ADAPTING SELF-SUPERVISED REPRESENTATIONS AS A LATENT SPACE FOR EFFICIENT GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce **Rep**resentation **Tok**enizer (RepTok), a generative modeling framework that represents an image using a single continuous latent token obtained from self-supervised vision transformers. Building on a pre-trained SSL encoder, we fine-tune only the semantic token embedding and pair it with a generative decoder trained jointly using a standard flow matching objective. This adaptation enriches the token with low-level, reconstruction-relevant details, enabling faithful image reconstruction. To preserve the favorable geometry of the original SSL space, we add a cosine-similarity loss that regularizes the adapted token, ensuring the latent space remains smooth and suitable for generation. Our single-token formulation resolves the spatial redundancies of the 2D latent space and significantly reduces training costs. Despite its simplicity and efficiency, RepTok achieves competitive results on class-conditional ImageNet generation and extends naturally to text-to-image synthesis, reaching competitive zero-shot performance on MS-COCO under extremely limited training budgets. Our findings highlight the potential of fine-tuned SSL representations as compact and effective latent spaces for efficient generative modeling. We will release our code to facilitate further research.

1 INTRODUCTION

In recent years, diffusion- (Ho et al., 2020; Kingma et al., 2021; Song & Ermon, 2019) and flow-based (Lipman et al., 2023; Liu et al., 2023b; Ma et al., 2024) models have emerged as powerful generative modeling frameworks, capable of synthesizing high-quality images (Ramesh et al., 2022; Rombach et al., 2022; Dhariwal & Nichol, 2021) and videos (Ho et al., 2022). However, these models typically come with substantial computational demands since they regress vector fields in the high-dimensional pixel space of images. Latent Diffusion Models (Rombach et al., 2022) address this challenge by decomposing the generative modeling task into two stages. By first compressing images into a lower-dimensional latent space via a pre-trained Variational Autoencoder (Kingma et al., 2013), LDMs abstract away imperceptible details, enabling the generation process to solely focus on semantic content and drastically reducing computational costs during training and inference (Esser et al., 2021; 2024; Fuest et al., 2024; Schusterbauer et al., 2024). However, despite these computational advantages, the latent space is still organized in a two-dimensional grid structure, which fails to exploit the high spatial redundancies inherent to natural images.

Recent efforts have sought to improve latent generative paradigms along two directions. *TiTok* (Yu et al., 2024a) tries to exploit spatial redundancies and replaces the default 2D spatial grid in latent diffusion with a transformer-based encoder–decoder that represents images as 1D latent sequences, achieving compact encodings with as few as 32 discrete tokens. In parallel, *REPA* (Yu et al., 2024b) leverages the rich representations of pre-trained self-supervised learning (SSL) models to accelerate the convergence of latent diffusion models, by distilling the semantic knowledge into the diffusion model via a cosine similarity loss between their respective feature representations.

In this work, we extend these two directions by exploring more powerful uses of SSL representations. While REPA accelerates training primarily through feature alignment on the 2D spatial grid, we demonstrate that self-supervised models can be leveraged more directly: with minimal but crucial fine-tuning, pooled 1D SSL representations themselves constitute effective latent spaces for generative modeling. These representations exhibit smooth, semantically structured geometry that is well-suited for generation, while simultaneously eliminating the spatial redundancies inherent in 2D grid-based

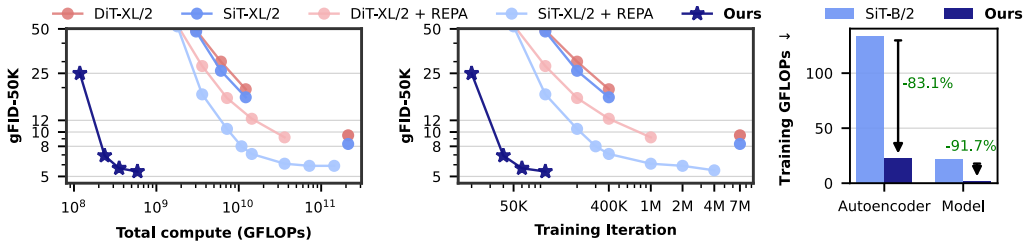

Figure 1: Comparison of our single-token MLP-Mixer generator against transformer-based baselines (DiT, SiT), as well as representation-aligned models like REPA. RepTok attains competitive generative performance while reducing training cost by over 90% owing to its compact latent space and lightweight architecture. All results reported without CFG. For fair comparison, we employ an encoder and decoder trained on general-domain data.

latents. Specifically, we show that the pooled 1D output from the `[cls]` token alone provides a compact yet expressive representation that not only captures high-level semantics but also preserves sufficient spatial detail to enable high-fidelity reconstruction.

Our **Rep**resentation **Tok**enization approach, termed RepTok, builds on a pre-trained SSL encoder that is lightly fine-tuned and trained jointly with a generative decoder. We train the decoder with a standard flow matching objective, complemented by a cosine-similarity loss that regularizes the latent representation to remain close to its original smooth and semantically structured space, which is well-suited for generation. Without auxiliary perceptual (Zhang et al., 2018) or adversarial (Esser et al., 2021) losses, the resulting model is able to faithfully decode the single-token latent representation into the pixel space. This design enables highly efficient image synthesis training, allowing us to use simple, attention-free architectures such as MLP-Mixers (Tolstikhin et al., 2021) for accelerated ImageNet training (see Figure 1). Furthermore, we show that the framework naturally extends to text-to-image (T2I) synthesis: by incorporating cross-attention to integrate textual conditioning, our model achieves competitive zero-shot performance on the COCO (Lin et al., 2014) benchmark under an extremely constrained training budget (see Figure 7). We state our contributions as follows:

