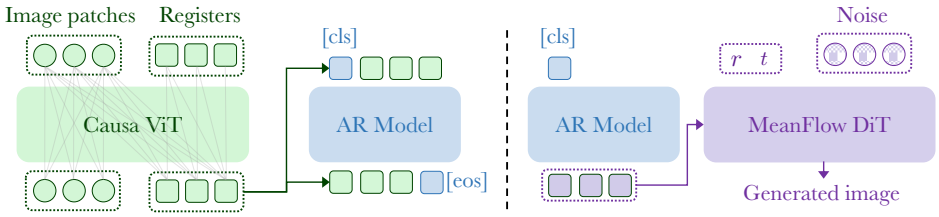


Figure 6: **Architecture of AR modeling.**

We illustrate AR training and sampling in Fig. 6, and visualize the causal mask mechanism in ViT in Fig. 6 a). After training CaTok, we freeze the encoder to extract 1D tokens. During AR training stage, these tokens are optimized with a class token prefix using teacher forcing under a diffusion loss (Li et al., 2024b). At sampling time, we input a learned class token, the AR model predicts the corresponding visual 1D tokens, and these tokens are then conditioned to the decoder for generation.

B More implementation details

Table 4: Detailed configuration of CaTok-B and CaTok-L for tokenization and AR modeling.

Training Config	CaTok-B	CaTok-L	AR modeling
Optimizer	AdamW		
Peak learning rate	1×10^{-4}		5×10^{-5}
Minimum learning rate	0		
Learning rate schedule	cosine decay		constant
Batch size	1024		2048
Weight decay	0.05		
Epochs	80	160	400
Warmup epochs	0		96
Gradient clipping	3.0		
EMA	0.999		

Training setup follows Wen et al. (2025), with detailed hyperparameters in Tab. 4. For reconstruction, we disable CFG in one-step sampling, and apply CFG with a scale of 2.0 in 25-step sampling. For 80-epoch training, we introduced the MeanFlow objective at epoch 10 and the selecting mechanism at epoch 40; for 160-epoch training, these corresponded to epochs 20 and 80, respectively. For generation, we do not use CFG with CaTok, and the CFG of AR model is the same as MUSE (Chang et al., 2023), MAR (Li et al., 2024b) and Semanticist (Wen et al., 2025), whici tunes down the guidance scale of small-indexed tokens to improve the diversity of generated sample.

C The use of Large Language Models (LLMs)

LLMs are utilized for language refinement and are not involved in any other component of this work.