REGLUE Your Latents with Global and Local Semantics for Entangled Diffusion

Giorgos Petsangourakis[1,2], Christos Sgouropoulos[1], Bill Psomas[3], Theodoros Giannakopoulos[1], Giorgos Sfikas[2], and Ioannis Kakogeorgiou[1]

[1] IIT, National Centre for Scientific Research "Demokritos"
[2] University of West Attica
[3] VRG, FEE, Czech Technical University in Prague

Abstract. Latent diffusion models (LDMs) achieve state-of-the-art image synthesis, yet their reconstruction-style denoising objective provides only indirect semantic supervision: high-level semantics emerge slowly, requiring longer training and limiting sample quality. Recent works inject semantics from Vision Foundation Models (VFMs) either externally via representation alignment or internally by jointly modeling only a narrow slice of VFM features inside the diffusion process, under-utilizing the rich, nonlinear, multi-layer spatial semantics available. We introduce `REGLUE` (Representation Entanglement with Global-Local Unified Encoding), a unified latent diffusion framework that jointly models (i) VAE image latents, (ii) compact local (patch-level) VFM semantics, and (iii) a global (image-level) `[CLS]` token within a single `SiT` backbone. A lightweight convolutional semantic compressor nonlinearly aggregates multi-layer VFM features into a low-dimensional, spatially structured representation, which is entangled with the VAE latents in the diffusion process. An external alignment loss further regularizes internal representations toward frozen VFM targets. On ImageNet 256×256, `REGLUE` consistently improves FID and accelerates convergence over `SiT-B/2` and `SiT-XL/2` baselines, as well as over `REPA`, `ReDi`, and `REG`. Extensive experiments show that (a) spatial VFM semantics are crucial, (b) non-linear compression is key to unlocking their full benefit, and (c) global tokens and external alignment act as complementary, lightweight enhancements within our global–local–latent joint modeling framework. The code is available at https://github.com/giorgospets/reglue.

Keywords: Latent Diffusion Models · Joint Image-Feature Synthesis

1 Introduction

Latent Diffusion Models (LDMs) have become a dominant choice for high-quality image synthesis, largely because they shift the generative paradigm to modeling a compact VAE latent space [50]. However, training such models is *challenging*: under a single denoising objective, the model must concurrently learn high-level semantics (*what* to generate, *e.g.* objects, layout, relations) and low-level visual details (*how* to generate it, *e.g.* fine-grained appearance) [63]. The reconstruction-style denoising loss provides semantic supervision only *indirectly*,

so semantic structure emerges slowly, limiting image quality and convergence speed [64].

To address these challenges, recent works leverages semantic representations from strong, pretrained Vision Foundation Models (VFMs), accelerating convergence and improving image quality [8, 29, 54, 62, 64, 71, 72]. `REPA` [64] proposes a representation *alignment* objective to distill VFM features as an external teacher to the diffusion model. `REG` [62] and `ReDi` [29] go a step further: they *jointly model* the image latent and the semantic signal inside the diffusion process. Regarding this signal, `REG` uses a single *image-level* (global) representation (*i.e.*, `[CLS]`), while `ReDi` employs *linearly* PCA-projected *patch-level* (local) VFM features.

Both `REG` and `ReDi` expose only a *narrow*, *low-capacity* semantic slice of the VFM to the diffusion model, under-utilizing the rich, nonlinear, multi-layer, and spatial semantics available. `REG`'s `[CLS]` offers informative image-level guidance, but is inherently non-spatial; fine-grained semantics are recovered via an external alignment loss as in `REPA` [64], which distills spatial semantics and boosts generative performance. In contrast, `ReDi` explicitly models patch-level semantics but its linear PCA projection restricts the representations to a low-dimensional linear subspace, limiting the richness and non-linear spatial information.

We argue that a critical design choice for effectively leveraging VFM features is to model spatial semantics within the diffusion model, while preserving their nonlinear, multi-layer information via compact learned representations produced by a *semantic compressor*. Specifically, jointly modeling patch-level features with VAE latents provides spatial guidance critical for capturing fine-grained structure and yields larger gains than either external feature alignment alone (`REPA`) or jointly modeling a single image-level token (`REG`). These signals remain complementary: an alignment loss and a global `[CLS]` token can be added as orthogonal auxiliary signals, but the primary improvements stem from spatial joint modeling of compressed patch features within the diffusion model (see Table 1).

In our work, we first train a lightweight convolutional *semantic compressor* that maps nonlinear, multi-layer VFM features into a compact, spatially structured, semantics-preserving representation. Then, we introduce a unified diffusion modeling approach that *jointly* models: (i) these compact patch-level (local) semantic features, (ii) an image-level (global) representation, and (iii) the VAE image latents. During training, we also apply a feature-alignment auxiliary loss that aligns diffusion internal features with VFM teacher representations, further improving image synthesis performance. The overview is shown in Figure 1. In summary, our contributions are:

1. We introduce `REGLUE` (*Representation Entanglement with Global–Local Unified Encoding*), a unified diffusion framework that jointly models image-level (global) and patch-level (local) VFM semantics with VAE latents, significantly boosting generative performance.
2. We propose a lightweight *semantic compressor* that aggregates multi-layer VFM features and maps them to a compact, semantics-preserving space. This compact representation enables significant gains via patch-level joint modeling with VAE latents, strongly improving synthesized image quality.

3. We show that, in REGLUE, (a) patch-level (local) semantics, (b) image-level (global) semantics, and (c) REPA-style representation alignment act synergistically, delivering substantial gains in image quality and training convergence, while keeping the diffusion model's parameters and inference-time compute essentially unchanged.
4. On ImageNet 256×256 generation benchmark, SiT-XL/2+REGLUE reaches the 1M-step performance of ReDi and REG using less than 30% and 80% of their iterations, respectively.

2 Related Work

Latent-variable generative modeling. Latent-variable models like Variational Autoencoders and Diffusion Denoising Probabilistic models are core building blocks of modern generative pipelines [50]. There are often underappreciated connections across these families [5, 45]: diffusion can be viewed as a hierarchical VAE with a fixed latent dimension and a frozen encoder [37], and both are trained via variational approximations [4]. Normalizing flows [27, 67] have likewise been analyzed through their links to diffusion [1, 65, 69]. VAEs are also closely related to probabilistic PCA, replacing the linear projection with a neural decoder [19, 59]. Analyzing models under a unified framework has repeatedly yielded progress: Scalable Interpolant Transformers (SiT) [38], alongside DiT [44] and Lightning DiT [63], an indispensable component of state-of-the-art generative frameworks [29, 53, 62–64]. Our work espouses an analogous rationale, focusing on a holistic, joint modeling of tokenizer (VAE latents) and VFM semantics within a single framework.

Representation alignment with VFM features. Latent diffusion typically uses a VAE first stage to obtain compact image latents, which are optimized for reconstruction rather than semantics; this can slow semantic emergence and curb the ability to represent abstract concepts [2, 32]. To mitigate this, works enhance the *first stage* by aligning or reshaping the latent space: VA-VAE [63] adds a VFM alignment loss, TexTok [66] injects text description embeddings, MAETok [9] argues for discriminative (non-variational) latents, and FA-VAE [39] separates low/high frequencies via wavelets. Complementarily, the *second stage* (the denoiser) can be aligned to VFMs: REPA [64] distills mid-block features to VFM targets and accelerates convergence; DDT [61] decouples encoder/decoder; REPA-E [33] pursues end-to-end VAE+denoiser alignment; and SVG [53] focuses on few-step generation. Our approach is complementary: instead of exposing a narrow semantic slice or relying solely on external representation alignment, we *jointly model* global and local VFM semantics with VAE latents via a frozen, lightweight semantic compressor, entangling them within a single diffusion backbone and thereby accelerating convergence and improving generative quality.