- We show that self-supervised vision transformers can be used more powerfully than just guiding generative training: with minimal adaptation of the semantic token, their smooth and semantically structured latent spaces can directly act as encoders for generative modeling. By injecting the necessary fine-grained information into this semantic token, we enable faithful reconstruction while simultaneously eliminating the spatial redundancies inherent in 2D grid-based latents. Coupled with a generative decoder, this setup allows accurate image reconstruction from a single continuous token.
- Exploiting this autoencoder design, we introduce a lightweight and optionally attention-free pipeline for latent generative modeling. This drastically reduces training compute while preserving quality, achieving competitive ImageNet generation at a fraction of the cost of transformer-based diffusion baselines.
- We show that RepTok scales effectively to text-to-image synthesis, achieving competitive zero-shot results on MS-COCO with under 20 hours of training on four A100 GPUs.

2 Related Work

Latent space generation Early approaches such as PixelVAE and VQVAE(Gulrajani et al., 2016; Razavi et al., 2019; Van Den Oord et al., 2017) demonstrated that generative modeling within compact latent spaces significantly improves sampling quality and efficiency. VQGAN (Esser et al., 2021) integrates vector-quantized variational autoencoders with adversarial losses to construct discrete latent codebooks. Subsequently, these discrete tokens are leveraged by autoregressive transformers for image generation tasks. Latent Diffusion Models (LDMs) (Rombach et al., 2022) brought this concept into the diffusion models, operating in learned spatial latent spaces that preserve semantic content and abstract away perceptual detail. This approach has since become foundational across modalities including images (Peebles & Xie, 2023; Ma et al., 2024; Pernias et al., 2024), audio (Liu et al., 2023a), and video (Ho et al., 2022; Blattmann et al., 2023b;a; Kong et al., 2024).

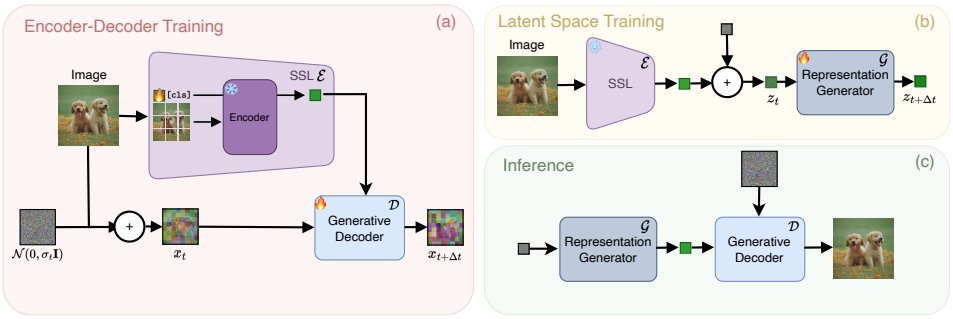

Figure 2: **Overview of our pipeline**. (a) Joint fine-tuning of the `[cls]` token of SSL encoder $\mathcal{E}$ and training of the generative decoder $\mathcal{D}$ for image reconstruction. (b) Training of the generation model $\mathcal{G}$ to synthesize frozen encoder outputs, which constitute the latent space $z = \mathcal{E}(x)$. (c) Inference pipeline, where the latent space z is first generated and subsequently decoded into the pixel space.

Pre-trained representations in diffusion models Leveraging pre-trained representations has been shown to improve image generation. REPA (Yu et al., 2024b) accelerates diffusion training by aligning diffusion features with DINO embeddings, with Wang et al. (2025) noting that careful scheduling is required for effective training. Closely related to our approach is RCG (Li et al., 2024), which employs a two-stage pipeline: first generating a predefined semantic representation and then transporting it to the pixel space. However, RCG primarily targets unconditional synthesis and thus leaves the representation space unchanged. In contrast, our objective is faithful reconstruction and generation, similar to the role of the latent space in VAEs. This requires not only semantic but also low-level visual information. We address this by injecting fine-grained details into the representation space, enabling both faithful reconstruction and high generative performance. Concurrent works like RAE (Zheng et al., 2025) and SVG (Shi et al., 2025) use the full spatial grid of SSL features, thus operating in a high-dimensional structured latent space. Our approach is fundamentally different: RepTok relies solely on the pooled semantic token, discarding all spatial tokens and learning to represent an image with a single compact vector. This yields a much more aggressive compression. Additionally, a key difference to SVG whose objective is aligning to the SSL representation, we directly employ the pooled semantic output as the latent representation itself.

Global information latent spaces Recent work has explored 1D tokenization beyond spatial grid latents. TiTok (Yu et al., 2024a) encodes images into compact sequences of as few as 32 discrete tokens with a ViT encoder, enabling efficient autoregressive generation. ElasticTok (Yan et al., 2024) extends this idea with adaptive token counts per frame, while FlexTok (Bachmann et al., 2025) introduces variable-length ordered tokens for coarse-to-fine generation. Our approach differs in the following key aspects: First, we operate in a continuous latent space, avoiding quantization and enabling fully differentiable diffusion training. Second, we directly exploit the `[cls]` token of SSL vision transformers as a compact latent, yielding smooth and semantically structured manifolds. Unlike discrete tokenizers, Diffusion Autoencoders (Preechakul et al., 2022) extract semantic information into a continuous latent space and utilize a jointly trained diffusion model for reconstruction. As the latent space is mostly semantic, image reconstruction requires an additional subcode x_T, obtained by mapping the image back to the Gaussian noise space using conditional DDIM sampling (Song et al., 2020). By contrast, our method reconstructs images faithfully from a single latent z alone. A concurrent work, AToken (Lu et al., 2025), proposes a unified visual tokenizer designed to operate consistently across multiple modalities.

