One Layer Is Enough: Adapting Pretrained Visual Encoders for Image Generation

Yuan Gao, **Chen Chen**, **Tianrong Chen**, **Jiatao Gu**

Apple

Visual generative models (e.g., diffusion models) typically operate in compressed latent spaces to balance training efficiency and sample quality. In parallel, there has been growing interest in leveraging high-quality pre-trained visual representations—either by aligning them inside VAEs or directly within the generative model. However, adapting such representations remains challenging due to fundamental mismatches between understanding-oriented features and generation-friendly latent spaces. Representation encoders benefit from high-dimensional latents that capture diverse hypotheses for masked regions, whereas generative models favor low-dimensional latents that must faithfully preserve injected noise. This discrepancy has led prior work to rely on complex objectives and architectures. In this work, we propose **FAE** (Feature Auto-Encoder), a simple-yet-effective framework that adapts pre-trained visual representations into low-dimensional latents suitable for generation using as little as a single attention layer, while retaining sufficient information for both reconstruction and understanding. The key is to couple two separate deep decoders: one trained to reconstruct the original feature space, and a second that takes the reconstructed features as input for image generation. FAE is generic—it can be instantiated with a variety of self-supervised encoders (e.g., DINO, SigLIP) and plugged into two distinct generative families–diffusion models and normalizing flows. Across class-conditional and text-to-image benchmarks, FAE achieves strong performance. For example, on ImageNet 256×256, our diffusion model with CFG attains a near–state-of-the-art FID of 1.29 (800 epochs) and 1.70 (80 epochs). Without CFG, FAE reaches the state-of-the-art FID of **1.48** (800 epochs) and 2.08 (80 epochs), demonstrating both high quality and fast learning.

Correspondence: Yuan Gao (ygao65@apple.com), Jiatao Gu (jgu32@apple.com)
Date: December 17, 2025

1 Introduction

In past years, diffusion models (Nichol and Dhariwal, 2021; Saharia et al., 2021; Rombach et al., 2022) have significantly advanced the quality and flexibility of visual generation, making them the dominant framework for producing high-resolution images and videos. A key recent change driving this progress is the integration of powerful pre-trained visual representations, typically obtained from large-scale self-supervised learning frameworks based on masked image prediction (Oquab et al., 2023; Tschannen et al., 2025). Such frameworks—exemplified by models like REPA (Yu et al., 2024b) and VA-VAE (Yao et al., 2025) utilize the rich semantic and structural information from the large models trained on unlabeled data. When incorporated into the diffusion pipeline—either within the denoising process or in variational autoencoders (VAEs)—these represen-
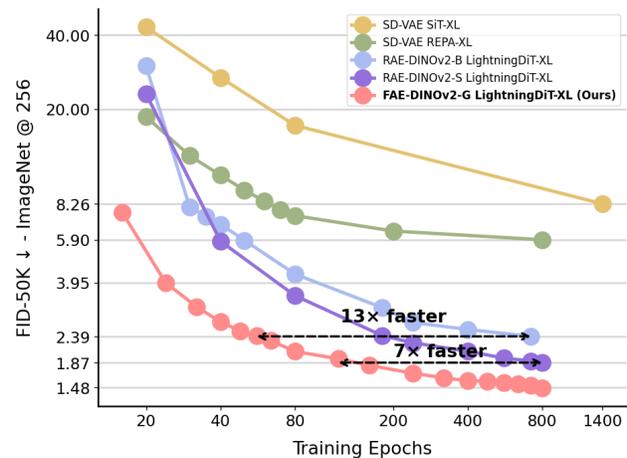

Figure 1 Training convergence on ImageNet 256×256. FAE achieves strong sample quality and converges 7–13× faster than concurrent baselines (Zheng et al., 2025).
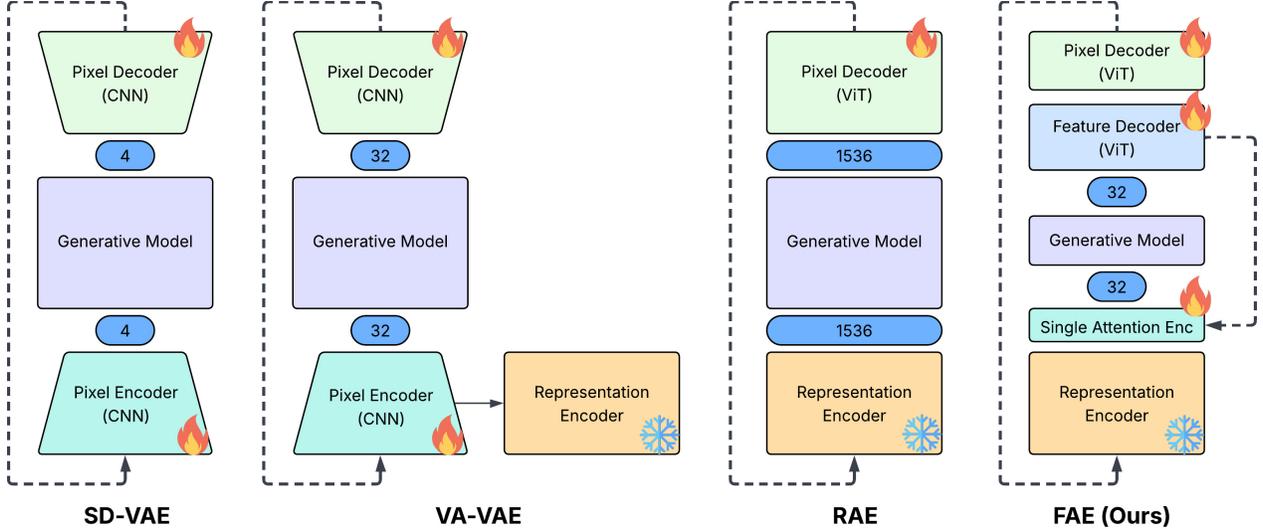
Figure 2 Comparison between standard VAE (Rombach et al., 2022), VA-VAE (Yao et al., 2025), RAE (Zheng et al., 2025) and our proposed FAE. The number shows the channel dimension of the generative modeling space.

tations substantially improve both training efficiency and generative fidelity.

Despite this progress, adapting pre-trained visual representations to generative models remains challenging due to the inherent mismatches between understanding-oriented representations and generation-friendly latent spaces. Self-supervised models, in order to build up a hard task with the unlabelled data, masking and prediction tasks are used, not only in the image area but also text and audio. To capture the complicated distribution of different possibilities of the masked regions, especially to simulate the distribution with simple embedding multiplication and softmax function, a large latent dimension is required. In contrast, generation models such as diffusion models and normalizing flow models, are often formulated as denoising processes, evolving a noisy input toward a clean signal through an iterative refinement process. In diffusion models, for instance, the input is perturbed by Gaussian noise and repeatedly denoised through multiple timesteps. To ensure trajectory stability throughout this process, the hidden representations must simultaneously encode the information of both the noised input and its clean predicted target. When the latent dimension is large, this dual encoding becomes more resource demanding, and the diffusion dynamics become sensitive to noise-level scheduling, often leading to instability or slower convergence. Therefore, generative models favor compact, low-dimensional latent spaces, which make the denoising trajectory smoother, reduce the training burden, and preserve generative fidelity under limited model capacity. This discrepancy often results in inefficiencies and necessitates complex architectural modifications when integrating pre-trained representations.

In this paper, we revisit this problem from a different perspective and ask: Is it truly necessary to preserve the high-dimensional structure of pre-trained visual representations when just zero-shot adapting them? In fact, although self-supervised models are trained on masked prediction tasks, the adaptation only involves unmasked inputs where the need for modeling diverse distribution is diminished. Instead, the goal is to leverage the rich semantics and spatial information from the pre-train features.

Building on this insight, we introduce **FAE** (**F**eature **A**uto-**E**ncoder), a simple yet effective framework that compresses the pre-train embedding into a compact, generation-friendly space. We employ only a single attention layer followed by a linear projection to map the embeddings into a continuous low-dimensional code, and use a lightweight decoder to reconstruct the original features. During experiments, we observed that the adaptation task is substantially weaker than the original self-supervised pre-training task; as a result, overly complex adapting frameworks tend to lose information from the pre-trained embeddings. Empirically maintaining a closer distance between the compressed code and the original embedding leads to higher reconstruction quality. Therefore, we adopt a minimal design, using only a single attention layer as the encoder to remove redundant global information shared across patch embeddings.

