SHOULD WE SKIP DIFFUSION?

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion models learn semantic representations while generating images. In the Decoupled Diffusion Transformer (DDT), a condition encoder provides features that guide a velocity decoder in denoising. To enable effective denoising at all noise levels, these features must capture both high-level abstract structures and low-level details. However, skip/residual connections in the encoder allow shallow features to bypass successive transformations, which may limit progressive abstraction, or at least make it difficult to disentangle different levels of abstraction. We propose DDT-RFE, which removes the residual connections around the Self-Attention and MLP operations in each encoder block while maintaining stable training. To retain the information that abstraction discards but that the decoder still needs, we fuse the input patch embedding with intermediate and final encoder features to form the encoder output. The decoder thus has access to information from multiple encoder depths, while each encoder block is able to learn more abstract representations. DDT-RFE achieves overall improvements over DDT across visual understanding tasks, including image classification, semantic segmentation, object discovery, and semantic correspondence, while using fewer encoder blocks. It also achieves a lower FID for image generation on ImageNet.

1 INTRODUCTION

Diffusion transformers (Peebles & Xie, 2023) have emerged as a dominant backbone for image generation, achieving remarkable fidelity and diversity across a wide range of visual domains. Beyond synthesis, their intermediate activations encode semantic information that supports downstream visual recognition and dense prediction tasks (Stracke et al., 2025; Gan et al., 2025; Helbling et al., 2025; Xiang et al., 2023). Crucially, representation learning and image synthesis are not independent processes: in standard diffusion transformers, semantic abstraction and fine-grained signal reconstruction are implicitly coupled within the same denoising network. Recent work has begun to explicitly decouple these two roles (Yu et al., 2026; Zheng et al., 2024). Decoupled Diffusion Transformer (DDT) (Wang et al., 2026), for example, introduces a "condition encoder" for learning representations and a "velocity decoder" for signal reconstruction. The condition encoder retains the standard diffusion transformer block design and transforms noisy inputs into representations that provide the decoder with information useful for reconstruction, including abstract structures and low-level image details.

However, how the encoder develops abstract features and combines features at different levels affect the quality of the representations that are provided to the decoder. In DDT, each encoder block transforms its input and integrates the resulting features with the input through residual connections (He et al., 2016). Stacking blocks allows features to undergo successive transformations, which supports abstract representations. At the same time, residual connections carry information from earlier blocks into the inputs of later transformations. Although this can retain information useful for reconstruction, we hypothesize that the repeated accumulation of shallow features may hinder abstraction from the noisy input (Ji et al., 2026a; Zhang et al., 2026). As shown in Fig. 1, object structure remains difficult to distinguish in the visualized DDT encoder features, even in deeper blocks. This raises a central question: can the encoder develop more abstract features through successive transformations, while combining them with earlier features at its output, to provide the decoder with information at multiple levels of abstraction?

To address this, we propose DDT with a Residual-Free Encoder (DDT-RFE), which combines a residual-free condition encoder with explicit fusion of intermediate and final encoder features to

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

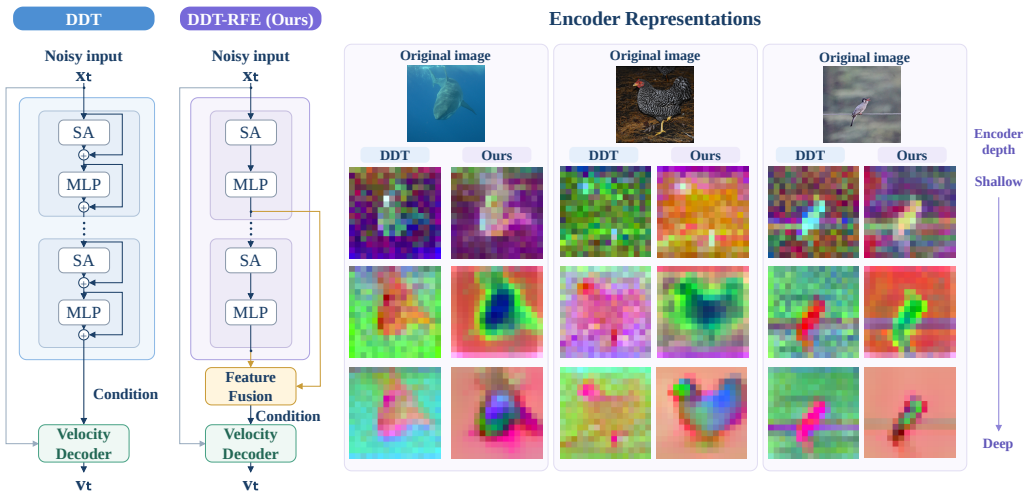


Figure 1: **Left:** Overall architectures of DDT and DDT-RFE (our method). **Right:** Encoder representations from shallow to deep layers at timestep $t = 0.5$. Features are projected into three dimensions using PCA and visualized as RGB images. Our method produces more abstract representations with clean backgrounds and salient objects. More visualizations are shown in Figs. 5 to 7