3 Method

3.1 Preliminaries

Flow Matching models learn vector fields that map between two terminal distributions: $p(x_0)$, typically a simple prior distribution such as a standard Gaussian distribution, and $p(x_1)$, the target data distribution. Let $\mathbb{R}^d$ be the space that x_0 and x_1 reside in, and let $v_\theta(t, x)$ represent the time-dependent vector field to be learned with $t \in [0, 1]$. The underlying dynamics of flow matching

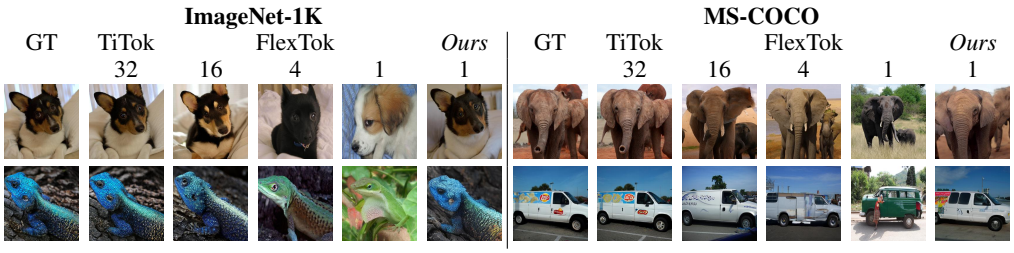

Figure 3: We introduce *RepTok*, a compact visual tokenizer that builds upon pre-trained SSL representations. Our approach augments these representations with additional necessary information to enable images to be faithfully encoded as a single continuous token, which allows for both high-fidelity image reconstruction and synthesis. The third row indicates the number of tokens for reconstruction.

models are then governed by the ordinary differential equation (ODE) $dx = v_\theta(x, t)$. A common choice for the interpolant between x_0 and x_1 is the linear interpolant (Liu et al., 2023b), defined as $x_t = tx_1 + (1-t)x_0$. The vector field v_θ can then be optimized using the following training objective with a randomly sampled t and the corresponding x_t (Lipman et al., 2023; Liu et al., 2023b; Schusterbauer et al., 2025):

$$\mathcal{L} = \mathbb{E}_{t,x_0,x_1}||v_\theta(x_t, t) - (x_1 - x_0)||. \tag{1}$$

To sample from a flow matching model, one simply integrates along the trajectory defined by the learned ODE. This can be accomplished using numerical integration techniques such as the forward Euler method, with the update rule given by $x_{t+t_\Delta} = x_t + t_\Delta \mathbf{v}_\theta(x_t, t)$, where $\forall t \in [0, 1)$, $t_\Delta = 1/N$, and N being the number of function evaluations (NFE).

3.2 RepTok: Representing Images As a Single Token

TiTok (Yu et al., 2024a) represents a significant advancement over traditional VAEs by overcoming their inherent 2D tokenization grid constraints. Unlike conventional approaches, where each token is restricted to attending only to a fixed image grid, TiTok enables tokens z to attend freely to the entire image. However, despite these improvements, TiTok typically still relies on multiple tokens to effectively encode an image. In this work, we show that continuous latent spaces can achieve even greater efficiency in few-token regimes. Specifically, we demonstrate that a single continuous token, derived from a pre-trained encoder, can be used together with a generative decoder to synthesize high-fidelity reconstructions.

Finetuned Self-supervised Models are Faithful Encoders It is well established that models such as CLIP (Radford et al., 2021), MAE (He et al., 2022) and DINO (Caron et al., 2021; Oquab et al., 2024) models encode highly informative representations and demonstrate a strong understanding of images, as evidenced by their effectiveness in various downstream tasks, including image classification (Radford et al., 2021; Caron et al., 2021; Oquab et al., 2024) and semantic segmentation (Zhang et al., 2023). This capability is further demonstrated by the existence of unCLIP models (Ramesh et al., 2022; Rombach et al., 2022), which can generate image variations from noise using only a single CLIP embedding. While this observation confirms that generative models can synthesize images from extremely compact bottlenecks (for unCLIP (Ramesh et al., 2022) $z \in \mathbb{R}^{1\times512}$), we hypothesize that the variations of the outputs arise from the fact that CLIP models are not explicitly trained to preserve exact pixel locations but instead optimize a contrastive loss with corresponding textual descriptions, thereby capturing only high-level semantic features.

Motivated by these observations, we explore and unlock the potential of leveraging a pretrained encoder $\mathcal{E}$ that already possesses a comprehensive understanding of image content. To this end, we introduce a novel training strategy that leverages a pretrained self-supervised learning (SSL) model with a transformer-based architecture as the encoder. These models typically incorporate a class token (typically referred to as the `[cls]` token) that is trained, either explicitly or implicitly, to aggregate information from image patches. However, such pretrained models are often optimized for downstream tasks and may consequently, as an example, *underrepresent* low-level visual details critical for image reconstruction. To address this limitation, we propose a targeted adaptation strategy that *only* updates the class token embedding while keeping the remainder of the encoder frozen. Remarkably, we find that this minimal intervention is sufficient to inject the necessary visual detail