Joint feature generative modeling. A complementary line of work directly *models* foundation features as observed variables, conceptually related to multimodal diffusion that learns joint spaces across modalities (*e.g.* text-image-video-audio) rather than distilling from them. Examples include CoDi [58] for any-to-

any generation across modalities, and video methods that entangle appearance with motion signals like VideoJam [7]. Closer to our setting, `REG` [62] proposes a `SiT`-based model that incorporates compressed tokens with the addition of also modeling a VFM `[CLS]` token; Representation Diffusion (`ReDi`) [29] takes this approach one step further and models spatially referenced high-level features on equal grounds to VAE latents by a Diffusion Transformer. We argue that these prior works model VFM suboptimally, either by discarding useful spatial cues [62] or working with constrained, linear projections of high-level semantics.

Representation learning. Supervised convolutional networks [23] and, more recently, Vision Transformers (ViTs) [18] have established strong baselines for transferable visual features, but the field has largely shifted toward pretraining regimes that reduce or remove manual labels. Early self-supervised work relies on hand-crafted pretext tasks (*e.g.* patch permutation [41] or rotation prediction [20]), while modern approaches favor contrastive objectives [11,42] and self-distillation [6,21,43], yielding highly transferable features at scale. Transformer-era enables masked image modeling (MIM): from BEiT [3] and MAE [22] to hybrid variants like iBOT [73] and AttMask [26]. In parallel, vision–language pretraining learns joint embeddings from web-scale image–text pairs [52]. CLIP [48] popularizes this paradigm by aligning images and captions with a contrastive objective, yielding strong zero-shot recognition [15], retrieval [28,47], and segmentation [57] performance. SigLIP [68] refines the recipe by replacing the softmax with independent sigmoid losses, and SigLIP-2 [60] further improves transfer via stronger data and training strategies.

3 Method

3.1 Preliminaries

Scalable Interpolant Transformer (SiT). We adopt the `SiT` [38] framework, built on the stochastic–interpolant formulation [35,38]. Given a clean image $\mathbf{x}_*$ and a pretrained VAE encoder $\mathcal{E}_z$ that produces image latents $\mathbf{z}_* \in \mathbb{R}^{D_z \times H_z \times W_z}$, we consider the continuous-time stochastic interpolant:

$$\mathbf{z}_t \;=\; \alpha_t \mathbf{z}_* + \sigma_t \boldsymbol{\epsilon}_z, \qquad \boldsymbol{\epsilon}_z \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad t \in [0,1], \tag{1}$$

where $\alpha_0 = \sigma_1 = 1$ and $\alpha_1 = \sigma_0 = 0$, with α_t decreasing and σ_t increasing in t. `SiT` adopts a Transformer-based architecture with $\mathcal{K}$ stacked blocks, parameterizes the velocity field $\mathbf{v}_\theta(\mathbf{z}_t, t)$ and is trained with the standard velocity objective:

$$\mathbb{E}_{\mathbf{z}_*, \boldsymbol{\epsilon}_z, t}\Big[\; \big\| \mathbf{v}_\theta(\mathbf{z}_t, t) \;-\; \dot{\alpha}_t \mathbf{z}_* \;-\; \dot{\sigma}_t \boldsymbol{\epsilon}_z \big\|^2 \Big]. \tag{2}$$

Unless stated otherwise, we adopt the linear schedule $\alpha_t = 1 - t$ and $\sigma_t = t$, which yields constant derivatives $\dot{\alpha}_t = -1$ and $\dot{\sigma}_t = 1$.

Vision Foundation Model (VFM). We denote by $\mathcal{E}_v(\cdot)$ a pretrained VFM (*e.g.* `DINOv2`), which provides *patch-level* semantic features $\mathbf{f}_*^{(\ell)} \in \mathbb{R}^{D_f \times H_f \times W_f}$

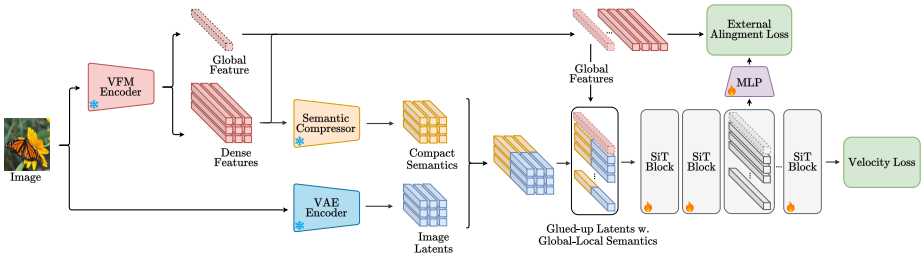

Fig. 1: Overview of REGLUE, Representation Entanglement with Global–Local Unified Encoding. The encoder of a vision foundation model (VFM) provides (i) *local* patch-level features and (ii) a *global* image-level feature (*i.e.* [CLS]). A lightweight semantic compressor (pre-trained offline) maps the patch features to *compact spatial semantics*. In parallel, a frozen VAE encoder produces *image latents*. We concatenate the latents, compressed semantics, and the global token and feed them to a SiT backbone [38], which *jointly models* all modalities with a velocity objective. Diffusion noise injection is omitted from the illustration. At a selected block, an MLP head applies an external alignment [64] to match hidden SiT features to clean VFM targets. At sampling, we decode the (jointly) generated VAE latents.

for each layer $\ell \in \{1, 2, \ldots, \mathcal{L}\}$, and a global *image-level* representation $\mathbf{cls}_* \in \mathbb{R}^{D_f}$. $H_f \times W_f$ denotes the patch grid size and D_f the feature dimensionality. In ViT-based encoders H_f, W_f, D_f remain fixed across layers.

3.2 Global–local representation entanglement

Our aim is to *jointly model* (i) VAE latents, (ii) local semantics (patch-level VFM features), and (iii) global semantics (*i.e.* image-level [CLS] token) within a single SiT model. We reuse the notation from Sec. 3.1. Given an image $\mathbf{x}_*$, $\mathcal{E}_z(\mathbf{x}_*) = \mathbf{z}_* \in \mathbb{R}^{D_z \times H_z \times W_z}$ denotes VAE latents, $\mathbf{cls}_* \in \mathbb{R}^{D_f}$ the global VFM token, and $\mathbf{s}_* \in \mathbb{R}^{D_s \times H_z \times W_z}$ patch-level *compressed* VFM features derived from $\{\mathbf{f}_*^{(\ell)}\}_{\ell=1}^{\mathcal{L}}$ and aligned to the VAE latents' grid (see Sec. 3.3 for a detailed description of the compressor).