We validate our method by integrating it into existing diffusion frameworks, including SiT (Ma et al., 2024)

and LightningDiT (Yao et al., 2025) and normaling-flow based models (e.g., STARFlow (Gu et al., 2025a)). On ImageNet 256×256 generation, with CFG, diffusion model using FAE attains a near–state-of-the-art FID of 1.29 in 800 epochs and reaches an FID of 1.70 within only 80 training epochs. Without CFG, diffusion model using FAE achieves a state-of-the-art FID of **1.48** in 800 epochs and reaches an FID of 2.08 with only 80 epochs, highlighting both its generation quality and its learning efficiency.

2 Related Work

Visual Representation Learning Self-supervised learning (SSL) has become a cornerstone for learning general visual representations without manual labels. Early contrastive frameworks such as MoCo (He et al., 2020) and SimCLR (Chen et al., 2020) maximized agreement between augmented views of the same image, while later non-contrastive approaches such as BYOL (Grill et al., 2020) and DINO (Caron et al., 2021) demonstrated that strong representations can emerge even without negative pairs. These methods trained large Vision Transformers (ViTs) (Dosovitskiy et al., 2021) to learn globally coherent, semantically rich feature spaces that transfer effectively to downstream tasks.

More recent works have explored parameter-efficient ways to adapt such pretrained ViTs. Adapter-based and prompt-tuning methods—including AdaptFormer (Chen et al., 2022a) and Visual Prompt Tuning (VPT) (Jia et al., 2022)—insert lightweight modules or learnable tokens into frozen backbones to tailor them to new domains with minimal fine-tuning overhead. However, most SSL-based adaptation studies focus on discriminative objectives—classification, segmentation, or retrieval—rather than generative modeling. Our work differs in that it repurposes a pretrained self-supervised ViT for visual generation, showing that a single attention layer suffices to bridge the gap between discriminative pretraining and generative diffusion training.

Visual Generative Models Diffusion models (Nichol and Dhariwal, 2021; Song et al., 2021) have emerged as a dominant paradigm for image generation, achieving remarkable realism and diversity. LDM (Rombach et al., 2022) further improved efficiency by performing denoising in a compressed latent space learned by a VAE. Subsequent work has explored various architectural improvements to transformers, including DiT (Peebles and Xie, 2023) and SiT (Ma et al., 2024), which are designed for scalability and enhanced global context modeling. More recently, normalizing flow based models such as TARFlow (Zhai et al., 2024) and STARFlows (Gu et al., 2025a,b) have emerged new type of generative models that become promising alternatives drastically different from standard diffusion models on visual generation.

Representation Alignment Aligning generative model representations with pretrained visual encoders has proven effective for stabilizing training and enhancing sample quality. REPA (Yu et al., 2024b) proposes to align noisy hidden features of diffusion transformers with clean image embeddings from a pretrained ViT, substantially accelerating convergence and improving fidelity. Follow-up work such as REPA-E(Wu et al., 2025) extends this idea by enforcing semantic consistency between latent tokens and image features throughout training. Recent analyses have emphasized the inherent tension between reconstruction quality and generation stability, underscoring the need to reconcile these competing objectives. VA-VAE (Yao et al., 2025) highlights that high-dimensional latent spaces may favor reconstruction but hinder generative convergence, motivating strategies that align latent encoders with pretrained vision models. Along this line, concurrent to our work, several studies have explored directly using pre-trained embeddings as tokenizer inputs to improve generation quality, including VFM-VAE (Bi et al., 2025) and RepTok (Gui et al., 2025). Contemporary work RAE (Zheng et al., 2025) directly adopts pre-trained embeddings as the diffusion latent space. While this avoids explicit alignment, it demands significant architectural changes to the generator (e.g., wider channels, additional heads) to accommodate high-dimensional feature maps.

Our work complements these findings by reusing pretrained ViTs directly as the generative backbone rather than modifying the latent space, and demonstrates that lightweight adaptation via a single attention layer can yield stable and high-quality diffusion generation. Where prior methods typically rely on external alignment losses or auxiliary projection heads to bridge discriminative and generative representations, our method internalizes the alignment process: by inserting a single-layer attention adapter into a pretrained ViT, we enable the model's own attention mechanisms to reconcile discriminative and generative objectives. This design unifies feature reuse and representation alignment within a single, minimal architectural modification, reducing both training cost and overall system complexity.

3   Motivation

A core challenge in adapting pre-trained visual representations for generative models is the inherent dimensional and functional mismatch between self-supervised understanding models and generation models. Representation encoders (He et al., 2021; Oquab et al., 2023; Tschannen et al., 2025; Radford et al., 2021), whose performance typically improves with higher-dimensional feature spaces, naturally favor large embeddings. For example, Dino-V2-G (Oquab et al., 2023) has a dimension of $1,536$. This has also become the de-facto choice in MLLMs (Liu et al., 2023; Bai et al., 2023). However, such high-dimensional representations are poorly suited for many generative models (Rombach et al., 2022; Peebles and Xie, 2023; Gu et al., 2025a) operating in a latent space. Unlike representation learning tasks, which only require encoding semantic information, generation tasks must accurately recover fine-scale details from noisy inputs, making the denoising process highly sensitive to the dimensionality and structure of the latent space, making it harder for the model to preserve and refine the injected noise, leading to instability and reduced sample quality. In normal cases, generation works in a much lower dimensional space, ranging from $4 \sim 64$.

Existing methods tackle this mismatch from two main directions: **(1) Feature alignment.** Methods such as REPA (Yu et al., 2024b) and VA-VAE (Yao et al., 2025) attempt to align the features of an external representation encoder with the generative model, either inside the generator or within a VAE. This typically requires carefully designed alignment losses and additional training stages. However, because the generator or VAE architecture is substantially different from the pre-trained encoder, such alignment inevitably discards information that is not immediately useful for generation, limiting the benefit of using the original representations. **(2) Direct modeling.** More recently, RAE (Zheng et al., 2025) directly adopts pre-trained embeddings as the diffusion latent space. While this avoids explicit alignment, it demands significant architectural changes to the generator (e.g., wider channels, additional heads) to accommodate high-dimensional feature maps. As a result, the model design becomes tightly coupled to the embedding dimensionality, making it difficult to scale or transfer across different encoders.

This motivates us to seek a design that simultaneously (1) keeps generation in a low-dimensional latent space, so that existing generative architectures can be reused without substantial ad-hoc modifications; and (2) remains as close as possible to the representation-encoder feature space, so that we fully inherit the strengths of pre-trained features and can ideally recover their semantics from the learned latents, rather than relying on lossy alignment losses.

To this end, we introduce **FAE**, a new Feature Auto-Encoding approach in which a lightweight encoder compresses high-dimensional representation features into a compact latent space tailored for generation. In the following, we formalize this design and describe the architecture and training objectives of FAE in detail.

4   Method

In this section, we introduce FAE, a simple-yet-effective frame to bridge visual representation learning and generative models using feature-level autoencoders.

4.1   Single-Attention Encoder

We first train a feature encoder to compress frozen pre-trained embeddings into a low-dimensional latent space. To preserve as much information as possible from the large pre-trained model, we deliberately use a minimal encoder: *a single self-attention layer followed by a linear projection* that maps pre-trained patch embeddings $\mathbf{x}$ to compact latents $\mathbf{z}$. This design reduces the parameter count and keeps the mapping close to the original feature space (see Figure 7 in Appendix).

Importantly, the adaptation objective (feature reconstruction) is substantially weaker than the original self-supervised pre-training task (e.g., masked region prediction). Over-parameterized encoders tend to overfit this easier objective, effectively re-encoding features for reconstruction and discarding information that is not directly supervised. Empirically, we find that keeping the encoder shallow and the latents close to the pre-trained embeddings leads to higher reconstruction quality and better downstream understanding. Notably, the self-attention layer is crucial: it operates across patch embeddings to remove redundant global information

and redistribute capacity, whereas a purely linear projection acts independently on each feature dimension and cannot adaptively de-redundantize patch-wise information. The original DINOv2 paper (Oquab et al., 2023) investigates several feature centering strategies, including moving-average centering after softmax and the Sinkhorn–Knopp algorithm. More recent work such as UniTok (Ma et al., 2025) also mentions the importance of attention mechanisms when compressing high-dimensional feature representations. Despite different formulations, these methods and our single-attention encoder may share a common underlying principle of removing redundant global information .

Our ablation in Table 5 confirms these observations: the single-attention encoder outperforms a purely linear encoder and yields significantly better reconstruction quality than a deep Transformer encoder.

4.2 Double Decoder

A central design in FAE is to separate *feature reconstruction* from *image synthesis*. Given compressed latents $\mathbf{z}$, we first reconstruct the original representation space, and only then decode pixels from the reconstructed features. This "double-decoder" design lets us preserve the semantics of the frozen encoder while giving the pixel decoder the flexibility needed for high-fidelity image generation.