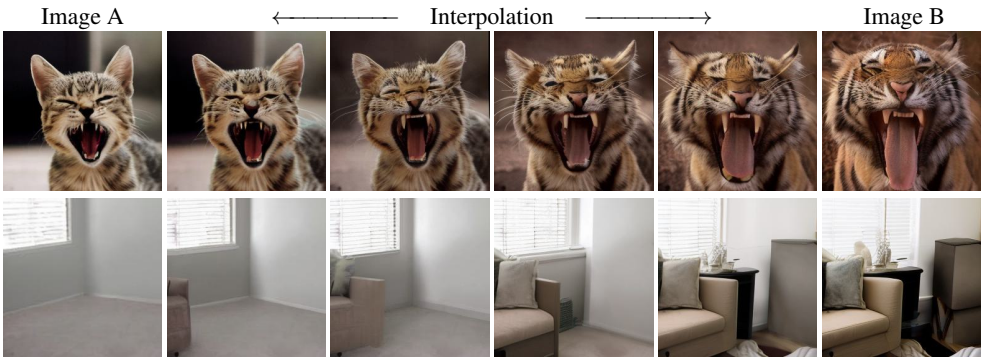

Figure 4: **Latent space interpolation.** We observe smooth transitions not only in semantic content but also in spatial configuration. This indicates that our method successfully integrates low-level spatial information while preserving the properties of the pretrained encoder's latent space, and facilitates generation within the learned representation. We provide more samples in the Appendix.

into the representation. Empirical results reveal that with only the class token being fine-tuned, the system is capable of producing reconstructions with high fidelity across a range of SSL backbones including DINOv2 (Oquab et al., 2024), MAE (He et al., 2022) and CLIP (Radford et al., 2021). We demonstrate our reconstructions in Figure 3.

Training the Encoder together with a Generative Decoder While the SSL-pretrained encoder $\mathcal{E}$ remains largely frozen, a supervisory signal is still required to inject reconstruction-relevant information into the class token. Additionally, a decoder is necessary to map the resulting single-token latent representation back into pixel space. To this end, we jointly train the encoder $\mathcal{E}$ and a generative decoder $\mathcal{D}$ in a continuous manner, using a simple but effective flow matching loss.

The generative decoder $\mathcal{D}$ is trained end-to-end alongside the encoder $\mathcal{E}$ to learn a mapping from randomly sampled Gaussian noise ϵ to the target image x. We follow principles similar to the conditioning mechanism employed in MMDiT (Esser et al., 2024) and concatenate the latent token $z = \mathcal{E}(x)$ with the noisy image tokens. The resulting training objective is formulated as a flow matching loss as in Equation (1), which optimizes both the encoder and the decoder:

$$\mathcal{L} = \mathbb{E}_{t,x_0,x_1} ||v_\theta(t, x_t, z) - (x_1 - x_0)||. \tag{2}$$

To improve computational efficiency and remain consistent with the SiT framework (Ma et al., 2024), we adopt a pretrained 2D VAE (Rombach et al., 2022) so that the generative decoding process operates within a learned latent space rather than directly in pixel space.

Cosine-Similarity Loss We observe that the `[cls]` tokens of self-supervised vision encoders already provide a smooth, semantically structured space. Hence, our goal during training is to maintain this well-regularized space while still allowing the token to integrate the fine-grained information the decoder needs for faithful reconstructions. Freezing the `[cls]` token leads to poor reconstruction quality, as indicated in Figure 5. Conversely, leaving the encoder completely unconstrained pulls the token far away from the well-regularized space, removing the prerequisite for later generative modeling. We find that only unfreezing the `[cls]` while fixing all other encoder weights strikes a good balance between integration of more information and maintaining the original regularization. To constrain the token from deviating its pre-trained representation, we introduce a cosine-similarity alignment term

Figure 5: **Fine-tuning the `[cls]` token**. From left: GT, frozen, finetuned.

$$\mathcal{L}_{\cos}(x) = \lambda(1 - \cos(z, z_{\text{frozen}})) \qquad z_{\text{frozen}} = \mathcal{E}_{\text{frozen}}(x),\ z = \mathcal{E}(x), \tag{3}$$

where z_{frozen} is the token output from the frozen SSL model, z is the fine-tuned counterpart, and λ explicitly controls the allowed deviation. Reducing λ relaxes the constraint; increasing it restricts the

token more tightly to its source. With this alignment mechanism, we retain the well-behaved SSL latent space for later generative modeling, while additionally enriching the token with the additional information the generative decoder needs to faithfully reconstruct. We observe that incorporating the cosine similarity loss prevents the embedding from drifting away from the well-regularized latent space, also under extended training, as illustrated in Figure 9. We directly condition the generative decoder on those representations and focus on preserving their structured properties while injecting additional information to enable both faithful reconstruction and generative abilities.

3.3 Single Token Generation for Image Synthesis

Since RepTok projects images into a continuous latent space z (typically in $\mathbb{R}^{1\times 768}$), it becomes feasible to model and sample from this space using a separate generative model $\mathcal{G}$. To this end, we again employ flow matching (Lipman et al., 2023) for latent space generation. We discover that utilizing a frozen SSL model, with only the class token finetuned, provides an effective alternative regularization mechanism to the conventional approaches using Kullback-Leibler (KL) divergence (Rombach et al., 2022) or vector quantization (Austin et al., 2021; Yu et al., 2024a; Tian et al., 2024). By preserving the structural properties of the learned feature space, the frozen encoder inherently constrains the latent representations and facilitates the generation process without requiring explicit KL or vector quantization regularization.