condition the decoder. We remove the residual connections around the Self-Attention and MLP operations in each encoder block, eliminating the identity shortcuts through which earlier features bypass these transformations. This design encourages progressive abstraction through successive transformations. In order to improve the convergence of the residual-free encoder, we propose an initialization scheme inspired by Ji et al. (2026a) but with fewer hyperparameters to tune. To retain access to information that is lost during abstraction but may still be needed for reconstruction, we fuse the input patch embedding and intermediate encoder features with the final encoder block output to form the decoder's condition. The decoder can therefore access representations at different levels, without requiring all useful information to persist in the final encoder block representation.

We evaluate DDT-RFE on four visual tasks and image generation. The results show that when DDT-RFE's encoder is ultilized for visual tasks, it achieves overall improvements in representation quality while remaining efficient with nearly 40% fewer encoder blocks. On ImageNet at 256×256 resolution, DDT-RFE reduces FID from 8.98 to 7.68 at the L scale and from 7.27 to 6.81 at the XL scale under the same training budget. Our main contributions are as follows:

- We propose DDT-RFE, which combines a residual-free condition encoder, a new initialization scheme, and feature fusion to improve representation learning and image generation.
- Qualitatively, our feature visualizations show clearer object structure and cleaner backgrounds in the representations learned by DDT-RFE than in those learned by DDT.
- Quantitatively, DDT-RFE achieves overall improvements over DDT on image classification, semantic segmentation, object discovery, and semantic correspondence with fewer encoder blocks. It also achieves a lower FID for image generation on ImageNet.

2 Related Work

Diffusion transformer architectures. DiT (Peebles & Xie, 2023) establishes a scalable backbone for latent diffusion built from a stack of transformer blocks. U-DiT (Tian et al., 2024) adds long skip connections that carry shallow features to deeper layers. Other designs separate the learning of semantic representations from reconstruction. In pixel-space diffusion, PixelDiT (Yu et al., 2026) models global semantics at the patch level in an encoder and refines details at the pixel level in a decoder. DDT (Wang et al., 2026) uses a condition encoder to learn representations that condition

a velocity decoder for signal reconstruction. We build on the condition encoder of DDT so that the representations learned in the encoder can be studied in isolation from the decoder.

Representation learning with diffusion models. Diffusion denoising has been studied as a self-supervised objective for learning visual representations that transfer effectively to downstream tasks. Through pretraining on image generation, Xiang et al. (2023) demonstrate that diffusion models learn linearly separable intermediate representations and achieve image classification performance comparable to that of contemporary contrastive methods. For diffusion transformers, DiTF (Gan et al., 2025) investigates the capabilities of DiTs for accurate dense correspondence and proposes a training-free method to extract semantically discriminative features from DiTs, achieving superior performance across multiple benchmarks. Furthermore, CleanDIFT (Stracke et al., 2025) introduces a lightweight, unsupervised fine-tuning method that enables diffusion backbones to provide high-quality, noise-free semantic features for downstream visual tasks. Our work examines how the condition encoder shapes representations, with a focus on progressive abstraction, and how information that may be discarded during this process can remain accessible to the decoder for reconstruction.

Residual connections in transformers. Residual connections affect both optimization and the learned representations. Ji et al. (2025) link the training difficulties caused by removing attention shortcuts to poor network conditioning in self-attention and analyze how residual pathways improve network conditioning. To address this optimization challenge, subsequent work develops initialization and optimization strategies that enable effective training of residual-free transformers (Ji et al., 2026a;b). Beyond optimization, residual pathways allow information from earlier blocks to persist in deeper representations (Raghu et al., 2021; Zhang et al., 2026). Ji et al. (2026a) further report that residual-free vision transformers learn more abstract representations and achieve improved performance on dense prediction tasks. An additional consideration in our setting is how to include more abstract features in the representations provided to the decoder while retaining other information useful for reconstruction. Our work examines this question in DDT by removing residual connections within the condition encoder and aggregating features from different depths at its encoder output.