Feature Decoder. Starting from the compact latent $\mathbf{z}$, a 6-layer Transformer feature decoder reconstructs the original embedding $\hat{\mathbf{x}}$. Each layer uses the same hidden dimension as the corresponding pre-trained backbone (e.g., DINOv2) so that the reconstructed features live in a compatible representation space. We employ Rotary Positional Embedding (RoPE) (Su et al., 2024), RMSNorm (Zhang and Sennrich, 2019), and SwiGLU activations (Shazeer, 2020), which we find empirically improve stability and reconstruction quality.

The feature decoder is trained with a standard VAE objective consisting of an ℓ_2 reconstruction term and a KL regularization term:

$$\mathcal{L}_{\text{VAE}} = \|\hat{\mathbf{x}} - \mathbf{x}\|_2^2 + \beta \, \text{KL}\big(q(\mathbf{z} \mid \mathbf{x}) \, \| \, p(\mathbf{z})\big). \tag{4.1}$$

This encourages $\mathbf{z}$ to remain close to a simple prior while retaining enough information for accurate recovery of the pre-trained embeddings. Compare to prior work like VA-VAE (Yao et al., 2025) and Wang and He (2025), our loss is quite simple, it's just L2 Loss. This make our reconstructed results can be easily zero-shot adapted into downstream task trained on original pre-train embedding, which we will show the results in the following subsection. In practice, we observe that high-quality feature reconstruction is a strong predictor of downstream generative and understanding performance.

Pixel Decoder. On top of the reconstructed features $\hat{\mathbf{x}}$, we attach a ViT-L–based pixel decoder (Dosovitskiy et al., 2021) that maps them to RGB images. Conceptually, the feature decoder restores the "language" of the pre-trained encoder, and the pixel decoder learns to translate that language into pixels. By letting the pixel decoder operate on rich, semantically meaningful embeddings rather than raw latents, we simplify the generation problem and improve visual fidelity. Following prior work (Yu et al., 2024a), we train the pixel decoder with a combination of adversarial, perceptual, and reconstruction losses:

$$\mathcal{L}_{\text{pix}} = \lambda_{\text{GAN}} \mathcal{L}_{\text{GAN}} + \lambda_{\text{perc}} \mathcal{L}_{\text{perc}} + \lambda_{\text{rec}} \mathcal{L}_{\text{rec}}. \tag{4.2}$$

Here, the adversarial term $\mathcal{L}_{\text{GAN}}$ encourages realistic textures and global coherence, the perceptual loss $\mathcal{L}_{\text{perc}}$ aligns high-level features with the ground-truth images, and the reconstruction term $\mathcal{L}_{\text{rec}}$ preserves low-level details. Together with the feature decoder, this yields a compact latent space that remains semantically faithful to the pre-trained encoder while supporting high-quality image synthesis.

We train the pixel decoder in two stages, entirely in the embedding space. In the first stage, we inject Gaussian noise into the frozen pre-trained embeddings and train the decoder to directly reconstruct images from these noisy features: $\tilde{\mathbf{x}} = \mathbf{x} + \epsilon, \quad \epsilon \sim \mathcal{N}\big(0, \sigma^2 I\big)$, where we set $\sigma = 0.4$ for DINOv2 and scale it according to the norm of the pre-trained embeddings. This produces a *Gaussian embedding decoder* that is robust to moderate perturbations in the representation space.

Pixel Fine-Tuning. Once the first stage converges, we fine-tune the same pixel decoder on the reconstructed embeddings $\hat{\mathbf{x}}$ produced by the feature decoder. Because the decoder only operates on embedding space, the Gaussian embedding decoder can be reused across different variants of FAE without architectural changes.
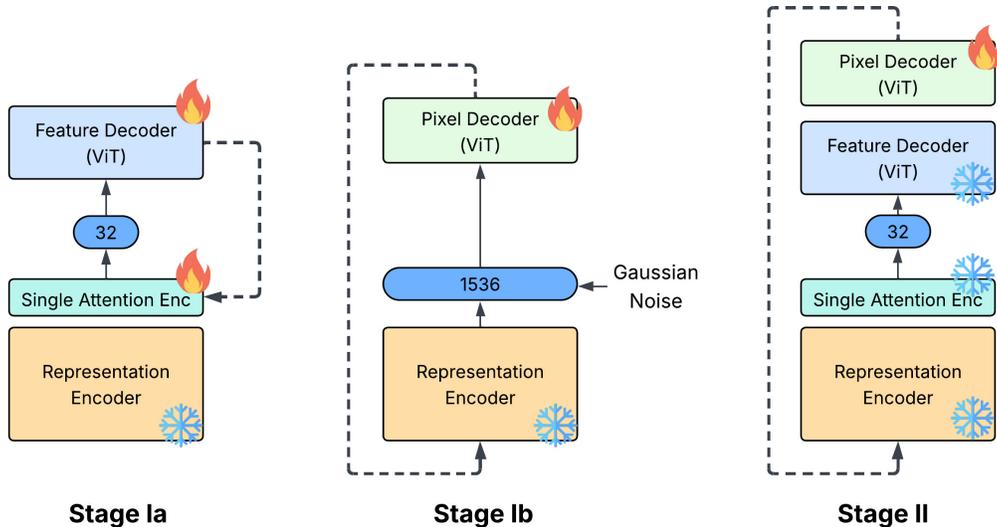
Figure 3 An illustration of Training Stages of FAE. Stage Ia and Ib can be trained independently.

Remarkably, even without fine-tuning on $\hat{\mathbf{x}}$, the Gaussian embedding decoder already achieves strong generation quality , indicating that our compressed latent space preserves the majority of the information in the original pre-trained embeddings. We visualize the overall training stages in Figure 3. The detailed parameters and pixel decoder reconstruction quality are available in the Appendix.

4.3 Generative Model Training

Once the latent space is ready, we in parallel train generative models *directly* on the *low-dimensional, compact* latents $\mathbf{z}$. During this stage, only the frozen backbone encoder and the single-layer feature encoder are used to map images into latent space, making generator training both memory- and compute-efficient. This decoupling turns the latent space into a modular interface: as long as a model can predict or transform $\mathbf{z}$, it can be used as a generator without any change to its core architecture. In this work, we instantiate two representative models on top of the same FAE latent space: SiT (Ma et al., 2024), a diffusion model, and STARFlow (Gu et al., 2025a), a normalizing flow. For both, we adopt the default parameterizations and training configurations from their original works, simply replacing their native latent representation with our learned $\mathbf{z}$, without additional architectural tricks or alignment losses.

4.4 Semantic Preservation in the Latent Space

A key property of FAE is that it can be directly applied to a variety of pre-trained visual representations (e.g., DINOv2 (Oquab et al., 2023), SigLIP (Tschannen et al., 2025)) without architectural changes, while largely preserving their understanding capabilities, thanks to an explicit feature-decoder reconstruction objective that encourages the latent space to stay semantically close to the original embeddings rather than collapsing into a purely generative code.

We verify this by examining patch-wise similarity structure (see Figure 8 and Figure 9 in Appendix). After passing through FAE, patches that are close in the original representation space remain close in our latent space, indicating that FAE largely preserves the relational geometry of the pre-trained features. Moreover, our latents retain the cross-image patch–matching behavior characteristic of DINOv2: semantically corresponding regions across different images (e.g., a player's hand, an animal's head) are still reliably matched using cosine similarity in the FAE's latent space (see Figure 4 and also Figure 10 Figure 11 in Appendix). This suggests that FAE preserves fine-grained, part-level semantics rather than only coarse global information.
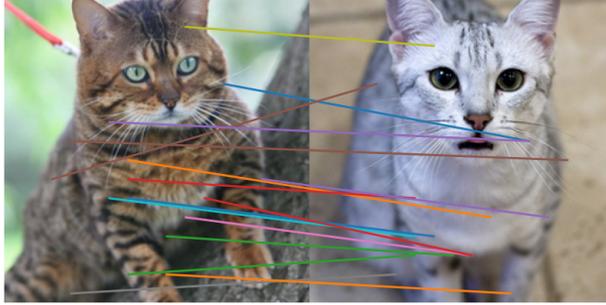
Figure 4 Matching across images. We match patch-level FAE features between images from different images that share similar semantic information. This exhibits the ability of our model to understand relations between similar parts of different objects.

Model	Res	ImageNet top-1
DinoV2-S/14 distilled	224	80.80%
DinoV2-B/14 distilled	224	84.40%
DinoV2-L/14 distilled	224	86.50%
DinoV2-g/14	224	87.00%
FAE (DinoV2-g/14)	224	86.17%

Table 1 ImageNet Linear Probing top-1 accuracy comparison for FAE and different DinoV2 variants (all at 224 resolution).