Forward process. We first adopt a shared schedule (α_t, σ_t) to inject noise for all three entangled modalities:

$$\begin{aligned} \mathbf{z}_t &= \alpha_t \mathbf{z}_* + \sigma_t \boldsymbol{\epsilon}_z, & \boldsymbol{\epsilon}_z &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \\ \mathbf{s}_t &= \alpha_t \mathbf{s}_* + \sigma_t \boldsymbol{\epsilon}_s, & \boldsymbol{\epsilon}_s &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \\ \mathbf{cls}_t &= \alpha_t \mathbf{cls}_* + \sigma_t \boldsymbol{\epsilon}_{\text{cls}}, & \boldsymbol{\epsilon}_{\text{cls}} &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \end{aligned} \tag{3}$$

with independent noise terms and $t \in [0, 1]$.

Velocity objective. SiT parameterizes a velocity field $\mathbf{v}_\theta(\mathbf{z}_t, \mathbf{s}_t, \mathbf{cls}_t, t)$ over the joint state $(\mathbf{z}_t, \mathbf{s}_t, \mathbf{cls}_t)$ and is trained with the multimodal velocity loss:

$$\begin{aligned}\mathcal{L}_{\mathrm{v}} = \mathbb{E}\Big[& \left\|\mathbf{v}_\theta^z(\mathbf{z}_t, \mathbf{s}_t, \mathbf{cls}_t, t) - \dot{\alpha}_t \mathbf{z}_* - \dot{\sigma}_t \boldsymbol{\epsilon}_z\right\|_2^2 \\ & + \lambda_s \left\|\mathbf{v}_\theta^s(\mathbf{z}_t, \mathbf{s}_t, \mathbf{cls}_t, t) - \dot{\alpha}_t \mathbf{s}_* - \dot{\sigma}_t \boldsymbol{\epsilon}_s\right\|_2^2 \\ & + \lambda_{\mathrm{cls}} \left\|\mathbf{v}_\theta^{cls}(\mathbf{z}_t, \mathbf{s}_t, \mathbf{cls}_t, t) - \dot{\alpha}_t \mathbf{cls}_* - \dot{\sigma}_t \boldsymbol{\epsilon}_{\mathrm{cls}}\right\|_2^2\Big],\end{aligned} \tag{4}$$

where $\mathbf{v}_\theta^z$, $\mathbf{v}_\theta^s$ and $\mathbf{v}_\theta^{cls}$ correspond to the predictions for the VAE latent, global VFM token, and patch-level VFM features velocity; λ_s and λ_{cls} are weighting coefficients.

Tokenization and fusion. The SiT diffusion backbone operates on a sequence of tokens with a shared width D, so we aim to bring the different modalities to a common dimensional space. Concretely, we first *patchify*[4] $\mathbf{z}_t$ and $\mathbf{s}_t$, forming $\mathbf{z}'_t = \mathrm{patch}(\mathbf{z}_t) \in \mathbb{R}^{N \times D_{z'}}$ and $\mathbf{s}'_t = \mathrm{patch}(\mathbf{s}_t) \in \mathbb{R}^{N \times D_{s'}}$ (where $N{=}H_z W_z / p^2$ is the number of $p \times p$ patches). We then project each modality to the model width D with linear embedding layers:

$$\tilde{\mathbf{z}}_t = \mathbf{z}'_t \mathbf{W}_z, \quad \tilde{\mathbf{s}}_t = \mathbf{s}'_t \mathbf{W}_s, \quad \tilde{\mathbf{cls}}_t = \mathbf{cls}_t \mathbf{W}_{\mathrm{cls}}$$

where $\mathbf{W}_z \in \mathbb{R}^{D_{z'} \times D}$, $\mathbf{W}_s \in \mathbb{R}^{D_{s'} \times D}$, and $\mathbf{W}_{\mathrm{cls}} \in \mathbb{R}^{D_f \times D}$ are learned embedding matrices.

To combine and jointly model $\tilde{\mathbf{z}}_t$ (VAE latents) and $\tilde{\mathbf{s}}_t$ (local semantics), there are two straightforward options: (i) concatenate image latents and semantic features along the *sequence* dimension and pass $2N$ patch tokens through SiT, or (ii) *merge channel-wise* and keep a single grid of N tokens. We adopt (ii), avoiding the $2\times$ longer sequence and quadratic self-attention overhead of (i) (see Suppl. Sec. A.9). The global [CLS] is inherently a single token; thus, we always keep it as a separate token, adding a negligible throughput overhead. Finally, the input sequence to the SiT Transformer is:

$$\mathbf{h}_t^0 = \Big[\underbrace{\tilde{\mathbf{cls}}_t}_{1\times D}; \underbrace{(\tilde{\mathbf{z}}_t + \tilde{\mathbf{s}}_t)}_{N\times D}\Big] \in \mathbb{R}^{(1+N)\times D}, \tag{5}$$

where "[;]" denotes concatenation along token dimension.

Prediction heads. Let $\mathbf{h}_t^{\mathcal{K}} \in \mathbb{R}^{(1+N)\times D}$ denote the hidden sequence after the last ($\mathcal{K}$-th) SiT block. We obtain the per-modality velocity predictions as:

$$\begin{aligned}\mathbf{v}_\theta^z &= \mathrm{unpatch}\big(\mathbf{h}_t^{\mathcal{K}}[1{:}N]\,\mathbf{W}_{\mathrm{dec}}^z\big), \\ \mathbf{v}_\theta^s &= \mathrm{unpatch}\big(\mathbf{h}_t^{\mathcal{K}}[1{:}N]\,\mathbf{W}_{\mathrm{dec}}^s\big), \\ \mathbf{v}_\theta^{\mathrm{cls}} &= \mathbf{h}_t^{\mathcal{K}}[0]\,\mathbf{W}_{\mathrm{dec}}^{\mathrm{cls}}.\end{aligned} \tag{6}$$

[4] Partitioning a tensor $x \in \mathbb{R}^{C\times H\times W}$ into non-overlapping $p \times p$ tiles (here $p{=}2$; denoted SiT/2), flattening each tile, and stacking them row-major into a sequence of patch tokens, resulting in $x' \in \mathbb{R}^{(HW/p^2)\times(Cp^2)}$.

where $\mathbf{W}^{z}_{\text{dec}} \in \mathbb{R}^{D\times D_{z'}}, \mathbf{W}^{s}_{\text{dec}} \in \mathbb{R}^{D\times D_{s'}}$ are linear prediction heads, unpatch$(\cdot)$ reshapes the N patch tokens back to spatial tensors of shape $D_z \times H_z \times W_z$ and $D_s \times H_z \times W_z$, respectively; $\mathbf{W}^{\text{cls}}_{\text{dec}} \in \mathbb{R}^{D\times D_f}$ projects the global token.

External representation alignment. At a selected Transformer block $k \in \{1,\dots,\mathcal{K}\}$, we encourage `SiT` hidden tokens to stay close to *clean*, frozen VFM targets via a lightweight projector $\phi : \mathbb{R}^D \to \mathbb{R}^{D_f}$ and a cosine loss, following [64]. Let $\mathbf{h}^k_t \in \mathbb{R}^{(1+N)\times D}$ be the hidden sequence at block k. We form the target token sequence by concatenating the global token with patch-level VFM features:

$$\mathbf{y}_* \;=\; \left[\mathbf{cls}_* \,;\, \tilde{\mathbf{f}}^{(\mathcal{L})}_*\right] \in \mathbb{R}^{(1+N)\times D_f}. \tag{7}$$

where $\tilde{\mathbf{f}}^{(\mathcal{L})}_*$ denotes flattened spatial dimension to tokens.