3 Preliminaries

3.1 Decoupled Diffusion Transformer

DDT (Wang et al., 2026) provides a dedicated condition encoder for semantic extraction and a velocity decoder for signal reconstruction, rather than relying on homogeneous stacks of DiT/SiT blocks. This explicit separation makes DDT particularly suitable for analyzing how information flow within the condition encoder affects representation learning.

DDT consists of a condition encoder E_ϕ and a velocity decoder D_θ, both operating in the latent space of a pretrained Variational Autoencoder (VAE)[1]. The VAE encoder maps each image to a latent representation, which we write in vectorized form as $\mathbf{x}_0 \in \mathbb{R}^{d_x}$.

The noisy latent $\mathbf{x}_t$ and noise $\boldsymbol{\epsilon}$ have the same dimension and a noisy latent is constructed as

$$\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \sigma_t \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{1}$$

where $t \in [0, 1]$ is the diffusion timestep, $\alpha_t = 1 - t$ and $\sigma_t = t$ specify the signal and noise coefficients so that $\mathbf{x}_0$ is the clean latent and $\mathbf{x}_1$ is pure noise. The encoder maps the noisy latent $\mathbf{x}_t$ to a representation under timestep and class conditioning. This representation is combined with the timestep embedding to condition the decoder, which takes $\mathbf{x}_t$ as its input:

$$\mathbf{Z}_t = E_\phi(\mathbf{x}_t; t, y), \tag{2}$$

$$\hat{\mathbf{v}}_t = D_\theta(\mathbf{x}_t; \mathbf{Z}_t, t). \tag{3}$$

where y is the class label and $\mathbf{Z}_t \in \mathbb{R}^{N \times D}$ is the encoder output feature matrix, with N tokens and feature dimension D. We focus on the condition encoder to examine how residual connections shape the representations it learns.

[1] https://huggingface.co/stabilityai/sd-vae-ft-ema

3.2 CONDITION ENCODER ARCHITECTURE

The noisy latent is partitioned into non-overlapping patches and each patch is flattened and linearly projected to a D-dimensional token, yielding $\mathbf{H}_0 \in \mathbb{R}^{N\times D}$, where N is the number of patches. Let $\mathbf{H}_\ell \in \mathbb{R}^{N\times D}$ denote the output of block ℓ, for $\ell = 1, \ldots, L$. For clarity, we omit timestep t.

Conditioning. Timestep and class information are injected through adaptive layer normalization (AdaLN) (Peebles & Xie, 2023). A conditioning vector $\mathbf{c} = \mathrm{SiLU}\big(e_t(t) + e_y(y)\big) \in \mathbb{R}^D$ is mapped by a per-block affine layer to shift, scale, and gate vectors $\boldsymbol{\beta}_i^\ell, \boldsymbol{\gamma}_i^\ell, \mathbf{g}_i^\ell \in \mathbb{R}^D$ for the attention ($i = 1$) and MLP ($i = 2$). Here, $e_t(\cdot)$ and $e_y(\cdot)$ denote the timestep embedding function and the class embedding function, respectively.

Residual Encoder Block. Each encoder block applies the following residual updates:

$$\tilde{\mathbf{H}}_\ell = \mathbf{H}_{\ell-1} + \mathbf{g}_1^\ell \odot \mathcal{A}_\ell(\mathbf{H}_{\ell-1}), \tag{4}$$

$$\mathbf{H}_\ell = \tilde{\mathbf{H}}_\ell + \mathbf{g}_2^\ell \odot \mathcal{M}_\ell(\tilde{\mathbf{H}}_\ell), \tag{5}$$

where the Self-Attention and MLP transformations are defined as

$$\begin{aligned} \mathcal{A}_\ell(\mathbf{H}) &= \mathrm{SA}_\ell\Big(\mathrm{Norm}_1^\ell(\mathbf{H}) \odot (\mathbf{1} + \boldsymbol{\gamma}_1^\ell) + \boldsymbol{\beta}_1^\ell\Big), \\ \mathcal{M}_\ell(\mathbf{H}) &= \mathrm{MLP}_\ell\Big(\mathrm{Norm}_2^\ell(\mathbf{H}) \odot (\mathbf{1} + \boldsymbol{\gamma}_2^\ell) + \boldsymbol{\beta}_2^\ell\Big). \end{aligned} \tag{6}$$

Here, $\odot$ denotes element-wise multiplication, with channel-wise modulation vectors broadcast across tokens.