5 Experiments

We evaluate our proposed method on two standard generation benchmarks: class-conditional image generation on ImageNet-1K (Deng et al., 2009) and text-to-image generation trained on CC12M (Changpinyo et al., 2021) and evaluated on MS-COCO (Lin et al., 2015). Our experiments show that the proposed approach substantially accelerates training convergence while improving overall generation quality. To further demonstrate the generality of our framework, we also apply it to the STARFlow (Gu et al., 2025a) training paradigm and observe consistent improvements. In addition, we investigate whether the learned latent representations preserve strong semantic understanding capabilities by performing zero-shot adaptation on common downstream tasks, including ImageNet-1K linear probing and MS-COCO image–text retrieval.

5.1 Class-conditional Image Generation

Implementation details. We processed images from ImageNet into resolution of 256×256 following ADM (Dhariwal and Nichol, 2021). Then each image is then encoded into a compressed vector of shape $16\times16\times32$ using FAE. For pre-train embedding, we explored DinoV2 (Oquab et al., 2023), we use a batch size of 1024 for training VAE. For latent diffusion model, we explored SiT following Ma et al. (2024)'s setting and LightningDit following Yao et al. (2025)'s setting, we use the XL model size from these papers. To ensure a fair comparison with DiTs and SiTs, we consistently use a batch size of 512 during training.

Evaluation and results. For generation without CFG (Ho and Salimans, 2022), we use SDE (Song et al., 2021) and runs 250 steps. We achieve an FID score of 2.08 in 80 epochs and an FID score of 1.48 in 800 epochs. For generation with CFG, we use ODE and runs 250 steps. We achieve an FID score of 1.70 in 80 epochs, and an FID score of 1.29 in 800 epochs. Detailed CFG and timesteps shifts are available in the appendix. Experiments on ImageNet demonstrate that our method can achieves state-of-the-art (SOTA) performance on image generation without CFG. And the CFG results also got improved comparing to VA-VAE. We also provide some examples for the class conditioned generation in Figure 5. Detailed hyper-parameter settings are provided in Appendix Section D.

Method	Training Epochs	#Params	Generation w/o CFG					Generation w/ CFG				
			gFID	sFID	IS	Pre.	Rec.	gFID	sFID	IS	Pre.	Rec.
AutoRegressive (AR)												
MaskGIT (Chang et al., 2022)	555	227M	6.18	–	182.1	0.80	0.51	–	–	–	–	–
LlamaGen (Sun et al., 2024)	300	3.1B	9.38	8.24	112.9	0.69	0.67	2.18	5.97	263.3	0.81	0.58
VAR (Tian et al., 2024)	350	2.0B	–	–	–	–	–	1.80	–	365.4	0.83	0.57
MagViT-v2 (Yu et al., 2023)	1080	307M	3.65	–	200.5	–	–	1.78	–	319.4	–	–
MAR (Li et al., 2024)	800	945M	2.35	–	227.8	0.79	0.62	1.55	–	303.7	0.81	0.62
Latent Diffusion Models												
MaskDiT (Zheng et al., 2023)	1600	675M	5.69	10.34	177.9	0.74	0.60	2.28	5.67	276.6	0.80	0.61
DiT (Peebles and Xie, 2023)	1400	675M	9.62	6.85	121.5	0.67	0.67	2.27	4.60	278.2	**0.83**	0.57
SiT (Ma et al., 2024)	1400	675M	8.61	6.32	131.7	0.68	0.67	2.06	4.50	270.3	0.82	0.59
FasterDiT (Yao et al., 2024)	400	675M	7.91	5.45	131.3	0.67	**0.69**	2.03	4.63	264.0	0.81	0.60
MDT (Gao et al., 2023a)	1300	675M	6.23	5.23	143.0	0.71	0.65	1.79	4.57	283.0	0.81	0.61
MDTv2 (Gao et al., 2023b)	1080	675M	–	–	–	–	–	1.58	4.52	**314.7**	0.79	0.65
REPA (Yu et al., 2024b)	800	675M	5.90	–	–	–	–	1.42	4.70	305.7	0.80	0.65
VA-VAE (Yao et al., 2025)	64	675M	5.14	4.22	130.2	0.76	0.62	2.11	4.16	252.3	0.81	0.58
	800	675M	2.17	4.36	205.6	0.77	0.65	1.35	**4.15**	295.3	0.79	0.65
RAE (DiT-XL) (Zheng et al., 2025)	800	676M	1.87	–	209.7	0.80	0.63	1.41	–	309.4	0.80	0.63
RAE (DiTDH-XL)	80	839M	2.16	–	214.8	0.82	0.59	–	–	–	–	–
	800	839M	1.51	–	**242.9**	0.79	0.63	**1.13**	–	262.6	0.78	**0.67**
FAE	64	675M	2.55	4.37	189.9	0.82	0.58	2.01	4.39	250.3	0.83	0.59
	80	675M	2.39	4.38	192.8	0.82	0.59	1.92	4.35	249.6	0.83	0.59
	800	675M	1.58	4.38	223.7	0.80	0.63	1.41	4.34	274.1	0.81	0.63
FAE w/ Timestep Shift	64	675M	2.34	4.38	206.6	**0.83**	0.58	1.87	4.39	241.1	0.82	0.59
	80	675M	2.08	**4.20**	207.6	0.82	0.59	1.70	4.33	243.8	0.82	0.61
	800	675M	**1.48**	4.24	239.8	0.81	0.63	1.29	4.32	268.0	0.80	0.64

Table 2 **Class-conditional Image Generation Performance on ImageNet 256×256.**

5.2 Text-to-Image Generation

Implementation details. For pre-train embedding, we reused the DinoV2 Encoder from Imagenet and train an extra Siglip2 (Tschannen et al., 2025) Encoder on ImageNet. For training latent diffusion, we only use CC12M dataset with 256x256 resolution and follow the data processing from starflow (Gu et al., 2025a)'s setting. We consistently use a batch size of 256 during training. For text tokenzier, we use t5-xl. We compare our model based on DinoV2 and Siglip2 embeddings. We also add SD-VAE as baseline.

Evaluation and results. We evaluate FAE on MS-COCO (Lin et al., 2015) following the data preprocessing protocol of U-ViT (Bao et al., 2023). For sampling *without* using CFG (Ho and Salimans, 2022), we use an SDE sampler (Song et al., 2021) with 250 steps and obtain an FID of 7.47 after 400 training epochs. When using CFG, we switch to an ODE sampler with 250 steps and further improve the FID to 6.90 at 400 epochs. Notably, these near–state-of-the-art results are achieved using only CC12M for pre-training, i.e., with significantly fewer images than typical web-scale text-to-image models.

We also provide qualitative text-to-image examples in Figure 5, and more examples in Appendix Figure 13. These samples are generated at 384 × 384 resolution using a SigLIP2-FAE backbone and an MMDiT decoder with 2B parameters. The model produces visually coherent images that follow short text prompts, indicating good text–image alignment.

Overall, experiments on COCO show that our method attains competitive, near-SOTA image generation quality while using substantially less training data. Detailed hyper-parameter settings are provided in the Appendix Section D.

5.3 Latent Normalizing Flows

We further validate the universality of FAE by training a different family of latent generative model. Specifically, we instantiate STARFlow (Gu et al., 2025a) – an end-to-end latent normalizing flow generator that maps noise directly to FAE's latents, using the default configuration of 1.4B parameters and a standard SD-VAE baseline. The original STARFlow uses patch size 1 for SD-VAE latents, yielding sequences that are 4× longer than those of FAE; for a fair comparison, we instead use patch size 2 for SD-VAE, resulting in a sequence length of 256 equivalent to FAE. As shown in Figure 6, the SD-VAE baseline attains a FID of

Figure 5 Random samples of ImageNet 256x256 and Text-to-Images using diffusion models.