Attention-free ImageNet Generation Typical diffusion models operate on high-dimensional image or latent spaces consisting of multiple tokens, where capturing global structure and local detail relies on modeling interactions across tokens. This is commonly achieved through attention (Vaswani et al., 2017). While effective, it introduces significant computational overhead. In contrast, when inputs are aggressively compressed into a single token, token-to-token interactions become unnecessary. We show that in this highly compressed regime, generative modeling can be effectively performed using an attention-free, pure MLP-based architecture such as MLP-Mixer (Tolstikhin et al., 2021). Despite its architectural simplicity and lack of self-attention, our MLP-only approach performs remarkably well. This highlights a novel and computationally efficient approach to diffusion modeling, where architectural complexity is shifted to the pre-trained compression stage without sacrificing flexibility or generality. For text-to-image synthesis, we still use attention for text conditioning, but the compactness of our latent space keeps the associated cost minimal. In particular, because the number of tokens in our latent space is small, the quadratic scaling of attention remains inexpensive.

4 Experiments

We evaluate RepTok on class-conditional ImageNet-1k (Deng et al., 2009) and show the scalability of our approach on text-to-image (T2I) generation. We evaluate reconstruction performance with reconstruction FID (*rFID*), PSNR, SSIM, and LPIPS, and generation performance with generation FID (*gFID*), consistent with prior work (Bachmann et al., 2025; Yu et al., 2024a). All models operate at 256^2 resolution; implementation and training details are provided in the Appendix.

4.1 Class-conditional Generation

We jointly train the SSL encoder (only the `[cls]` token parameters are trainable) and a generative flow matching-decoder for reconstruction in a first stage. We use DINOv2 (Oquab et al., 2024) as our SSL encoder, but show in section 4.3 that our method also generalizes to other SSL methods. For latent space synthesis, we train a lightweight, attention-free generator (MLP-Mixer) over the continuous `[cls]` token, where we encode images using the previously trained SSL encoder model. We inject class information by concatenating a learned class embedding, which we randomly drop during training to enable classifier-free guidance (Ho & Salimans, 2021).

Quantitative Comparison Table 3 compares our method against recent, state-of-the-art transformer-based generative models on ImageNet 256×256. For each model, we report the FID score, number of training iterations, parameter count, per-iteration compute in GFLOPs, and the resulting total training compute in Peta-FLOPs. FLOPs are estimated from a single forward pass (batch size 1), and

Table 1: Reconstruction performance on ImageNet 256^2.

	FID@50K ↓	PSNR ↑
RCG	3.20	9.31
Ours	**1.85**	**14.94**

Table 2: State-of-the-art comparison between tokenizers for reconstruction and class-conditional ImageNet generation. † metrics sourced from (Bachmann et al., 2025).

Tokenizer	# tokens	global	continuous	rFID	gFID
LDM (Rombach et al., 2022)	32x32	✗	✓	0.90	7.76
LlamaGen† (Sun et al., 2024)	16x16	✓	✗	2.19	3.06
TiTok-L† (Yu et al., 2024a)	32	✓	✗	2.21	2.77
TiTok-B† (Yu et al., 2024a)	64	✓	✗	1.70	2.48
TiTok-S† (Yu et al., 2024a)	128	✓	✗	1.71	1.97
FlexTok† `d12-d12` (Bachmann et al., 2025)	1-256	✓	✗	4.20	3.83
FlexTok† `d18-d18` (Bachmann et al., 2025)	1-256	✓	✗	1.61	2.02
FlexTok† `d18-d28` (Bachmann et al., 2025)	1-256	✓	✗	1.45	1.86
RepTok (*ours*)	1	✓	✓	1.85	1.88

scaled linearly with the effective batch size and the number of training steps; we follow the convention of counting only the forward pass. Our model achieves highly competitive FID scores while requiring significantly less total compute than other baselines such as DiT and SiT. We note that classifier-free guidance (CFG) yields only limited improvements in our setting, a phenomenon also reported by RCG. Table 2 compares RepTok with spatial and 1D tokenizers for both reconstruction and class-conditional generation on ImageNet. Despite using just *one* continuous token, RepTok matches or even outperforms several spatial and non-spatial baselines in rFID while remaining competitive on gFID relative to recent discrete tokenizers. Additional results in Table 1 compare RepTok to RCG (Li et al., 2024), a method which relies on purely semantic codes. RepTok achieves significantly higher PSNR and lower FID, indicating that our continuous token preserves more information than pure semantics and delivers stronger performance across both perceptual and pixel-level metrics.

Efficiency We measure training compute in floating point operations (FLOPs). In the single-token latent space, token-to-token interactions are unnecessary. We therefore adopt a pure MLP-Mixer as the latent space generator model. The combination of representing an image with a single token and the MLP-only architecture reduces training FLOPs by an order of magnitude compared to attention-based diffusion in latent space, as shown in Figure 1. Despite a comparable number of parameters across both models, our approach still achieves a substantially lower computational footprint, requiring only 1.7% of the FLOPs consumed by SiT (Ma et al., 2024). Our overall FLOPs remain significantly lower, also when accounting for the inference cost of the corresponding first-stage encoder.

Qualitative Comparison Figure 3 shows high-fidelity reconstructions from a single token on ImageNet validation images and strong out-of-distribution reconstructions on MS-COCO (Lin et al., 2014), despite training only on ImageNet. Figure 6 presents class-conditional samples; despite the simple architecture and low compute budget, quality remains competitive with attention-based image generation models. We provide more uncurated, qualitative samples in the Appendix.