The representation alignment loss is

$$\mathcal{L}_{\text{REPA}} = -\,\mathbb{E}\left[\frac{1}{N+1}\sum_{n=1}^{N+1} \text{sim}\Big(\mathbf{y}^{[n]}_*,\; \phi\big(\mathbf{h}^k_t\big)^{[n]}\Big)\right], \tag{8}$$

where we apply alignment on the `[CLS]` position and the N semantic patch tokens and sim is cosine similarity.

Total objective. We train with the multimodal velocity loss in Eq. (4) and the auxiliary alignment loss in Eq. (8):

$$\mathcal{L}_{\text{total}} \;=\; \mathcal{L}_{\text{v}} \;+\; \lambda_{\text{rep}}\,\mathcal{L}_{\text{REPA}}. \tag{9}$$

Sampling. To generate new samples, we employ the reverse-time SDE Euler-Maruyama [56] using $\mathbf{v}_\theta$ to obtain $(\mathbf{z}_0, \mathbf{s}_0, \mathbf{cls}_0)$ from Gaussian noise, and reconstruct the image via the (frozen) VAE decoder.

3.3 A lightweight spatial semantic compressor

Our goal is to jointly model VAE latents with VFM semantics. Naïvely fusing patch-level VFM features with image latents substantially widens the dimensionality to $D_z{+}D_f$ with $D_f \gg D_z$, biasing `SiT` capacity toward the representation modality and hurting generative quality (a phenomenon also addressed in `ReDi` [29] via PCA).

Convolutional autoencoder. Instead of a linear subspace, we introduce a *nonlinear*, lightweight *semantic compressor* $\mathcal{E}_\psi$ that preserves spatial structure while re-balancing dimensionality. We instantiate $\mathcal{E}_\psi$ as a shallow convolutional autoencoder, pretrain it once to reconstruct VFM features and then keep it frozen. Our compressor finally projects compressed features of dimension $D_s \ll \sum_\ell D_\ell$. An overview of the architecture can be found at Suppl. Figure 4.

Multi-layer aggregation. Since leveraging representations across depths has shown benefits to many scene understanding tasks [10, 12, 34, 36, 49, 70], we typically aggregate the multi-layer patch-level VFM features by channel-wise concatenation:

$$\mathbf{f}_* = \left[\mathbf{f}_*^{(1)}, \mathbf{f}_*^{(2)}, \ldots, \mathbf{f}_*^{(\mathcal{L})}\right] \in \mathbb{R}^{(\mathcal{L}\cdot D_f)\times H_f \times W_f}, \tag{10}$$

where "[,]" denotes concatenation along the channels.

Training and inference. A lightweight convolutional autoencoder $(\mathcal{E}_\psi, \mathcal{D}_\psi)$ is trained (offline) to reconstruct $\mathbf{f}_*$:

$$\min_{\psi} \mathbb{E}\Big[\, \|\mathcal{D}_\psi(\mathcal{E}_\psi(\mathbf{f}_*)) - \mathbf{f}_*\|_2^2 \,\Big], \tag{11}$$

We then *freeze* $\mathcal{E}_\psi$, produce compact spatial semantics $\mathcal{E}_\psi(\mathbf{f}_*) \in \mathbb{R}^{D_s \times H_f \times W_f}$, and spatially resample them to the VAE latent grid $(H_z \times W_z)$ resulting in $\mathbf{s}_* \in \mathbb{R}^{D_s \times H_z \times W_z}$, typically using bilinear resampling (See Suppl. Sec. A.10).

4 Experiments

4.1 Setup

Implementation details. We strictly follow the standard training protocols of `SiT` [38]. Our experiments are conducted on the ImageNet [16] dataset. Following the ADM preprocessing pipeline [17], all images are center-cropped and resized to 256×256 resolution. Each image is then encoded into a latent representation $\mathbf{z}_* \in \mathbb{R}^{4\times 32\times 32}$ using the pre-trained `SD-VAE-FT-EMA` [50]. Our main experiments are based on `SiT-B/2` models, which use a 2×2 patch size and are trained for 400K steps. To assess the impact of our approach at larger scales and longer training, we additionally train `SiT-XL/2` models for 1M steps. We maintain a batch size of 256 for all experiments.

Unless stated otherwise, for semantic feature extraction, we employ `DINOv2-B` [14,43], concatenating features from blocks 9-12 to obtain a 3072-channel (768×4) feature map at 16×16 spatial resolution. Our lightweight convolutional autoencoder $(\mathcal{E}_\psi, \mathcal{D}_\psi)$, with a hidden layer of 256, is pretrained for 25 epochs using a reconstruction loss (Eq. 11) to compress these maps. The encoder $\mathcal{E}_\psi$ produces a compact 16-channel, 16×16 latent representation.

For external representation alignment we follow the formulation in Sec. 3. We apply the external alignment using the local and global representations derived from the VFMs last layer, with layer k of the `SiT` backbone. We use k = 4 for `SiT-B/2` and k = 8 for `SiT-XL/2`. Regarding the total objective coefficients we set $\lambda_s = 1$, $\lambda_{cls} = 0.03$, and $\lambda_{rep} = 0.5$. More implementation details regarding the semantic compressor and `SiT`, are provided in Suppl. Sec. B.1 and Suppl. Sec. B.2.

Evaluation. In order to evaluate image generation quality, we report a standard set of quantitative metrics. These include Fréchet Inception Distance (FID) [24] for perceptual quality, sFID [40] for spatial coherence, Inception Score (IS) [51] for diversity, as well as Precision (Pre.) and Recall (Rec.) [30] to measure sample fidelity and distribution coverage, respectively. All metrics are computed using 50,000 generated samples, following the standard ADM evaluation suite [17].

Table 1: Impact of VFM semantics on `SiT-B/2` for improved generation. Results at 400K training steps. (a) Baseline `SiT-B/2` in Gray, (b) `REPA` in Purple, (d) `ReDi` in Light Green, (h) `REG` in Yellow, and (k-p) `REGLUE` (ours) in Light Cyan. ☑ denotes novel components proposed in our work. While the nonlinear patch-level semantics alone yield substantial gains, the other listed components provide additional improvements. †Setting (p) uses a stronger `DINOv3-B` VFM; for fairness and consistency with prior work, all other experiments adopt `DINOv2-B` as the default VFM.

	Local (Patch-level)			Global (Image-level)	External Alignment		FID
	Non-linear	Linear	Multi-Layer	[CLS]	Patch-Level	[CLS]	
(a)							33.0
(b)					✓		24.4
(c)				✓			25.7
(d)		✓					21.4
(e)		✓			✓		18.8
(f)				✓		✓	33.7
(g)				✓	✓		15.5
(h)				✓	✓	✓	15.2
(i)		✓		✓	✓	✓	17.4
(j)		✓	✓	✓	✓	✓	23.1
(k)	☑						14.3
(l)	☑				✓		14.1
(m)	☑			✓	✓	✓	13.7
(n)	☑		☑				13.3
(o)	☑		☑	✓	✓	✓	**12.9**
(p)	☑		☑	✓	✓	✓	**12.3**†

We report additional generation quality metrics in Suppl. Sec. A.5. For all experiments, we use Euler–Maruyama SDE sampling with 250 steps. When using Classifier-Free Guidance (CFG) [25], we set CFG scale at $w = 2.3$ and guidance interval to $[0, 0.9]$, following [31].