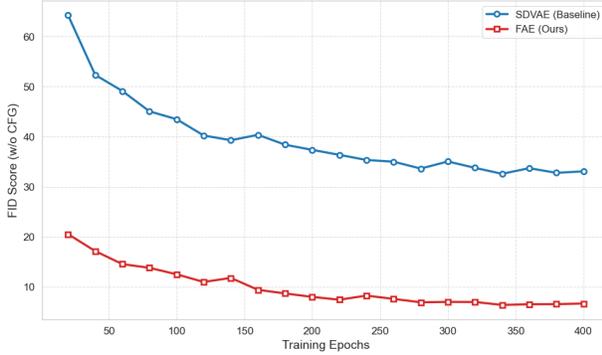| Model	FID	Type	Training datasets	#Params
DALL-E (Ramesh et al., 2021)	∼ 28	Autoregressive	DALL-E dataset (250M)	12B
CogView (Ding et al., 2021)	27.1	Autoregressive	Internal dataset (30M)	4B
LAFITE (Zhou et al., 2021)	26.94	GAN	CC3M (3M)	75M + 151M (TE)
GLIDE (Nichol et al., 2021)	12.24	Diffusion	DALL-E dataset (250M)	3.5B + 1.5B (SR)
Make-A-Scene (Gafni et al., 2022)	11.84	Autoregressive	Union datasets (without MS-COCO) (35M)	4B
DALL-E 2 (Ramesh et al., 2022)	10.39	Diffusion	DALL-E dataset (250M)	4.5B + 700M (SR)
Imagen (Saharia et al., 2022)	7.27	Diffusion	Internal dataset (460M) + LAION (400M)	2B + 4.6B (TE) + 600M (SR)
Parti (Yu et al., 2022)	7.23	Autoregressive	LAION (400M) + FIT (400M) + JFT (4B)	20B + 630M (AE)
Re-Imagen (Chen et al., 2022b)	**6.88**	Diffusion	KNN-ImageText (50M)	2.5B + 750M (SR)
SDVAE+T5 (w/o CFG)	21.25	Diffusion	CC12M	604M
FAE (SigLIPV2)+T5 (w/o CFG)	7.57	Diffusion	CC12M	604M+ 514M (FAE)
FAE (SigLIPV2)+T5 (w/ CFG)	7.11	Diffusion	CC12M	604M+ 514M (FAE)
FAE (DinoV2)+T5 (w/o CFG)	7.47	Diffusion	CC12M	604M+ 514M (FAE)
FAE (DinoV2)+T5 (w/ CFG)	6.90	Diffusion	CC12M	604M+ 514M (FAE)

Table 3 FID results of different models on MS-COCO validation (256 × 256). All the models are trained on external dataset and zero-shot evaluated on MS-COCO using 30K example.

4.51 under 400 training epoch, whereas the FAE-based variant (DinoV2-g/14) achieves a FID of **2.67** and converges substantially faster for both guided and unguided scenarios. Additional visual results are provided in the Appendix.
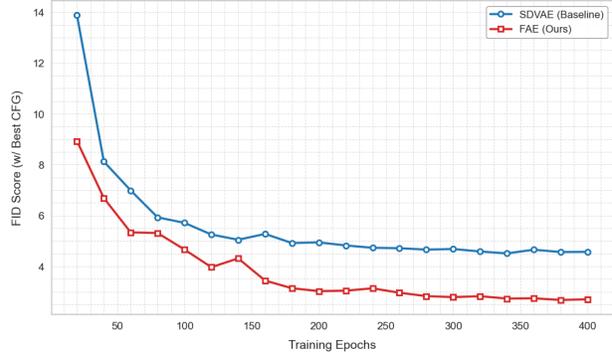
5.4 Image Understanding

We further validate the reconstructed embedding quality by zero-shot adapting it to the downstream tasks. **Linear Probing.** We evaluate the DINOv2 (Oquab et al., 2023) linear probing using existing layer and weights from the origin model. We directly passed the FAE reconstructed embedding to the layer. For data preprossing, we follow DINOv2. Espeically, we use the register version of FAE (The generation results are similar to the one we listed in the main experiments, which is based on the non-register one.) Results are shown in Table 1

Text-Image Retrieval. We evaluate the MS-COCO text-image dataset on COCO2014, following Siglip2's original setting. We compute the text embedding using SigLip2 (Tschannen et al., 2025) text encoder. And compute the image embedding using image encoder and pass it to FAE and use the reconstructed embedding. Then we compute the cosine similarity between the pair and find the Top-1. Results are shown in Table 4.

(a) Results without CFG.　　　　　　　　　　　　**(b)** Results with CFG.

Figure 6 Comparison of STARFlow (Gu et al., 2025a) with SDVAE and the proposed FAE under the same settings.

Model	ViT	Res	COCO T→I	COCO I→T
SigLIP2	g-opt/16	256	55.45%	73.10%
FAE (SigLIP2)	g-opt/16	256	55.79%	72.94%

Table 4 Text & Image Retrieval on COCO dataset.

6 Ablation Study

In this section, we conduct several ablation experiments to examine the impact of each component in our single-layer adaptation framework. For CFG-guidance, we grid search the CFG in 0.1 level and report the best results. All the CFG are enabled from t=0.7 to t=0.0 in the ablation experiments. For pixel decoder, we use a decoder fine-tuned with a 6 layer transformer reconstructed DinoV2 embedding. Detailed experiment settings and results for each ablation study are available in Section D in Appendix.

FAE Model Structure. We evaluate the different FAE structure and architectural design in this subsection. We show that the shallower networks yield better generation quality and faster convergence. Our single attention layer model outperforms both linear and 4/6 layer transformers in FID, while achieving comparable embedding-reconstruction similarity to deeper transformers—and substantially higher similarity than the linear baseline. This suggests that a compact attention design not only accelerates optimization but also preserves semantic representation quality, potentially benefiting downstream zero-shot probing. Results are available in Table 5

LDM Model Structure. We further analyze the model structure changes in the latent diffusion model. Starting from the base SiT architecture, we sequentially integrate SwiGLU, ROPE, and RMSNorm, resulting in a structure equivalent to LightningDiT. Our conclusion is consistent with the Lightning paper, each component contributes positively to both convergence speed and overall generation quality, with the largest gains observed when all three are combined. See Table 6

Token Dimension. We test VAE latent dimensions of 32, 48, and 64. After fine-tuning from the same decoder trained on Gaussian noise, the 64-dim model shows lower rFID than 48-dim and 32-dim variants However, the final generation results indicate that the 32-dim setting achieves the best FID scores and IS scores, as well as the fastest convergence. Notably, when time-shift is enabled (see below), the performance gap between different token dimensions narrows substantially. Results are available in Table 7

Time Shift. Finally, we ablate the time-shift parameter in the diffusion process. By introducing time-shift, loss weighting and diffusion trajectory changes, thus effectively bridge the quality differences across VAE latent dimensions and significantly accelerate convergence. With time-shift, our model reaches state-of-the-art convergence speed and generation quality within only 64 epochs. Results are available in appendix Table 8

Model	Linear Probing	CFG	64 Epochs	160 Epochs	320 Epochs
Single Attention	86.17%	w/o	2.98	2.27	1.98
		w/	–	1.79	1.61
Linear	85.74%	w/o	3.03	2.38	2.07
		w/	–	1.92	1.76
6-Layer Transformer	–	w/o	3.31	2.47	2.13
	–	w/	–	1.84	1.65
Direct Predict	–	w/o	15.37	12.99	12.72
DinoV2	–	w/	–	17.85	16.53

Table 5 Ablation results comparing different encoder structure.

Model	CFG	64 Epochs	160 Epochs	320 Epochs
SiT	w/o	2.98	2.27	1.98
	w/	–	1.79	1.61
SiT + SwiGLU	w/o	3.02	2.26	1.97
	w/	–	1.75	1.60
SiT + SwiGLU + ROPE	w/o	2.86	2.182	1.89
	w/	–	1.78	1.63
SiT + SwiGLU + ROPE + RMSNorm	w/o	2.74	2.15	1.86
	w/	–	1.71	1.55

Table 6 Ablation results comparing LDM model structure

7 Conclusion

We presented FAE, a simple yet powerful framework for adapting high-quality self-supervised visual representations for generative modeling. In contrast to prior approaches that rely on complex objectives or substantial architectural modifications to diffusion models, FAE uses an extremely simple design: a single attention layer paired with two lightweight decoders. Across class-conditional and text-to-image benchmarks, FAE demonstrates strong and efficient performance. On ImageNet 256×256, our diffusion model with CFG achieves a near–state-of-the-art FID of 1.29 with 800 training epochs and 1.70 with only 80 epochs. Without CFG, FAE attains a state-of-the-art FID of 1.48 (800 epochs) and 2.08 (80 epochs), highlighting both its high sample quality and fast learning behavior.

Despite these promising results, FAE still has limitations. Because the encoder is trained without an explicit image reconstruction loss, the rFID and tokenizer fidelity lag behind methods such as VA-VAE that directly optimize reconstruction quality. Overall, FAE provides a simple and general mechanism for leveraging pretrained vision encoders in generative modeling, offering a compelling balance between architectural minimalism, adaptability, and performance.

Model	CFG	64 Epochs	160 Epochs	320 Epochs
32-dim	w/o	2.67	2.02	1.76
	w/	–	1.70	1.52
48-dim	w/o	2.73	2.10	1.88
	w/	–	1.73	1.56
64-dim	w/o	2.86	2.25	1.99
	w/	–	1.76	1.64

Table 7 Ablation results comparing different latent dimension.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

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

Tianci Bi, Xiaoyi Zhang, Yan Lu, and Nanning Zheng. Vision foundation models can be good tokenizers for latent diffusion models, 2025.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640, 2021.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11315–11325, 2022.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts, 2021.

Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: adapting vision transformers for scalable visual recognition. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 2022a.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. Re-imagen: Retrieval-augmented text-to-image generator. *ArXiv preprint*, 2022b.

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*, pages 248–255. Ieee, 2009.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021.

Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. In *Advances in Neural Information Processing Systems*, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. *ArXiv preprint*, 2022.

Shanghua Gao, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. Masked diffusion transformer is a strong image synthesizer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23164–23173, 2023a.

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

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent a new approach to self-supervised learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

Jiatao Gu, Tianrong Chen, David Berthelot, Huangjie Zheng, Yuyang Wang, Ruixiang Zhang, Laurent Dinh, Miguel Angel Bautista, Josh Susskind, and Shuangfei Zhai. Starflow: Scaling latent normalizing flows for high-resolution image synthesis. *arXiv preprint arXiv:2506.06276*, 2025a.

Jiatao Gu, Ying Shen, Tianrong Chen, Laurent Dinh, Yuyang Wang, Miguel Angel Bautista, David Berthelot, Josh Susskind, and Shuangfei Zhai. Starflow-v: End-to-end video generative modeling with normalizing flow. *ArXiv preprint arXiv:2511.20462*, 2025b.

Ming Gui, Johannes Schusterbauer, Timy Phan, Felix Krause, Josh Susskind, Miguel Angel Bautista, and Björn Ommer. Adapting self-supervised representations as a latent space for efficient generation, 2025.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv:2111.06377*, 2021.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*, page 709–727, 2022.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *arXiv preprint arXiv:2406.11838*, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.

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

Chuofan Ma, Yi Jiang, Junfeng Wu, Jihan Yang, Xin Yu, Zehuan Yuan, Bingyue Peng, and Xiaojuan Qi. Unitok: A unified tokenizer for visual generation and understanding. *ArXiv*, abs/2502.20321, 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. *arXiv preprint arXiv:2401.08740*, 2024.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *ArXiv preprint*, 2021.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

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*, pages 8748–8763. PMLR, 2021.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *ArXiv preprint*, 2022.

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*, pages 10684–10695, 2022.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *ArXiv preprint*, 2021.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, 2022.

Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

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

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

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

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *arXiv preprint arXiv:2404.02905*, 2024.

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

Runqian Wang and Kaiming He. Diffuse and disperse: Image generation with representation regularization, 2025.

Ge Wu, Shen Zhang, Ruijing Shi, Shanghua Gao, Zhenyuan Chen, Lei Wang, Zhaowei Chen, Hongcheng Gao, Yao Tang, Jian Yang, Mingqiang Cheng, and Xiang Li. Representation entanglement for generation:training diffusion transformers is much easier than you think. *ArXiv*, abs/2507.01467, 2025.

Jingfeng Yao, Wang Cheng, Wenyu Liu, and Xinggang Wang. Fasterdit: Towards faster diffusion transformers training without architecture modification. *arXiv preprint arXiv:2410.10356*, 2024.

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

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *ArXiv preprint*, 2022.

Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, et al. Language model beats diffusion–tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023.

Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. An image is worth 32 tokens for reconstruction and generation, 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.

Shuangfei Zhai, Ruixiang Zhang, Preetum Nakkiran, David Berthelot, Jiatao Gu, Huangjie Zheng, Tianrong Chen, Miguel Angel Bautista, Navdeep Jaitly, and Josh Susskind. Normalizing flows are capable generative models. *arXiv preprint arXiv:2412.06329*, 2024.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 32, 2019.

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

Hongkai Zheng, Weili Nie, Arash Vahdat, and Anima Anandkumar. Fast training of diffusion models with masked transformers. *arXiv preprint arXiv:2306.09305*, 2023.

Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Lafite: Towards language-free training for text-to-image generation. *ArXiv preprint*, 2021.

A   FAE Encoder Structure

We merge the consecutive linear layers in the attention module and use a larger per-head dimension for the encoder. The structure of our encoder is shown in Figure 7.

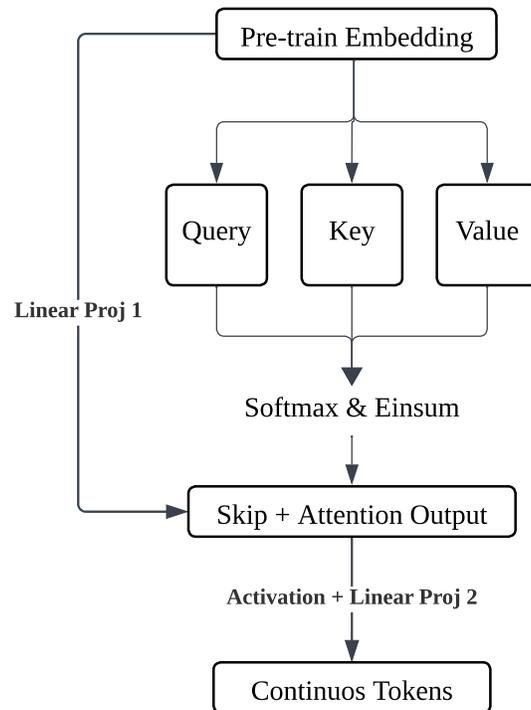**Figure 7**   Modified Attention

B Ablation on FlowMatching Timesteps Shift

Our ablation experiment demonstrates that applying Timesteps Shift accelerates convergence, mitigates the discrepancy across different latent token dimensions, and provides a modest improvement in the final FID score.

Model	CFG	64 Epochs	160 Epochs	320 Epochs
32-dim, ts=0.7	w/o	2.4087	1.9060	1.6836
	w/	–	1.6786	1.5131
32-dim, ts=0.5	w/o	2.3233	1.8501	1.7086
	w/	–	1.6735	1.5682
32-dim, ts=0.3	w/o	2.3220	1.8800	1.7743
	w/	–	1.7125	1.6227
48-dim, ts=0.5	w/o	2.4329	1.9546	1.6952
	w/	–	1.6797	1.5312
48-dim, ts=0.3	w/o	2.3599	1.9105	1.6911
	w/	–	1.6694	1.5423
64-dim, ts=0.2	w/o	2.4398	1.9549	1.7581
	w/	–	1.7563	1.5402

Table 8 Ablation results comparing different timesteps shift across different token dimension.

C rFID

Because the encoder training is disentangled with the image reconstruction loss, our rFID and tokenizer reconstruction fidelity lag behind methods such as VA-VAE that directly optimize reconstruction quality.

	SD-VAE	VA-VAE	FAE 32-dim	FAE 64-dim
rFID	0.73	0.28	0.68	0.66

Table 9 Reconstruction rFID comparison.

D FAE Hyper Parameters

Category	Field	Encoder	Decoder	Pixel Decoder	LDM	MMDiT	MMDiT 384x384
Architecture	Input dim.	16×16×1536	16×16×32	16×16×1536	16×16×32	16×16×32	24×24×64
	Output dim.	16×16×64	16×16×1536	256×256×3	16×16×32	16×16×32	24×24×64
	Hidden dim.	6144	1536	1024	1152	1024	1536
	Num. layers	1	6	24	28	16	24
	MLP Ratio	–	4	4	4	4	4
	Dim. per head	256	64	64	72	64	64
	Num. heads	24	24	16	16	16	24
	Total Params (M)	38.17	170.43	305.36	675.26	603.46	2017.84
Optimization	Training iters	1M		1M	2M	1M	1M
	Batch size	1024		512	512	512	512
	Optimizer	AdamW		AdamW	AdamW	AdamW	AdamW
	Peak LR	1e-4		1e-4	1e-4	1e-4	1e-4
	LR Scheduler	Cosine		Cosine	Constant	Constant	Constant
	Warmup	1000		1000	–	–	–
	(β_1, β_2)	(0.9, 0.999)		(0.9,0.999)	(0.9,0.999)	(0.9,0.999)	(0.9,0.999)
Interpolants	α_t	–	–	–	1-t	1-t	1-t
	σ_t	–	–	–	t	t	t
	w_t	–	–	–	σ_t	σ_t	σ_t
	Training objective	–	–	–	v-prediction	v-prediction	v-prediction
	Sampler	–	–	–	Euler-Maruyama (w/o CFG) Euler (w/ CFG)	Euler-Maruyama (w/o CFG) Euler (w/ CFG)	Euler-Maruyama (w/o CFG) Euler (w/ CFG)
	Sampling steps	–	–	–	250	250	250
	Guidance	–	–	–	0.9 (t=1∼0.9) 2.5 (t=0.7∼0)	1.5 (t=0.9∼0)	1.5 (t=0.9∼0)

The MMDiT 384×384 and its FAE encoder are only used for generating high quality examples provided in the paper. The three-partitioned CFG are only used for generating Main Results. The main results uses timesteps shift=0.4. All ablation experiments use single cfg scale for $t = 0.7 \sim 0.0$, and the cfg scale is grid searched in 0.1 fineness. The ablation experiments for FAE Model Structure are using SiT. The linear probing results are got from a separate encoder trained with DinoV2 register version with same parameters. And the ablation experiments for Token Dimension and Time Shift are using LightningDiT.