Latent Space Interpolation A key advantage of self-supervised encoders is the smoothness of their latent spaces, yielding a geometry well-suited for generation. Figure 4 shows that our training preserves this property, where we linearly interpolate between latent representations, which produces gradual transitions in both high-level semantics and low-level visual details. We observe continuous changes in object shape, size, emergence, and rotation (see more samples in the Appendix).

4.2 Enabling RepTok for T2I generation

We scale RepTok to text-to-image generation using 120M image–text pairs from COYO (Byeon et al., 2022), recaptioned using InternVL3-1B (Zhu et al., 2025). We first train the language-agnostic

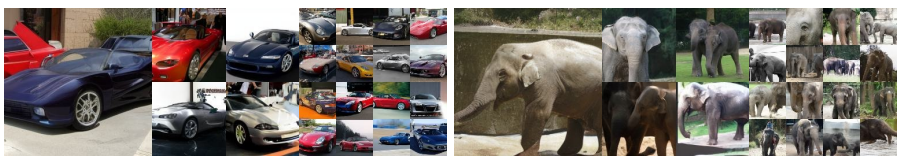

Figure 6: Uncurated MLP-Mixer ImageNet generations (CFG=3.5). More samples in the Appendix.

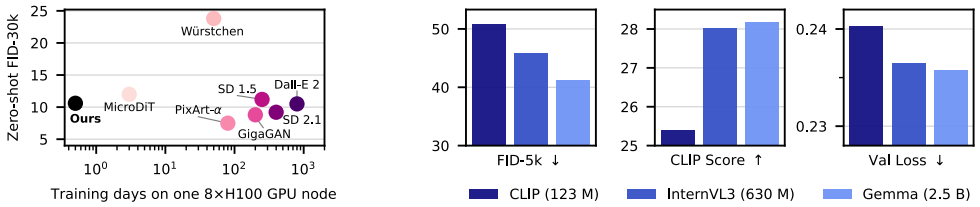

Figure 7: *Left:* Training days vs gFID, zero-shot evaluation on MS-COCO (Lin et al., 2014). Data sourced from MicroDiT (Sehwag et al., 2024). *Right:* Scaling the frozen language backbones results in improved performance. Language models: CLIP, InternVL, and Gemma-2B.

encoder-decoder using DINOv2 as our SSL encoder and a Flow Matching transformer as decoder. During generative model training, we concatenate four learnable tokens with the noisy `[cls]` token from the SSL encoder and apply cross-attention to the frozen outputs of the language model. Similar to prior work, we evaluate our method on the COCO validation set (Lin et al., 2014). We report FID, CLIP Score (Hessel et al., 2021), as well as validation loss, as (Esser et al., 2024; Polyak et al., 2024) found that it correlates with human evaluations.

Quantitative Results Figure 7 (*left*) shows that our method achieves substantially lower training cost than prior text-to-image models while maintaining competitive zero-shot FID. Since the language backbone is frozen and only provides conditioning, it can be scaled independently without impacting the training cost of the generative model. Figure 7 (*right*) shows the performance for language backbones with increasing scale: CLIP (Radford et al., 2021), InternVL (Zhu et al., 2025), and Gemma-2B (Team-Gemma et al., 2024). Larger language models consistently improve performance across all metrics. All results are obtained after 200k training iterations with a batch size of 256.

Qualitative Results Figure 8 shows qualitative text-to-image results. Our model is able to produce realistic images after only 200k training iterations. Despite the short training time (< 20 hours on $4\times$ A100 GPUs), the generations capture fine details and adhere closely to the prompt. This highlights the efficiency and scalability of RepTok for text-to-image synthesis. Interestingly, we observe that the SSL encoder and generative decoder trained exclusively on ImageNet can already be repurposed for text-to-image generation. We show qualitative samples and discuss this further in the Appendix.

Table 3: FID comparison on the ImageNet 256×256 benchmark, including parameter counts and training FLOPs. *Stage 1* refers to the training of the generative decoder, while *Stage 2* corresponds to the main generative model training. As all models rely on the SD-VAE and REPA uses DINOv2 as well, we exclude these shared pre-training costs from FLOP estimates.

		Stage 1	Stage 2				
Model	FID	PFlops	Train Steps	Params (M)	GFlops/Iter	PFlops	Total PFlops
DiT-XL/2	19.5	–	400K	675	118.6	12.1K	12.1K
+REPA	12.3	–	400K	675	140.5	14.4K	14.4K
SiT-L/2	18.8	–	400K	458	77.5	7.9K	7.9K
+REPA	9.7	–	400K	458	99.4	10.2K	10.2K
SiT-XL/2	17.2	–	400K	675	118.6	12.1K	12.1K
+REPA	7.9	–	400K	675	140.5	14.4K	14.4K
SiT-XL/2	8.3	–	7M	675	118.6	212.5K	212.5K
+CFG=1.5	2.06	–	7M	675	118.6	212.5K	212.5K
+REPA	5.9	–	4M	675	140.5	143.9K	143.9K
+REPA, CFG=1.5	1.42	–	4M	675	140.5	143.9K	143.9K
RepTok	5.4	30.4K	100K	276	23.0	0.6K	31.0K
RepTok	3.4	30.4K	700K	276	23.0	4.1K	34.5K
+CFG=1.5	3.22	30.4K	700K	276	23.0	4.1K	34.5K
RepTok	2.06	30.4K	460K	516	25.0	11.7K	42.1K
+CFG=1.5	1.88	30.4K	460K	516	25.0	11.7K	42.1K