4.2 Analyzing VFM semantics for generation

We leverage our framework to investigate how diffusion modeling in `SiT-B/2` benefits from: (i) *which* semantics are modeled (global vs. local), (ii) *how* local semantics are compressed (linear vs. non-linear), and (iii) the *degree to which* an external representation-alignment objective contributes to generation quality. The corresponding design choices and their impact on FID are shown in Table 1.

Local (patch-level) outperform global (image-level) semantics. We first focus on representations modeled *directly* by the diffusion model, without any external alignment. We find that patch-level semantics clearly outperform global-only signals. Relying solely on the global `[CLS]` token (setting **(c)**) attains 25.7 FID, whereas modeling patch-level features (setting **(d)**, `ReDi`, linear PCA) improves to 21.4 FID. Both settings substantially outperform the baseline `SiT-B/2` backbone (33.0 FID, setting **(a)**), but the gap between **(c)** and **(d)** underscores that fine-grained *spatial* semantics are pivotal for improved generative modeling.

Non-linear compression unlocks local guidance. Replacing the linear PCA used in `ReDi` with our lightweight *non-linear* semantic compressor boosts patch-

level joint modeling: setting **(k)** reaches 14.3 FID without any alignment loss, an absolute 7.1 FID reduction over `ReDi` (setting **(d)**). Notably, this also surpasses the state-of-the-art `REG` baseline (setting **(h)**, 15.2 FID), even though `REG` combines global modeling with external alignment. Moreover, enriching local guidance by aggregating multi-layer VFM patch-level features before compression (setting **(n)**) further reduces FID to 13.3, *without* using any global token or external alignment. The results indicate that rich spatial semantics are a crucial signal, and non-linear compression is key to unlocking their full benefit. For more analysis about why our *non-linear* compression enhances diffusability, see Sec. 4.4 and Figure 3.

External alignment under local and global modeling. We analyze how `REPA` behaves when applied to *local* and/or *global* semantics. When the backbone does not jointly model patch tokens, only aligning *local* VFM features already provides a strong boost: the original `REPA` configuration (setting **(b)**) improves the default `SiT-B/2` baseline from 33.0 **(a)** to 24.4 FID. `REPA` also improves the patch-only setting of `ReDi` **(d)** to **(e)** (from 21.4 to 18.8 FID). A similar pattern appears when starting from a model that only jointly models the global `[CLS]` token: adding *local-only* external alignment (setting **(g)**) reduces FID from 25.7 **(c)** to 15.5, and including the global component in the alignment as well (setting **(h)**) further improves it to 15.2. This indicates that local patch alignment is the dominant source of improvement, while global alignment provides a smaller, complementary gain. In contrast, aligning *only* the global information without any local alignment (setting **(f)**) degrades performance from 25.7 **(c)** to 33.7 FID, suggesting that alignment on global features alone is unstable without spatial anchors. Finally, in our setting, adding `REPA` on top of non-linear patch-level modeling improves FID from 14.3 **(k)** to 14.1 **(l)**, showing that once strong spatial semantics are jointly modeled, external alignment acts as a mild but consistent performance complement.

`REGLUE`: Joint local-global-latent modeling. Building on these observations, we progressively add the global token and alignment to compressed local modeling. Adding REPA-style alignment on top of **(k)** yields setting **(l)** with 14.1 FID (a modest but consistent gain), indicating that external supervision *complements* joint local modeling. Incorporating the global `[CLS]` and aligning both local and global signals (setting **(m)**) further improves to 13.7 FID. Finally, aggregating multi-layer patch features (from the last four VFM blocks) before compression (setting **(o)**) forms our final `REGLUE` unified setting, achieving 12.9 FID. However, richer VFM semantics improve latent diffusion only when integrated appropriately. Our non-linear semantic compressor is the key mechanism; without it, simply adding more signals can be counterproductive. For instance, naïvely stacking the existing components (setting **(i)**) or just using multi-layer PCA (setting **(j)**) yields significantly worse results (17.4 and 23.12 FID, respectively) compared to `REGLUE` at 12.9 FID. To study the dependence of `REGLUE` on the underlying VFM, we also examine a stronger VFM, `DINOv3-B`. As reported in setting **(p)**, `DINOv3-B` yields the best result (FID 12.3), improving over `DINOv2-B` (FID 12.9) and indicating that `REGLUE` can effectively exploit a

more powerful semantic encoder. Nevertheless, to remain consistent with prior work and enable fair comparison, we adopt `DINOv2-B` as our default VFM in all main experiments. For a more in-depth analysis of our compressor, see Sec. 3.3.

4.3 Enhancing diffusion models

Accelerating convergence. Table 2(a) reports conditional ImageNet 256×256 results without classifier-free guidance (CFG) with a `SiT-B/2` backbone. `REGLUE` reaches 14.5 FID at 300K steps surpassing `REG` (15.2 at 400K) with 25% *fewer* iterations, and further improves to 12.9 at 400K. At 400K, `REGLUE` reduces FID by 60.9% vs. vanilla `SiT-B/2` (33.0), 47.1% vs. `REPA` (24.4), and 39.7% vs. `ReDi` (21.4). Notably, `REGLUE` inference-time throughput and parameters remain unchanged. Moving to a larger `SiT-XL/2` backbone and more training steps, in Table 3 we show conditional ImageNet 256 × 256 results (no CFG). At 200K steps, `REGLUE` achieves 4.6 FID, outperforming `REG` (5.0) and substantially surpassing `REPA` (11.1) and `ReDi` (12.5). Notably, `REGLUE` reaches 2.7 FID at 700K, matching `REG`'s 1M performance (2.7) with 30% fewer iterations. At 1M, `REGLUE` sets the best score (2.5) vs. `REG` (2.7), `ReDi` (5.1), and `REPA` (6.4). Importantly, `REGLUE` improves performance without incurring computational overhead, as measured by throughput (Table 2(a)). See also Suppl. Sec. A.8 for more details on computational efficiency.

Table 2: Conditional and unconditional generation. Comparison of `SiT-B/2` with `REPA`, `ReDi`, `REG`, and `REGLUE` on ImageNet 256 × 256 without classifier-free guidance (CFG). We report parameter count, inference-time throughput (img/s), iterations, and FID. Throughput measured on 1×A100 with batch size 64 and 250 sampling steps per image.

Model	#Params	Thr.	Iter.	FID↓
(a) Conditional Generation				
`SiT-B/2`	130M	4.1	400K	33.0
+ `REPA`	130M	4.1	400K	24.4
+ `ReDi`	130M	4.1	400K	21.4
+ `REG`	132M	3.8	400K	15.2
+ `REGLUE` (ours)	132M	3.8	300K	**14.5**
+ `REGLUE` (ours)	132M	3.8	400K	**12.9**
(b) Unconditional Generation				
`SiT-B/2`	130M	4.1	400K	59.8
+ `ReDi`	130M	4.1	400K	43.6
+ `REG`	132M	3.8	400K	29.7
+ `REGLUE` (ours)	132M	3.8	400K	**28.7**

Unconditional generation. We evaluate our method in the unconditional setting and summarize the results in Table 2(b). The findings are consistent with the analysis in the previous section. Our `REGLUE` achieves 52%, 34.2%, and 3.4% improvements over `SiT-B/2`, `ReDi`, and `REG`, respectively, demonstrating the effectiveness of nonlinear compression and joint local–global feature modeling. Remarkably, even in this more challenging unconditional setting, `REGLUE` (28.7 FID) substantially outperforms the conditional `SiT-B/2` baseline (33.0 FID).