The SA_ℓ operation uses four learnable projection matrices, $\mathbf{W}_{\mathrm{Q}}^\ell, \mathbf{W}_{\mathrm{K}}^\ell, \mathbf{W}_{\mathrm{V}}^\ell, \mathbf{W}_{\mathrm{O}}^\ell \in \mathbb{R}^{D\times D}$, corresponding to the query, key, value, and output projections, respectively. The MLP_ℓ operation uses two learnable weight matrices, $\mathbf{W}_1^\ell \in \mathbb{R}^{D_{\mathrm{ff}}\times D}$ and $\mathbf{W}_2^\ell \in \mathbb{R}^{D\times D_{\mathrm{ff}}}$, where D_{ff} is the MLP hidden dimension. Detailed definitions of the SA and MLP operations are provided in Section A.2.

Simply, we combine the residual updates within block ℓ into a function f_ℓ. The block output can then be written as

$$\mathbf{H}_\ell = \mathbf{H}_{\ell-1} + f_\ell(\mathbf{H}_{\ell-1}), \tag{7}$$

Encoder Output. The final encoder output, $\mathbf{Z}_t = \mathbf{H}_L$, is supplied to the decoder. Expanding the residual updates gives

$$\mathbf{Z}_t = \mathbf{H}_0 + f_1(\mathbf{H}_0) + f_2(\mathbf{H}_0 + f_1(\mathbf{H}_0)) + \cdots + f_L\left(\mathbf{H}_0 + \sum_{j=1}^{L-1} f_j(\mathbf{H}_{j-1})\right). \tag{8}$$

This expansion shows that shallow features persist as the encoder blocks go deeper.

4 METHOD

In this section, we introduce our approach to improving the representations learned by the DDT encoder, including a residual-free encoder architecture and initialization strategies.

Residual-Free Encoder We remove the residual connections around the attention and MLP sublayers within each encoder block while retaining the timestep- and class-dependent modulation. Using $\mathcal{A}_\ell$ and $\mathcal{M}_\ell$ defined in Eqs. (4) and (5), each block computes

$$\tilde{\mathbf{H}}_\ell = (\mathbf{1} + \mathbf{g}_1^\ell) \odot \mathcal{A}_\ell(\mathbf{H}_{\ell-1}), \tag{9}$$

$$\mathbf{H}_\ell = (\mathbf{1} + \mathbf{g}_2^\ell) \odot \mathcal{M}_\ell(\tilde{\mathbf{H}}_\ell). \tag{10}$$

We parameterize the output gates as $\mathbf{1} + \mathbf{g}_1^\ell$ and $\mathbf{1} + \mathbf{g}_2^\ell$. We initialize the gate parameters to zero so that both output gates equal $\mathbf{1}$ at initialization.

Let f_ℓ^{RF} denote the gated transformation of residual-free encoder block ℓ. In contrast to Eq. (7), each residual-free encoder block is computed as $\mathbf{H}_\ell = f_\ell^{\mathrm{RF}}(\mathbf{H}_{\ell-1})$, without additive identity pathways. We expect these successive transformations to encourage progressive semantic abstraction.

Encoder Output. To provide the decoder with both detailed and abstract information from multiple encoder depths, we aggregate the input patch embeddings and the outputs of all encoder blocks:

$$\mathbf{Z}_t = \sum_{\ell=0}^{L} \mathbf{H}_\ell, \tag{11}$$

where $\mathbf{H}_0$ denotes the input patch embeddings and $\mathbf{H}_\ell$ denotes the output of encoder block ℓ. In contrast to the residual accumulation in Eq. (8), this aggregation is performed only at the encoder output, and does not introduce identity pathways within the encoder. It gives the decoder direct access to features from different depths, without requiring deeper encoder blocks to retain all the information needed for reconstruction.