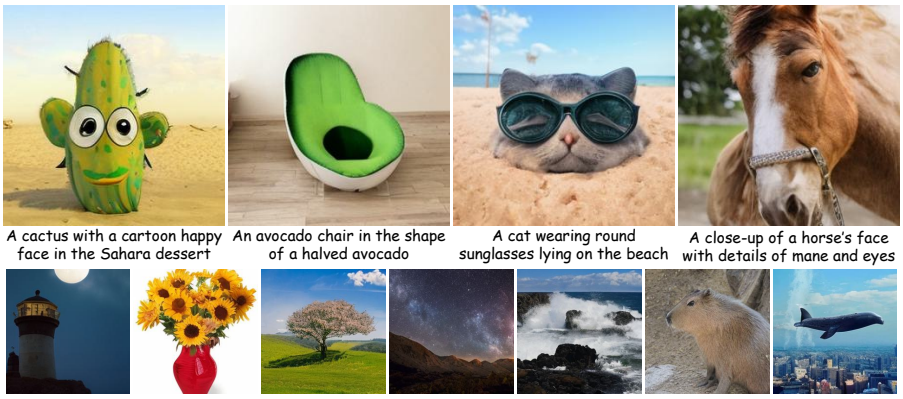

Figure 8: RepTok text-to-image results with a transformer-based latent space model (CFG scale 3.5).

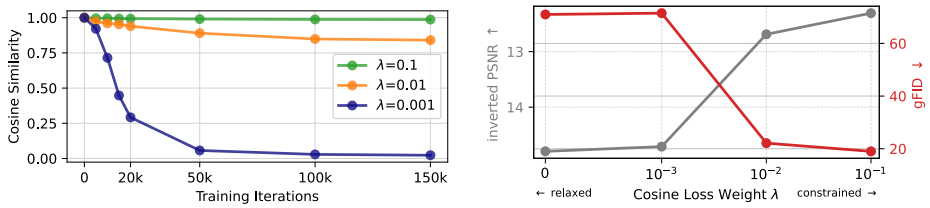

Figure 9: The parameter λ of the cosine similarity loss in Equation (3) allows us to trade off between pixel-wise reconstruction and generation capabilities. Relaxed constraints (low λ) improve pixel-wise reconstruction (PSNR in right plot), but result in poor generation capabilities (gFID in right plot).

4.3 Ablations

Generalization to other SSL methods Our method generalizes to a number of self-supervised vision encoders, as shown in Table 4. While the main results are based on DINOv2, we observe similarly strong reconstruction quality and generative performance when using alternative SSL methods such as MAE and CLIP. In contrast, when using a randomly initialized encoder with no prior information, the generative decoder loss enforces a strong pixel-wise reconstruction but leaves the resulting latent space completely unstructured and hard to capture for the generative model, as reflected in the high generation FID. A semantic prior enforces a geometry in which semantically similar images are drawn together and dissimilar images are pushed apart. This naturally induces smooth, low-dimensional manifolds which promotes stable generations.

Cosine Similarity Loss We introduced a cosine similarity loss in Equation (3) that incentivizes the semantic token to remain close to the SSL encoder's original to preserve the beneficial properties of the pre-trained space. Here, similar to previous work (Yao et al., 2025; Tschannen et al., 2024), we observe a trade-off between generation and reconstruction, visualized in Figure 9. Stronger regularization improves the generative performance (gFID), but at the cost of reduced pixel-wise reconstruction (PSNR). Mild regularization significantly improves the generative quality, indicating a better latent space for generation, while minimally degrading reconstruction quality. λ allows us to balance between preserving high-level semantic content and reconstructing low-level visual details.

Table 4: Our approach generalizes to other self-supervised encoders. We compare 10k FID on class-conditional ImageNet (Deng et al., 2009).

SSL method	rFID ↓	PSNR ↑	SSIM ↑	LPIPS ↓	gFID ↓
w/o prior	13.99	19.64	47.19	0.23	128.54
CLIP (Radford et al., 2021)	13.66	14.24	31.69	0.44	30.56
MAE (He et al., 2022)	9.09	13.79	30.28	0.45	28.48
DINOv2 (Oquab et al., 2024)	7.95	14.94	33.26	0.41	20.75

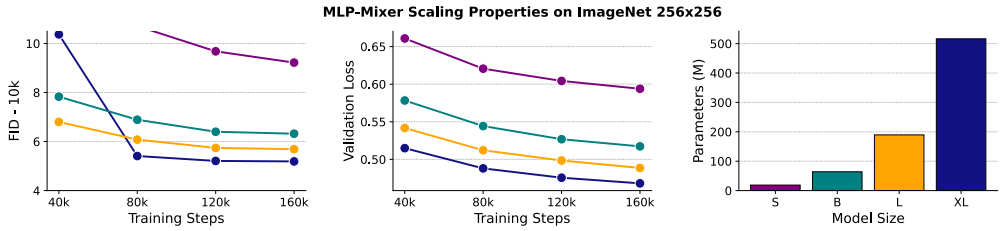

Figure 10: Scaling analysis of the MLP Mixer (Tolstikhin et al., 2021). We observe that the model scales: larger variants consistently yield improved performance, and continued training further improves results. The largest XL model evaluated contains 516M parameters.

Figure 11: Fine-tuned encoder-decoder model on 512^2 resolution. *Left*: original validation images. *Right*: reconstructions at 512^2 resolution.

Scaling Analysis of MLP-Mixer Since parameter scaling is a key feature of Transformers, we further evaluate whether this also holds for our latent space MLP-Mixer architecture (Tolstikhin et al., 2021). In Figure 10 we show that the MLP Mixer also scales with parameter size and FLOPs.