State-of-the-art comparison. Table 4 reports quantitative results on ImageNet with classifier-free guidance. `REGLUE` improves over `REG` at matched epochs and closes the gap to longer-trained baselines. At 80 epochs, `REGLUE` lowers FID to 1.59 vs 1.86 for `REG`. Notably, `REGLUE` at 80 epochs matches the performance of 160-epoch `REG`. At 160 epochs, it further improves to 1.46 vs 1.59. Although trained for 5× fewer epochs than the 800-epoch variants (`REPA`, `ReDi`, `REG`), the

Table 3: Conditional generation. Comparison of SiT-XL/2 with REPA, ReDi, REG, and REGLUE on ImageNet 256 × 256 without classifier-free guidance (CFG) under comparable settings. We report parameter count, iterations, and FID.

Model	#Params	Iter.	FID↓
SiT-XL/2	675M	7M	8.3
+ REPA	675M	200K	11.1
+ ReDi	675M	200K	12.5
+ REG	677M	200K	5.0
+ REGLUE (ours)	677M	200K	**4.6**
+ REPA	675M	400K	7.9
+ ReDi	675M	400K	7.5
+ REG	677M	400K	3.4
+ REGLUE (ours)	677M	400K	**3.2**
+ ReDi	675M	700K	5.6
+ REGLUE (ours)	677M	700K	**2.7**
+ REPA	675M	1M	6.4
+ ReDi	675M	1M	5.1
+ REG	677M	1M	2.7
+ REGLUE (ours)	677M	1M	**2.5**

Table 4: Comparison with state-of-the-art. Quantitative results on ImageNet 256 × 256 with classifier-free guidance (CFG). REPA, ReDi, REG and REGLUE employ an SiT-XL/2 model. *Does not use SD-VAE latents. †Uses class-balanced evaluation protocol.

Model	Epochs	FID↓	sFID↓	IS↑	Pre.↑	Rec.↑
Autoregressive Models						
VAR	350	1.80	-	365.4	0.83	0.57
MagViTv2	1080	1.78	-	319.4	0.83	0.57
MAR	800	1.55	-	303.7	0.81	0.62
Latent Diffusion Models						
LDM	200	3.60	-	247.7	0.87	0.48
U-ViT-H/2	240	2.29	5.68	263.9	0.82	0.57
DiT-XL/2	1400	2.27	4.60	278.2	0.83	0.57
MaskDiT	1600	2.28	5.67	276.6	0.80	0.61
SD-DiT	480	3.23	-	-	-	-
SiT-XL/2	1400	2.06	4.50	270.3	0.82	0.59
FasterDiT	400	2.03	4.63	264.0	0.81	0.60
MDT	1300	1.79	4.57	283.0	0.81	0.61
Leveraging VFM's Representations						
ReDi	800	1.61	4.66	295.1	0.78	0.64
REPA	800	1.42	4.70	305.7	0.80	0.65
REG	800	1.36	4.25	299.4	0.77	0.66
RAE*	800	1.13	—	262.6	0.78	0.67
REPA-E*†	800	1.12	4.09	302.9	0.79	0.66
REG	80	1.86	4.49	321.4	0.76	0.63
REPA-E*	80	1.67	4.12	266.3	0.80	0.63
REGLUE (ours)	80	1.59	4.22	282.9	0.78	0.64
REG	160	1.59	4.36	304.6	0.77	0.65
REGLUE (ours)	160	1.46	4.23	293.9	0.78	0.65

160-epoch REGLUE remains competitive with models that leverage VFM representations and are trained for substantially longer (REPA, FID 1.42; REG, FID 1.36). Classifier-free guidance ablations are presented in Suppl. Table 12. We provide qualitative results in Sec. E, along with small-scale dataset and higher-resolution results in Suppl. Sec. A.7.

4.4 Semantic compressor impact

As we highlight in Sec. 3.3, the channel dimensionality of VFM representations is substantially higher than that of image latents, which can lead to degraded performance when fused naïvely. To mitigate this, ReDi [29] employs linear PCA to project the representations into an low-dimensional latent space; however, as we show in Table 1, this design choice is suboptimal. In contrast, we show that our non-linear CNN-based semantic compressor (Suppl. Figure 4) can substantially improve generation quality. In this section, we examine its main design choices: the compression dimensionality (Table 5), the compressor capacity (Table 6), the set of VFM layers used as input (Table 7), and quantify their effect on both sample quality and efficiency. We further measure how much semantic information is preserved under compression using downstream probing tasks (Table 2).

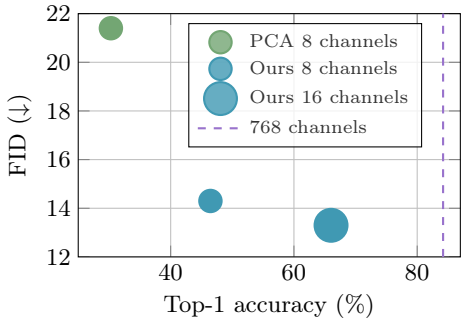

Fig. 2: Attentive probing accuracy vs. generation quality on ImageNet for `DINOv2` patch-level compression variants. Each point reports top-1 attentive probing accuracy [46] and FID, bubble size reflects feature dimensionality. Our nonlinear compressors (8/16 channels) achieve lower FID at higher accuracy compared to PCA-based compression, while the dashed line indicates the full 768-channel representation.

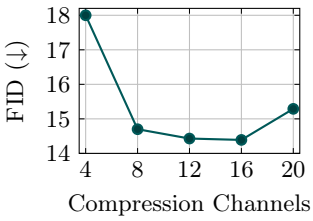

Table 5: Performance vs. compression channels. Ablation of the final compression channels, in `DINOv2` last layer's representation, using `SiT-B/2` trained for 400K steps without `REPA` loss.

Semantic preservation under compression. Figure 2 evaluates how well compressed patch-level features retain VFM semantics via attentive probing accuracy on ImageNet [46] and how this relates to generative quality (FID). Our non-linear semantic compressor preserves semantics much better than linear PCA: with only 8 channels it achieves substantially higher probing accuracy and lower FID than the PCA-compressed `ReDi` features, and increasing to 16 channels further improves both metrics, approaching the full 768-channel `DINOv2` baseline. In contrast, the PCA-based compression in `ReDi` yields low probing accuracy and only modest FID gains, indicating that non-linear, spatially structured compression is key to preserving semantic information while improving generation quality. Additional semantic preservation analysis with semantic segmentation experiments on Cityscapes [13] are presented in Suppl. Sec. A.1.

Semantic compression dimensionality. Table 5, examines how the number of compression channels affects generative performance. Aggressive compression (i.e., 4-channels) removes too much information, leading to degraded FID. Performance improves as we increase the number of channels up to 16, but degrades again at 20 channels. This suggests an optimal intermediate subspace: the compressed features preserve essential information to guide generation, yet are compact enough to stay balanced with the 4-channel image latents and not dominate the model's capacity. We therefore adopt 16 channels for `REGLUE` configuration.

Semantic compressor design. Our aim is to keep the compressor lightweight while preserving essential information. In Table 6, we present an ablation over different hidden-layer widths. Reducing the hidden dimensionality from 256 to 128 channels degrades FID, indicating that overly constrained bottlenecks limit the ability to preserve essential semantic information. Increasing the hidden size from 256 to 512 channels yields no further improvement in FID, while doubling

the model size and significantly reducing throughput, making this configuration inefficient. Pushing the capacity further (e.g., 1024 hidden size) leads to unstable compressor training. Moreover, increasing the model depth, via additional convolutional layers or residual blocks, exhibits the same instability. Overall, a shallow compressor with 256 hidden size offers the best balance between stability, efficiency, and generative performance.