Attention initialization. Self-attention is disproportionately ill-conditioned compared with other components of a transformer block (Ji et al., 2025), and residual connections are commonly used to stabilize its training. To improve its conditioning and convergence in the absence of residual connections, one approach is mimetic initialization (Trockman & Kolter, 2023; Ji et al., 2026a), which initializes the query and key projections to satisfy

$$(\mathbf{W}_{\mathrm{Q}}^{\ell})(\mathbf{W}_{\mathrm{K}}^{\ell})^{\top} = \alpha\mathbf{Z} + \beta\mathbf{I}, \tag{12}$$

where $\mathbf{Z}_{ij} \sim \mathcal{N}(0, 1/D)$, $\mathbf{I}$ is the identity matrix, and α and β are scalar constants. This initialization encourages diagonal dominance in the query–key product, which can improve the conditioning of self-attention and support stable and fast convergence.

However, applying this approach to new models involves tuning two hyperparameters, α and β. We therefore propose a new strategy that mixes a shared orthogonal component with the default query and key initializations. For each encoder block ℓ, we sample a random orthogonal matrix $\mathbf{O}_\ell \in \mathbb{R}^{D\times D}$ and set

$$\begin{aligned} \mathbf{W}_{\mathrm{Q}}^{\ell} &\leftarrow \sqrt{1-\rho}\,\mathbf{O}_\ell + \sqrt{\rho}\,\mathbf{W}_{\mathrm{Q,default}}^{\ell}, \\ \mathbf{W}_{\mathrm{K}}^{\ell} &\leftarrow \sqrt{1-\rho}\,\mathbf{O}_\ell + \sqrt{\rho}\,\mathbf{W}_{\mathrm{K,default}}^{\ell}, \end{aligned} \tag{13}$$

The entries of $\mathbf{W}_{\mathrm{Q,default}}^{\ell}$ and $\mathbf{W}_{\mathrm{K,default}}^{\ell}$ are sampled independently from the uniform distribution $\mathcal{U}(-1/\sqrt{D},\, 1/\sqrt{D})$, where D is the input feature dimension. The parameter $\rho \in [0, 1]$ controls the contribution of these random components relative to the shared orthogonal component. The orthogonal component is shared between the query and key projections, whereas the default components are drawn independently. At $\rho = 0$, both projections start from the same orthogonal matrix with condition number $\kappa = 1$. At $\rho = 1$, they retain the default initialization. The projections are tied only at initialization and trained as separate parameters. The value and output projections $\mathbf{W}_{\mathrm{V}}^{\ell}$, $\mathbf{W}_{\mathrm{O}}^{\ell}$ are initialized as orthogonal matrices and thus have condition number $\kappa = 1$.

MLP initialization. Using an approach similar to Martens et al. (2021) and He et al. (2023), we initialize weight matrices in the MLP blocks according to the scale-corrected uniform orthogonal (SUO) distribution, which ensures that these layers preserve the vector RMS of their inputs, whilst also having a condition number of 1. For the depths we trained at, we found that it wasn't necessary to transform the activation functions as recommended in Martens et al. (2021).

5 Experiments

We evaluate our methods in training diffusion transformers from scratch on the ImageNet-1k dataset at a resolution of 256×256 and patch size of 2 for 80 epochs, using a total training batch size of 1024. Following Zhou et al. (2026), we use the Muon optimizer with a constant learning rate of 0.0002 and a momentum of 0.95 throughout training. We set the gradient clipping threshold to 1, weight decay to 0, and the exponential moving average (EMA) decay rate to 0.9999. We use the off-the-shelf VAE-ft-EMA from Hugging Face[2], with a downsampling factor of 8. For sampling, we use the Euler solver with 250 steps. We train both DDT and DDT-RFE in both Large (L) and Extra Large (XL) configurations, with 458M and 675M parameters, respectively. For DDT-RFE, we set $\rho = 0.1$ for the L configuration and $\rho = 0$ for the XL configuration. The original DDT

[2]https://huggingface.co/stabilityai/sd-vae-ft-ema

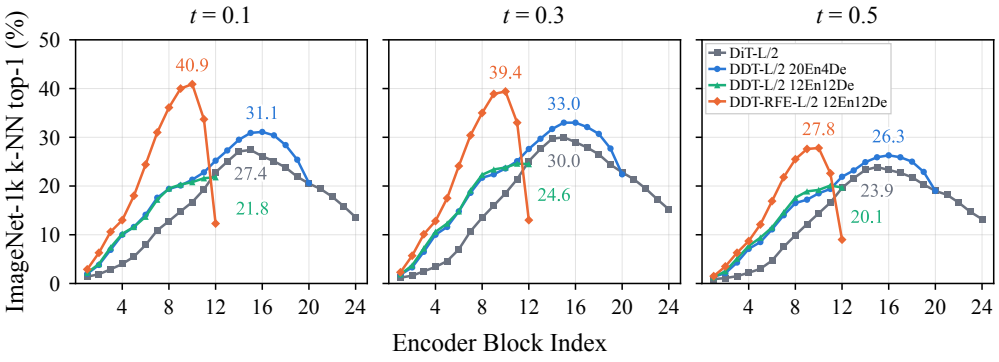


Figure 2: kNN top-1 accuracy on ImageNet-1k for the representation of every (encoder) block, at noise levels $t \in \{0.1, 0.3, 0.5\}$.