E Patch Embedding Similarity Maps

We compare the similarity between different patches inside single images. For each triplet of visualizations, the first image shows the similarity map computed from the DINOv2 embeddings, while the second shows the corresponding similarity map derived from our FAE latents. The third image displays the original image patch used as the query. The selected query patches are highlighted with a red rectangle, and darker colors in the similarity maps indicate higher similarity values.
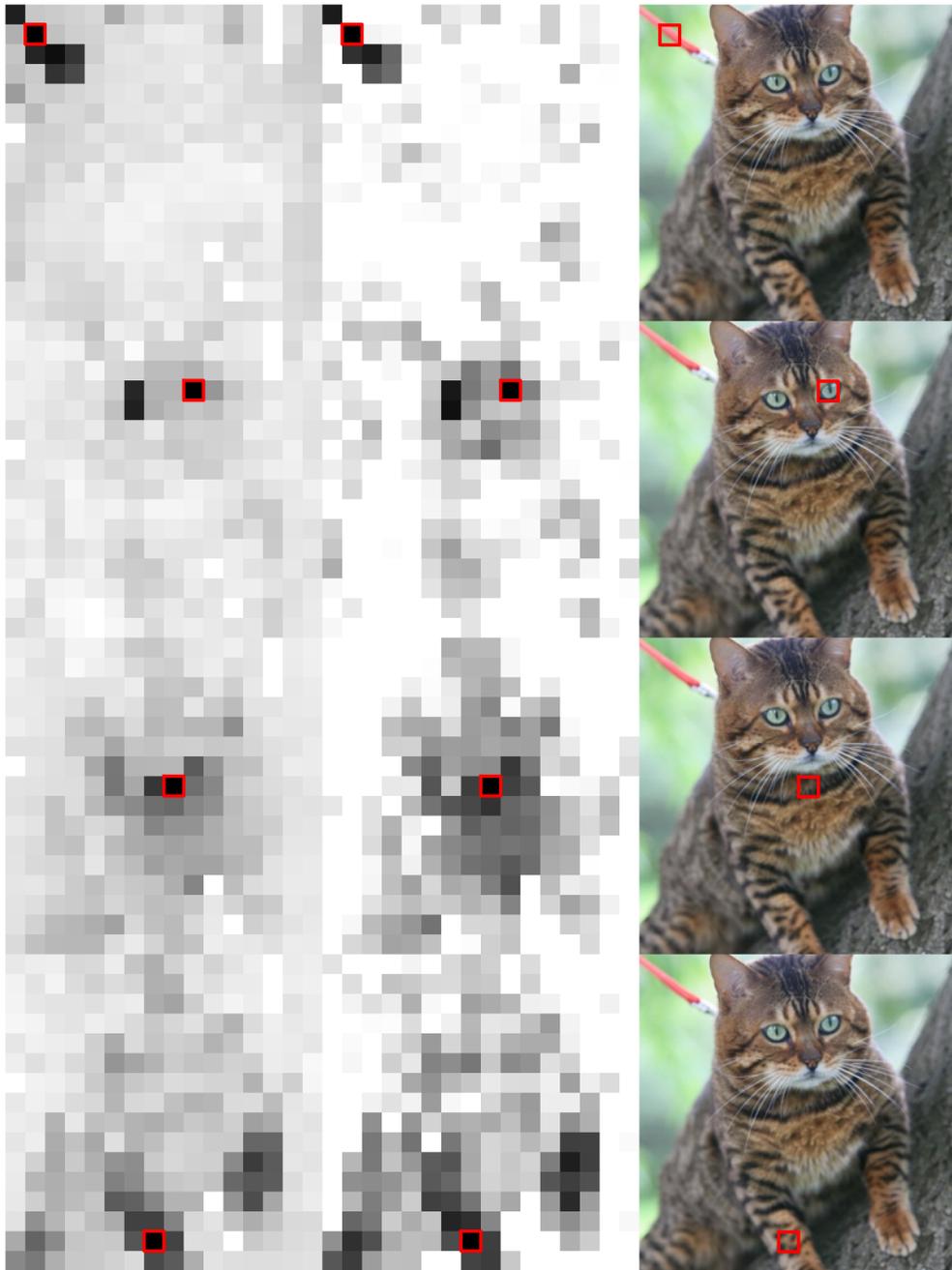

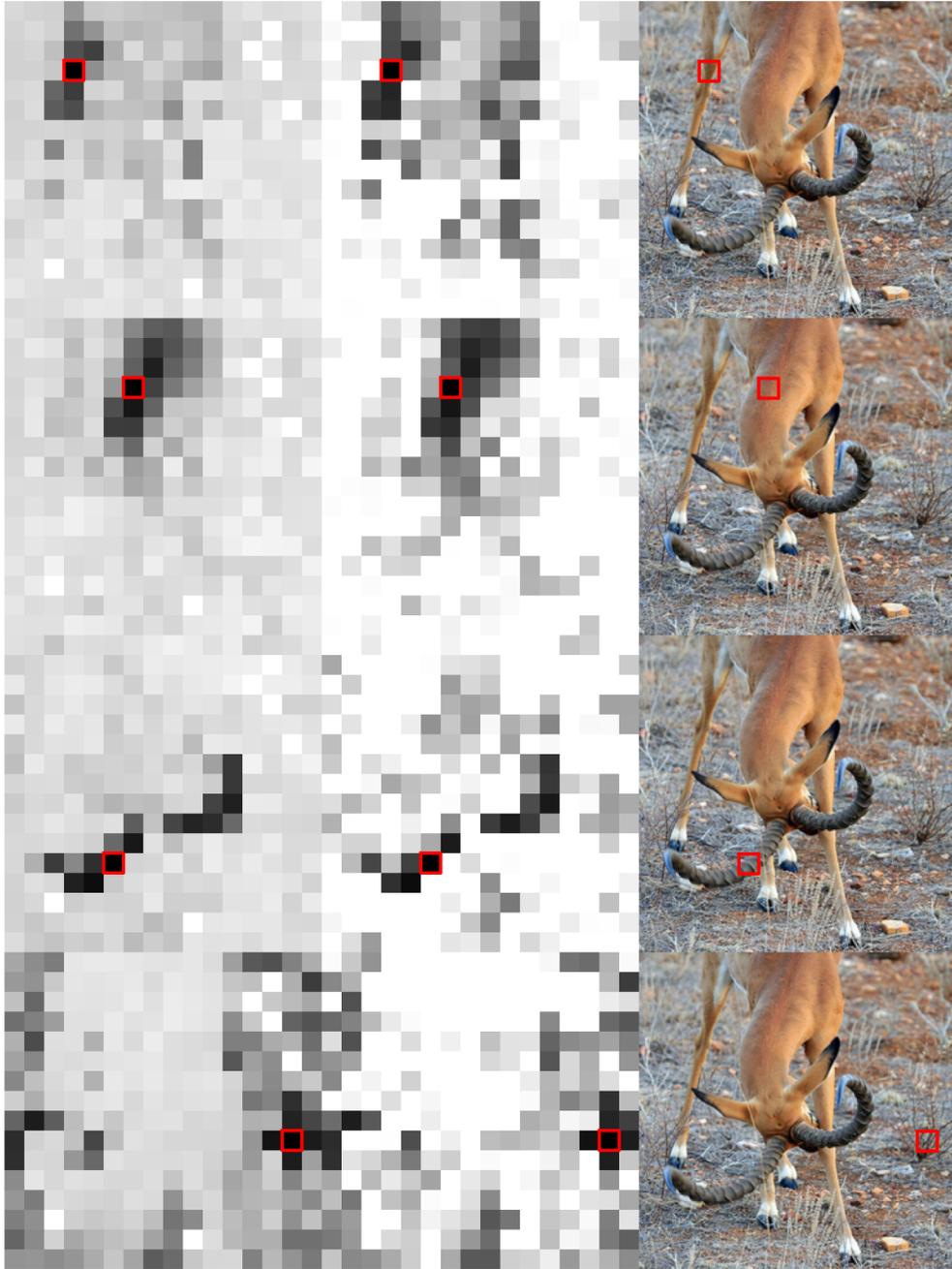**Figure 8** Similarity of a photo of cat.