Scaling the Decoder to 512px We further fine-tune our encoder–decoder model for 100K iterations at 512^2 resolution, starting from weights pre-trained at 256^2. The model successfully adapts to the higher-resolution setting as visualized in Figure 11.

5 Conclusion

In this work, we introduced RepTok, a framework that adapts self-supervised representations into a compact latent space for generative modeling. By fine-tuning only the class token of an SSL encoder and regularizing it with a cosine-similarity loss, we obtain a single continuous token that retains the smooth geometry of the original space while enriching it with reconstruction-relevant information. Coupled with a generative decoder trained via flow matching, this setup enables faithful reconstructions and efficient image synthesis without reliance on costly attention mechanisms or auxiliary losses. Our experiments demonstrate that this single-token formulation achieves competitive results in class-conditional generation at a fraction of the computational cost. We further show that RepTok scales to more complex text-to-image settings. Overall, these findings highlight the potential of leveraging SSL representations themselves to build lightweight but effective generative models.

REFERENCES

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in neural information processing systems*, 34:17981–17993, 2021.

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. *arXiv preprint arXiv:2502.13967*, 2025.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023a.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22563–22575, 2023b.

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

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

Katherine Crowson, Stefan Andreas Baumann, Alex Birch, Tanishq Mathew Abraham, Daniel Z. Kaplan, and Enrico Shippole. Scalable high-resolution pixel-space image synthesis with hourglass diffusion transformers, 2024. URL https://arxiv.org/abs/2401.11605.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *The Twelfth International Conference on Learning Representations*, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Michael Fuest, Pingchuan Ma, Ming Gui, Johannes Schusterbauer, Vincent Tao Hu, and Bjorn Ommer. Diffusion models and representation learning: A survey. *arXiv preprint arXiv:2407.00783*, 2024.

Ishaan Gulrajani, Kundan Kumar, Faruk Ahmed, Adrien Ali Taiga, Francesco Visin, David Vazquez, and Aaron Courville. Pixelvae: A latent variable model for natural images. *arXiv preprint arXiv:1611.05013*, 2016.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022.

Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021.

Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Tianhong Li, Dina Katabi, and Kaiming He. Return of unconditional generation: A self-supervised representation generation method. *Advances in Neural Information Processing Systems*, 37: 125441–125468, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

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

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. Audioldm: text-to-audio generation with latent diffusion models. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 21450–21474, 2023a.

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

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Jiasen Lu, Liangchen Song, Mingze Xu, Byeongjoo Ahn, Yanjun Wang, Chen Chen, Afshin Dehghan, and Yinfei Yang. Atoken: A unified tokenizer for vision. *arXiv preprint arXiv:2509.14476*, 2025.

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 *European Conference on Computer Vision*, pp. 23–40. Springer, 2024.

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. *Transactions on Machine Learning Research Journal*, pp. 1–31, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Pablo Pernias, Dominic Rampas, Mats Leon Richter, Christopher Pal, and Marc Aubreville. Würstchen: An efficient architecture for large-scale text-to-image diffusion models. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*. OpenReview, 2024.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10619–10629, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. URL https://arxiv.org/abs/2204.06125.

Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.

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

Johannes Schusterbauer, Ming Gui, Pingchuan Ma, Nick Stracke, Stefan Andreas Baumann, Vincent Tao Hu, and Björn Ommer. Fmboost: Boosting latent diffusion with flow matching. In *European Conference on Computer Vision*, pp. 338–355. Springer, 2024.

Johannes Schusterbauer, Ming Gui, Frank Fundel, and Björn Ommer. Diff2flow: Training flow matching models via diffusion model alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28347–28357, 2025.

Vikash Sehwag, Xianghao Kong, Jingtao Li, Michael Spranger, and Lingjuan Lyu. Stretching each dollar: Diffusion training from scratch on a micro-budget. *arXiv preprint arXiv:2407.15811*, 2024.

Noam Shazeer. Glu variants improve transformer, 2020. URL https://arxiv.org/abs/2002.05202.

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. *arXiv preprint arXiv:2510.15301*, 2025.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2023. URL https://arxiv.org/abs/2104.09864.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation, 2024. URL https://arxiv.org/abs/2406.06525.

Gemma Team-Gemma, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024.

Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34:24261–24272, 2021.

Michael Tschannen, Cian Eastwood, and Fabian Mentzer. Givt: Generative infinite-vocabulary transformers. In *European Conference on Computer Vision*, pp. 292–309. Springer, 2024.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Ziqiao Wang, Wangbo Zhao, Yuhao Zhou, Zekai Li, Zhiyuan Liang, Mingjia Shi, Xuanlei Zhao, Pengfei Zhou, Kaipeng Zhang, Zhangyang Wang, et al. Repa works until it doesn't: Early-stopped, holistic alignment supercharges diffusion training. *arXiv preprint arXiv:2505.16792*, 2025.

Wilson Yan, Volodymyr Mnih, Aleksandra Faust, Matei Zaharia, Pieter Abbeel, and Hao Liu. Elastictok: Adaptive tokenization for image and video. *arXiv preprint arXiv:2410.08368*, 2024.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. *arXiv preprint arXiv:2501.01423*, 2025.

Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. An image is worth 32 tokens for reconstruction and generation. *Advances in Neural Information Processing Systems*, 37:128940–128966, 2024a.

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. *arXiv preprint arXiv:2410.06940*, 2024b.

Biao Zhang and Rico Sennrich. Root mean square layer normalization, 2019. URL https://arxiv.org/abs/1910.07467.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *Advances in Neural Information Processing Systems*, 36:45533–45547, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.