training setup (Wang et al., 2026) incorporates REPA (Yu et al., 2025) to align intermediate encoder features with representations from a pretrained visual encoder. We omit this auxiliary alignment objective in our experiments to examine the encoder representations learned through the generative objective. We evaluate both representation learning tasks using the intermediate features learned by the diffusion transformer models and the generation performance of our trained models.

5.1 Representation Learning

To assess how our architectural design affects the representations learned by diffusion models, we evaluate representations on four visual tasks: image classification, object discovery, semantic correspondence and semantic segmentation. For all evaluations, we extract features using conditioning on the null class.

5.1.1 Image Classification

Setting. We assess the representations of the frozen models by classification on ImageNet-1k. Given a noised latent $\mathbf{x}_t$, we extract token features from an encoder block and apply global average pooling to obtain an image-level representation. The same representation is used for both evaluations below. First, following the DINO protocol (Caron et al., 2021), we evaluate k-nearest-neighbor (kNN) top-1 accuracy for every encoder block at noise levels $t \in \{0.1, 0.3, 0.5\}$, using the same timestep t to construct the noisy input and to condition the network. Second, following DDAE (Xiang et al., 2023), we evaluate linear separability by training a classification head consisting of BatchNorm and a linear layer for 30 epochs. For each model we select the encoder block and noise level with the best validation accuracy and report top-1 and top-5 accuracy, and we additionally probe the encoder output that conditions the decoder with the same procedure. We compare DiT-L, DDT-L in its best configuration (20en4de), and DDT-RFE-L (12en12de). Because DDT-RFE uses fewer encoder blocks than the best DDT configuration, we also train DDT-L with the same encoder depth as ours (12en12de) for a fair comparison.

Results and analysis. Figure 2 reports kNN accuracy across blocks. As expected, DDT-L outperforms DiT-L at every noise level, reflecting the benefit of decoupling the encoder from the decoder. The DDT-L configuration with more encoder blocks also achieves higher accuracy. For example, at $t = 0.1$, DDT-L (20en4de) reaches 31.1%, compared with 21.8% for DDT-L (12en12de). Despite using fewer encoder blocks, DDT-RFE-L improves over DDT-L by 9.8% at $t = 0.1$ and 6.4% at $t = 0.3$. The linear probing results in Table 1 also show that DDT-RFE outperforms DDT in both top-1 and top-5 accuracy at the L and XL scales. At both scales, DDT-RFE achieves its best linear probing performance at a lower noise level ($t = 0.05$ versus $t = 0.15$ for DDT). Moreover, the drop in top-1 accuracy from the best encoder block to the encoder output is 7.8 and 4.7 percentage points for DDT-L and DDT-XL, respectively, compared with only 1.2 and 1.1 percentage points for DDT-RFE-L and DDT-RFE-XL, respectively. The encoder output of DDT-RFE thus remains nearly as semantically discriminative as its best intermediate representations for the classification

Table 1: Linear probing accuracy (%) on ImageNet-1k. We report the best over encoder blocks and noise levels t ($t = 0$ is the clean latent, $t = 1$ pure noise). "Encoder Output" refers to the representation supplied to the decoder: the final encoder-block representation for the DDT, the aggregated encoder representation for the DDT-RFE.

		Best Encoder Block				Encoder Output		
Model	Blocks (En/De)	Block	t	top-1	top-5	t	top-1	top-5
DDT-L/2	20 / 4	15	0.15	59.6	82.1	0.20	51.8	75.7
DDT-RFE-L/2 (Ours)	12 / 12	8	0.05	**60.9**	**83.5**	0.15	**59.7**	**82.4**
DDT-XL/2	22 / 6	17	0.15	61.8	84.0	0.15	57.1	80.0
DDT-RFE-XL/2 (Ours)	14 / 14	9	0.05	**62.4**	**84.7**	0.10	**61.3**	**83.7**

task, whereas DDT exhibits a more pronounced gap between the two. DDT-RFE achieves these improvements with substantially fewer encoder blocks (12 versus 20 at the L scale and 14 versus 22 at the XL scale), corresponding to reductions of 40.0% and 36.4% in encoder depth, respectively.