Table 6: Compression model size comparison. Evaluation of compression encoder hidden layer sizes using the last 4 `DINOv2` layers and `SiT-B/2` as the diffusion model, trained for 400K steps. Representations are compressed to 16 channels. `REPA` loss is not applied. Input and Output layers are Conv2D(*in*, *out*) layers. Each middle block is a Residual Block which comprises two 3×3 convolutional layers (stride=1, padding=1) with Batch Normalization and ReLU activation. Samples in Throughput column are VFM local representations.

Input	Middle	Output	#Params	Throughput	FID↓
(3072,128)	(128,128)	(128,16)	3.8	32,304	14.2
(3072,256)	(256,256)	(256,16)	8.3	18,578	**13.3**
(3072,512)	(512,512)	(512,16)	18.9	9,510	**13.3**

Table 7: Multi-layer features effect. We compare the generation performance without using `REPA` loss across different sets of `DINOv2` layers. Results are reported at 400K training steps. In all runs VFM patch tokens are compressed to 16 channels and `SiT-B/2` is used. The compressor input layer is denoted by CNN In Dim column.

Dino Layers	CNN In Dim	FID↓
12	768	14.3
3,6,9,12	3072	16.9
9,10,11,12	3072	**13.3**

Choosing VFM layers for compression. In Table 7, we study how the choice of VFM layers fed into the semantic compressor affects generation performance. Motivated by [43], we compare three configurations of `DINOv2-B` patch features: (i) using only the last layer (12), (ii) using four intermediate layers (3, 6, 9, 12), and (iii) using the last four layers (9–12). In all cases, the compressed features are mapped to 16 channels and fed into the `SiT-B/2` diffusion model. Using only the final layer yields 14.3 FID, while including shallow intermediate layers (i.e., 3 & 6) degrades performance to 16.9 FID, indicating that early-layer features do not provide useful semantic guidance to the generation. In contrast, aggregating the last four layers (9–12) leads to 13.3 FID, suggesting that jointly compressing semantically rich, deeper VFM features provides the most beneficial signal for our framework. Ablations about KL regularization of the compressor can be found in Suppl. Table 8 and utilization of different VFMs in Suppl. Table 9.

Semantic compressor effectiveness. In Figure 3, we provide a *diffusability* diagnostic. Following [55], which identifies inordinate high-frequency components in latents as harmful for diffusion modeling, we compute DCT *frequency profiles* and observe that our compressor suppresses high-frequency components that diffusion models tend to mis-model. PCA exhibits a pronounced high-frequency tail, which is harder to model, whereas our semantic compressor strongly attenuates this tail and yields a spectrum closer to the RGB reference, indicating a more diffusion-friendly representation. Crucially, this does not sacrifice semantics

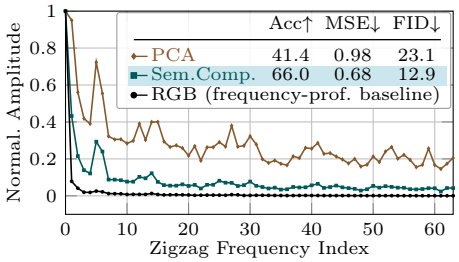

Fig. 3: Diffusability of Semantic Compressor *vs*. PCA. Frequency-profile comparison [55] of compression strategies, with RGB as reference. The inset shows that our compressor preserves more semantic signal (Acc.,MSE) while yielding a more diffusion-friendly representation by suppressing excess high-frequency components, leading to better generation quality (FID). Both map `DINOv2` last-4-layer channels to 16-D.

(higher ImageNet probing Acc., lower MSE). Collectively, these improvements enable better generative modeling (lower FID).

5 Conclusion

We introduced `REGLUE`, a unified generative model for latent diffusion that enables efficient entanglement of reconstruction-optimized and semantics-optimized image representations. By *jointly* modeling VAE latents with VFM patch-level & global semantics, coupled with *lightweight compression and aggregation* components, we have shown that `REGLUE` improves generation FID and accelerate convergence on ImageNet baselines by significant margins.

Acknowledgements We thank our colleague Theodoros Kouzelis for fruitful discussions. The project has received funding from the European High-Performance computing Joint Undertaking (JU) under grant agreement No 101234269 and the Greek Ministry of Digital Governance. Bill was supported by the EU Horizon Europe programme MSCA PF RAVIOLI (No. 101205297). AWS resources were provided by the National Infrastructures for Research and Technology GRNET and funded by the EU Recovery and Resiliency Facility. The publication/registration fees were partially covered by the University of West Attica. Also, we acknowledge the EuroHPC Joint Undertaking for awarding this project access to the EuroHPC supercomputer LEONARDO, hosted by CINECA (Italy) and the LEONARDO consortium through an EuroHPC Development Access call (Projects ID EHPC-DEV-2026D02-285 & EHPC-DEV-2026D03-137).

References

1. Albergo, M.S., Vanden-Eijnden, E.: Building normalizing flows with stochastic interpolants. In: ICLR (2023) 3

2. Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., Ballas, N.: Self-supervised learning from images with a joint-embedding predictive architecture. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15619–15629 (2023) 3
3. Bao, H., Dong, L., Piao, S., Wei, F.: Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254 (2021) 4
4. Bishop, C.M.: Pattern recognition and machine learning. Springer (2006) 3
5. Bond-Taylor, S., Leach, A., Long, Y., Willcocks, C.G.: Deep generative modelling: A comparative review of VAEs, GANs, normalizing flows, energy-based and autoregressive models. IEEE transactions on pattern analysis and machine intelligence **44**(11), 7327–7347 (2021) 3
6. Caron, M., Touvron, H., Misra, I., Jegou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV (2021) 4
7. Chefer, H., Singer, U., Zohar, A., Kirstain, Y., Polyak, A., Taigman, Y., Wolf, L., Sheynin, S.: Videojam: Joint appearance-motion representations for enhanced motion generation in video models. arXiv preprint arXiv:2502.02492 (2025) 4
8. Chen, B., Bi, S., Tan, H., Zhang, H., Zhang, T., Li, Z., Xiong, Y., Zhang, J., Zhang, K.: Aligning visual foundation encoders to tokenizers for diffusion models. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=ajnBafpqmE 2
9. Chen, H., Han, Y., Chen, F., Li, X., Wang, Y., Wang, J., Wang, Z., Liu, Z., Zou, D., Raj, B.: Masked autoencoders are effective tokenizers for diffusion models. In: ICML (2025) 3
10. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017) 7
11. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PmLR (2020) 4
12. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: CVPR (2022) 7
13. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR (June 2016) 13
14. Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers. arXiv preprint arXiv:2309.16588 (2023) 8
15. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009). https://doi.org/10.1109/CVPR.2009.5206848 4
16. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255 (2009) 8
17. Dhariwal, P., Nichol, A.Q.: Diffusion models beat GANs on image synthesis (2021) 8
18. Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 4
19. Ghojogh, B., Ghodsi, A., Karray, F., Crowley, M.: Factor analysis, probabilistic principal component analysis, variational inference, and variational autoencoder: Tutorial and survey. arXiv preprint arXiv:2101.00734 (2021) 3

20. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. arXiv preprint arXiv:1803.07728 (2018) 4
21. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems **33**, 21271–21284 (2020) 4
22. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022) 4
23. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 4
24. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017) 8
25. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022) 9
26. Kakogeorgiou, I., Gidaris, S., Psomas, B., Avrithis, Y., Bursuc, A., Karantzalos, K., Komodakis, N.: What to hide from your students: Attention-guided masked image modeling. In: European Conference on Computer Vision. pp. 300–318. Springer (2022) 4
27. Kingma, D.P., Dhariwal, P.: Glow: Generative flow with invertible 1x1 convolutions. Advances in neural information processing systems **31** (2018) 3
28. Kordopatis-Zilos, G., Stojnić, V., Manko, A., Suma, P., Ypsilantis, N.A., Efthymiadis, N., Laskar, Z., Matas, J., Chum, O., Tolias, G.: Ilias: Instance-level image retrieval at scale. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 14777–14787 (2025) 4
29. Kouzelis, T., Karypidis, E., Kakogeorgiou, I., Gidaris, S., Komodakis, N.: Boosting generative image modeling via joint image-feature synthesis. In: NeurIPS (2025) 2, 3, 4, 7, 12
30. Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., Aila, T.: Improved precision and recall metric for assessing generative models. Advances in neural information processing systems **32** (2019) 8
31. Kynkäänniemi, T., Aittala, M., Karras, T., Laine, S., Aila, T., Lehtinen, J.: Applying guidance in a limited interval improves sample and distribution quality in diffusion models (2024) 9
32. LeCun, Y.: A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. Open Review **62**(1), 1–62 (2022) 3
33. Leng, X., Singh, J., Hou, Y., Xing, Z., Xie, S., Zheng, L.: REPA-E: Unlocking VAE for end-to-end tuning with latent diffusion transformers. In: ICCV (2025) 3
34. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR (2017) 7
35. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=PqvMRDCJT9t 4
36. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (June 2015) 7
37. Luo, C.: Understanding diffusion models: A unified perspective. arXiv preprint arXiv:2208.11970 (2022) 3

38. Ma, N., Goldstein, M., Albergo, M.S., Boffi, N.M., Vanden-Eijnden, E., Xie, S.: Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In: ECCV. p. 23–40 (2024) 3, 4, 5, 8
39. Medi, T., Wang, H.Y., Rampini, A., Keuper, M.: FAVAE-effective frequency aware latent tokenizer. In: NeurIPS 2025 Workshop: Reliable ML from Unreliable Data (2025) 3
40. Nash, C., Menick, J., Dieleman, S., Battaglia, P.W.: Generating images with sparse representations. arXiv preprint arXiv:2103.03841 (2021) 8
41. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: European conference on computer vision. pp. 69–84. Springer (2016) 4
42. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018) 4
43. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024), https://openreview.net/forum?id=a68SUt6zFt 4, 8, 14
44. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4195–4205 (2023) 3
45. Prince, S.J.: Understanding deep learning. MIT press (2023) 3
46. Psomas, B., Christopoulos, D., Baltzi, E., Kakogeorgiou, I., Aravanis, T., Komodakis, N., Karantzalos, K., Avrithis, Y., Tolias, G.: Attention, please! revisiting attentive probing for masked image modeling. arXiv preprint arXiv:2506.10178 (2025) 13
47. Psomas, B., Retsinas, G., Efthymiadis, N., Filntisis, P., Avrithis, Y., Maragos, P., Chum, O., Tolias, G.: Instance-level composed image retrieval. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025) 4
48. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021) 4
49. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: CVPR (2021) 7
50. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022) 1, 3, 8
51. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. Advances in neural information processing systems **29** (2016) 8
52. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: Laion-5b: An open large-scale dataset for training next generation image-text models. vol. 35, pp. 25278–25294 (2022) 4
53. Shi, M., Wang, H., Zheng, W., Yuan, Z., Wu, X., Wang, X., Wan, P., Zhou, J., Lu, J.: Latent diffusion model without variational autoencoder. arXiv preprint arXiv:2510.15301 (2025) 3

54. Shi, M., Wang, H., Zheng, W., Yuan, Z., Wu, X., Wang, X., Wan, P., Zhou, J., Lu, J.: Latent diffusion model without variational autoencoder. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=kdpeJNbFyf 2
55. Skorokhodov, I., Girish, S., Hu, B., Menapace, W., Li, Y., Abdal, R., Tulyakov, S., Siarohin, A.: Improving the diffusability of autoencoders. In: ICML (2025) 14, 15
56. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020) 7
57. Stojnić, V., Kalantidis, Y., Matas, J., Tolias, G.: Lposs: Label propagation over patches and pixels for open-vocabulary semantic segmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 9794–9803 (2025) 4
58. Tang, Z., Yang, Z., Zhu, C., Zeng, M., Bansal, M.: Any-to-any generation via composable diffusion. Advances in Neural Information Processing Systems **36**, 16083–16099 (2023) 3
59. Tipping, M.E., Bishop, C.M.: Probabilistic principal component analysis. Journal of the Royal Statistical Society Series B: Statistical Methodology **61**(3), 611–622 (1999) 3
60. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., Hénaff, O., Harmsen, J., Steiner, A., Zhai, X.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025) 4
61. Wang, S., Tian, Z., Huang, W., Wang, L.: DDT: Decoupled diffusion transformer. arXiv preprint arXiv:2504.05741 (2025) 3
62. Wu, G., Zhang, S., Shi, R., Gao, S., Chen, Z., Wang, L., Chen, Z., Gao, H., Tang, Y., Yang, J., et al.: Representation entanglement for generation: Training diffusion transformers is much easier than you think. NeurIPS (2025) 2, 3, 4
63. Yao, J., Yang, B., Wang, X.: Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025) 1, 3
64. Yu, S., Kwak, S., Jang, H., Jeong, J., Huang, J., Shin, J., Xie, S.: Representation alignment for generation: Training diffusion transformers is easier than you think. In: ICLR (2025) 2, 3, 5, 7
65. Zand, M., Etemad, A., Greenspan, M.: Diffusion models with deterministic normalizing flow priors. TMLR (2024) 3
66. Zha, K., Yu, L., Fathi, A., Ross, D.A., Schmid, C., Katabi, D., Gu, X.: Language-guided image tokenization for generation. In: CVPR. pp. 15713–15722 (2025) 3
67. Zhai, S., Zhang, R., Nakkiran, P., Berthelot, D., Gu, J., Zheng, H., Chen, T., Bautista, M.A., Jaitly, N., Susskind, J.: Normalizing flows are capable generative models. In: ICML (2025) 3
68. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid Loss for Language Image Pre-Training . In: ICCV. pp. 11941–11952 (2023) 4
69. Zhang, Q., Chen, Y.: Diffusion normalizing flow. In: NeurIPS. pp. 16280–16291 (2021) 3
70. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR (2017) 7
71. Zheng, B., Ma, N., Tong, S., Xie, S.: Diffusion transformers with representation autoencoders. arXiv preprint arXiv:2510.11690 (2025) 2

72. Zheng, B., Ma, N., Tong, S., Xie, S.: Diffusion transformers with representation autoencoders. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=0u1LigJaab 2
73. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: ibot: Image bert pre-training with online tokenizer. International Conference on Learning Representations (ICLR) (2022) 4