4

Figure 9 Similarity of a photo of impala.

F Matching most Similar patch pair across two images

Our latents retain the cross-image patch–matching behavior characteristic of DINOv2: semantically corresponding regions across different images are still reliably matched using cosine similarity in the latent space. This suggests that FAE preserves fine-grained, part-level semantics rather than only coarse global information.

We first identify animal-related patches using K-Means clustering. From these, we randomly select patches in the first image and match each one to the patch in the second image with the highest cosine similarity. For each example, 16 patch pairs are selected and visualized.

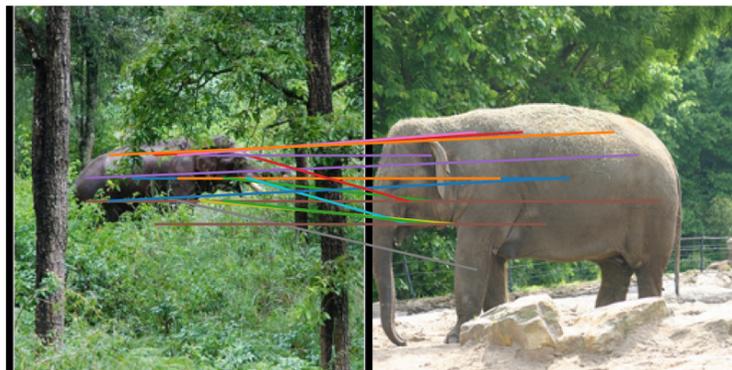**Figure 10** Matching most similar patch pair from two photo of bird.

Figure 11 Matching most similar patch pair from two photo of elephant.

G   Text-to-Image Prompts

The prompts for the text to image examples in Figure 5 are:
"a wooden bench under a large oak tree with warm sunlight streaming through branches and fallen leaves scattered below",
"a panoramic mountain ridge under soft morning clouds",
"a wooden arrow sign reading 'north trail' pointing into the woods",
"an alpine lake surrounded by steep cliffs, water perfectly still except for faint circular ripples, floating pollen creating tiny shimmering patterns on the surface",
"a window sticker reading 'welcome'",
"a snow leopard walking across snowy slope, faint pawprints trailing behind",
"a winter woodland with heavy snow draped asymmetrically across branches, faint animal tracks weaving between tree shadows under pale blue light",
"a sticky note attached to a monitor reading 'finish draft by 5 pm'",
"a small shop window sign reading 'local goods'",
"a tortoise lumbering through sunlit shrubs, shell etched with age patterns",
"a wooden produce crate stamped 'farm fresh' positioned in a sunlit garden shed beside tools",
"a rocky mountain meadow scattered with boulders and wildflowers under bright daylight"

H   STARFlow Examples

Figure 12   Random samples of ImageNet 256x256 generated by STarFlow model.

I Extra Examples from Siglip2 MMDiT 384×384 Model
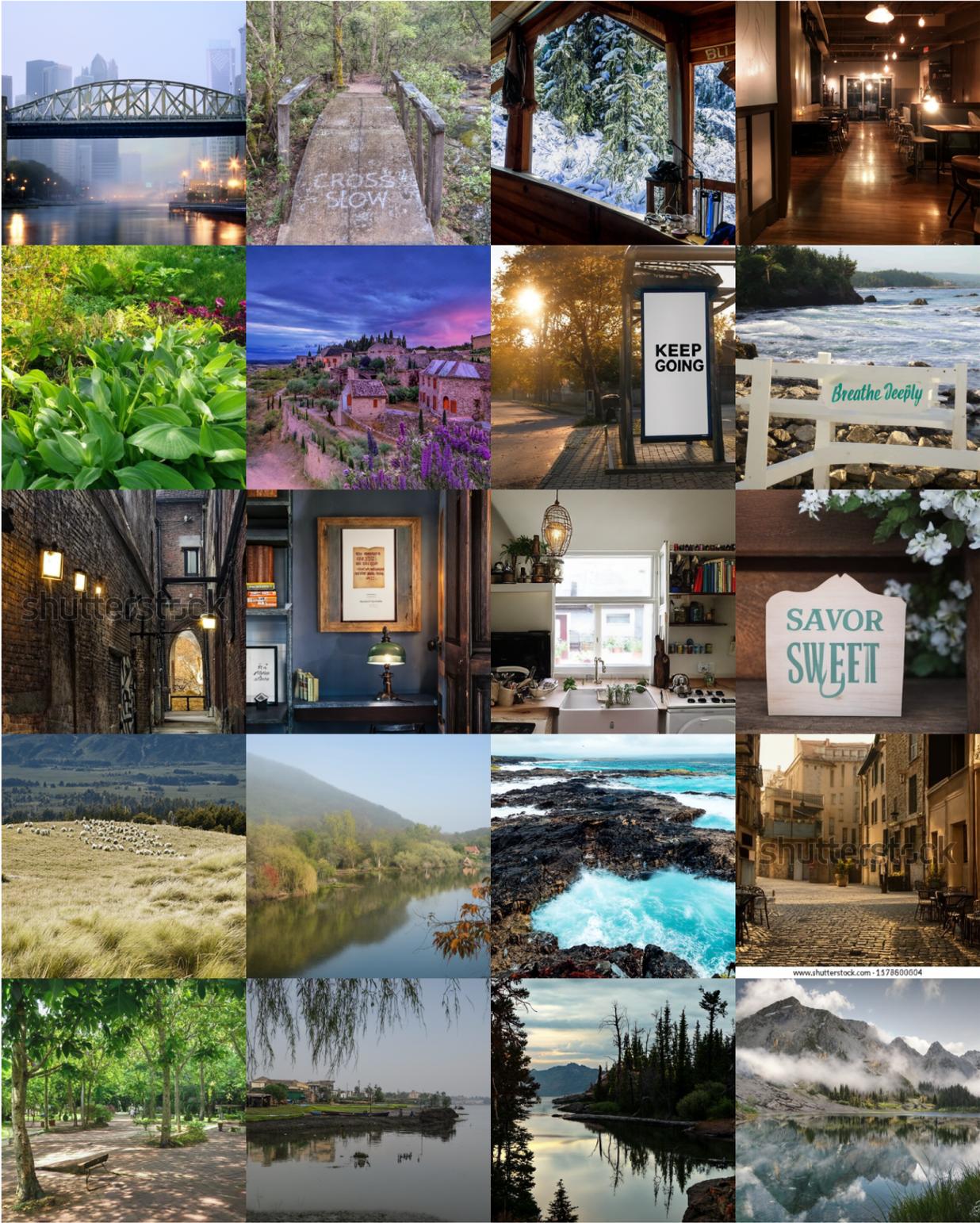

Figure 13 Random samples of Siglip2 MMDiT 384×384 Model.

The prompts for the text to image examples are:
"a foggy bridge spans a calm river reflecting muted city lights.",
"a forest bridge plank etched: 'cross slow'. water murmurs beneath the boards.",
"a hiking lodge interior has a carved plaque: 'rest; the mountains will wait.' snow drifts past the windows.",
"a small café interior glows softly as candles flicker on wooden tables.",
"a quiet backyard garden with warm sun patterns filtering through leaves.",
"a hillside dotted with tiny cottages beneath a lavender evening sky.",
"a bus stop poster saying: 'keep going'. golden morning light warms the street.",
"a seaside cabin porch sign saying: 'breathe deeply'. waves pulse against the rocks below.",
"a lantern-lit alleyway glowing softly between old brick walls.",
"a quiet library alcove features a framed message: 'seek answers, but also seek the calm between them.' warm lamplight glows against tall wooden shelves.",
"a kitchen window frames sun-washed herbs, bowls, and warm shelves.",
"an orchard bench plaque reading: 'savor sweet'. blossoms float in warm breezes.",
"a windy hillside covered in tall dry grass, each stalk catching light differently, distant sheep forming small irregular white clusters",
"a river winds beside a sleepy village, reflecting pale morning skies and drifting willow branches.",
"a stormy coastline where winds whip through rugged rocks and dark water.",
"a quiet european street lined with stone buildings glows under early dawn light as café chairs sit empty on cobblestones.",
"a quiet university quad filled with shaded benches and tree-lined paths.",
"a calm river flows beside a small town, reflecting the pale sky while fishermen prepare their nets and willow branches trail across the water's surface.",
"a serene lake mirrors the surrounding pines and layered mountain ridges during early dawn.",
"a remote mountain lake reflects drifting clouds and jagged peaks.",