5.1.2 Object Discovery

Setting. Detecting salient objects is a fundamental problem in computer vision with applications in real-world vision systems. To address this problem, recent methods, such as LOST (Siméoni et al., 2021) and TokenCuT (Wang et al., 2023), use self-supervised pretrained models to obtain high-quality, abstract feature representations without requiring manual labels. We evaluate unsupervised object discovery using LOST, which localizes the salient object in an image using token features from a frozen model. We apply LOST to representations extracted from each transformer block and report the best results. We evaluate on VOC2012 and COCO20k and report correct localization (CorLoc), defined as the fraction of images for which the predicted bounding box has an intersection-over-union (IoU) greater than 0.5 with a ground-truth bounding box.

Table 2: Unsupervised object discovery with LOST on frozen encoders: CorLoc (%) on VOC2012 and COCO20k. We report the best over encoder blocks and noise levels.

Model	Blocks (En/De)	VOC2012	COCO20k
DDT-L/2	20 / 4	42.5	31.1
DDT-RFE-L/2 (Ours)	12 / 12	**58.4**	**41.4**
DDT-XL/2	22 / 6	41.3	29.8
DDT-RFE-XL/2 (Ours)	14 / 14	**57.8**	**39.0**

Results and analysis. Table 2 reports CorLoc results. At both model scales, DDT-RFE improves over DDT by about 16 % on VOC2012 and 10 % on COCO20k. LOST relies on a clear separation between foreground and background features. When features retain too much low-level information from shallow blocks, patch similarities may reflect local appearance more than object structure. This can make it harder to select a foreground seed and identify the full object. With the residual-free encoder, tokens from the same object have more similar features, as also shown in Fig. 1. These observations suggest that DDT-RFE learns features that better capture object semantics and distinguish objects from the background.

5.1.3 Semantic correspondence

Setting. Semantic correspondence aims to match points on corresponding parts of different instances of the same object category. These instances may differ in appearance, pose, viewpoint, and scale. This task tests whether local features capture consistent semantic information across images and support accurate localization of object parts. We evaluate zero-shot semantic correspondence on SPair-71k (Min et al., 2019), which is the most challenging semantic correspondence dataset, containing diverse variations in viewpoint and scale with 12,234 image pairs on 18 categories for

Table 3: Zero-shot semantic correspondence on SPair-71k (PCK@0.1_{bbox}, %), following the DIFT protocol at 256×256 input with the features of one encoder block. No training is involved; we report the best over encoder blocks and noise levels.

Model	Blocks (En/De)	SPair-71k
DDT-L/2	20 / 4	33.3
DDT-RFE-L/2 (Ours)	12 / 12	**37.2**
DDT-XL/2	22 / 6	35.2
DDT-RFE-XL/2 (Ours)	14 / 14	**39.9**

testing, following the protocol of DIFT (Tang et al., 2023). We report PCK@0.1_{bbox}, the percentage of keypoints whose predicted matches lie within $0.1 \times \max(w, h)$ of their ground-truth locations, where w and h are the width and height of the target object's bounding box.

Results and analysis. Table 3 shows that DDT-RFE improves PCK over DDT by 3.9 and 4.7 percentage points at the L and XL scales, respectively. These gains are achieved by directly matching frozen features, without training a task-specific prediction head. They suggest that DDT-RFE learns features that more reliably identify corresponding object parts across different instances. This complements the object discovery results that the features support both locating whole objects and matching their parts across images. DDT-RFE achieves these improvements with fewer encoder blocks, using 12 versus 20 at the L scale and 14 versus 22 at the XL scale.

5.1.4 Semantic Segmentation

Setting. We evaluate linear probing on intermediate features for the semantic segmentation task. A classifier is trained on top of the representations, with performance measured by mean intersection-over-union (mIoU). Our linear probing protocol largely follows DDAE++ (Xiang et al., 2025) and sweep over all best blocks and noise levels. We evaluate both single representation and multiscale representations from different blocks. We report results on VOC2012 (Everingham et al., 2015), ADE20K (Zhou et al., 2019) and COCO-Stuff-10K (Caesar et al., 2018).

Results and analysis. Table 4 reports mIoU on the three datasets. On VOC2012, DDT-RFE outperforms DDT by about 6 mIoU at both L and XL scales, with single-block and with multi-scale features. On ADE20K and COCO-Stuff, the single-block gains are smaller. At the XL scale, DDT-RFE matches DDT on COCO-Stuff and is 0.3 mIoU below it on ADE20K. We attribute this to the greater scene complexity of these datasets, where multi-scale information is critical. The residual connections of DDT mix representations implicitly across depth, so a single block already carries multi-level information. However, the residual-free encoder carries more abstract but less multi-scale features. Multi-scale probing is consistent with this and it benefits DDT-RFE more than DDT on both scene datasets (for example, +2.9 versus +1.8 mIoU on ADE20K at the L scale), and with multi-scale features DDT-RFE leads on every dataset at both scales. In the residual-free encoder, abstraction arises within the hierarchy and multi-scale context is recovered by aggregation. DDT-RFE achieves these results with a shallower encoder than DDT, obtaining better semantic representations at lower cost.

5.2 Image Generation

Setting. We evaluate class-conditional image generation on ImageNet at 256×256 resolution to assess whether the representation improvements are accompanied by better generation performance. For each model trained for 80 epochs, we generate 50,000 samples using EMA weights and a 250-step Euler ODE sampler, without classifier-free guidance. We report Fréchet Inception Distance (FID), spatial FID (sFID), Inception Score (IS), precision, and recall.

Results and analysis. Table 5 shows that DDT-RFE achieves the lowest FID among the compared models at both scales. DDT-RFE-L reaches an FID of 7.68, compared with 8.98 for DDT-L and 9.34 for DiT-L. The improvement remains when DDT-L uses the same encoder and decoder depths

Table 4: Dense linear probing for semantic segmentation (mIoU) on VOC2012, ADE20K and COCO-Stuff 10K. We report the best over encoder blocks and noise levels with single and multiscale representations.

		VOC2012		ADE20K		COCO-Stuff	
Model	Blocks (En/De)	single	multi	single	multi	single	multi
DDT-L/2	20 / 4	50.5	54.2	25.8	27.6	25.5	26.7
DDT-RFE-L/2 (Ours)	12 / 12	**57.0**	**60.5**	**26.8**	**29.7**	**26.0**	**28.2**
DDT-XL/2	22 / 6	53.8	56.9	**27.9**	29.8	**27.0**	28.2
DDT-RFE-XL/2 (Ours)	14 / 14	**59.8**	**63.1**	27.6	**30.0**	**27.0**	**29.0**

Table 5: Benchmarking class-conditional image generation on ImageNet 256×256 generation. We use 50k samples, 250-step Euler ODE sampler without classifier-free guidance. We report FID, IS, sFID, Precision and Recall.

Model	Blocks (En/De)	FID↓	IS↑	sFID↓	Prec.↑	Rec.↑
DiT-L/2	24 (–)	9.34	111.6	**4.63**	0.699	**0.634**
DDT-L/2	20 / 4	8.98	112.9	5.37	0.708	0.628
DDT-L/2	12 / 12	8.82	108.8	4.80	0.712	0.626
DDT-RFE-L/2 (Ours)	12 / 12	**7.68**	**118.4**	5.37	**0.731**	0.615
DDT-XL/2	22 / 6	7.27	**125.6**	5.89	0.724	**0.624**
DDT-RFE-XL/2 (Ours)	14 / 14	**6.81**	**125.6**	**5.50**	**0.740**	0.610

as DDT-RFE-L. At the XL scale, DDT-RFE outperforms DDT in terms of FID (6.81 vs. 7.27), sFID (5.50 vs. 5.89), and precision (0.740 vs. 0.724). The two models achieve the same IS, although DDT-RFE has slightly lower recall. Overall, the gains in representation quality are accompanied by improved FID at both scales under the same training budget. We demonstrate generated images by our model in Fig. 3

6 Limitations and Future Work

Due to limited computational resources, our experiments focus on DDT models trained on ImageNet at 256×256 resolution for 80 epochs. We leave experiments with longer training schedules, higher resolutions (e.g., 512×512), and other diffusion architectures to future work to assess the broader applicability of our approach.

7 Conclusion

In this work, we studied how residual connections in the condition encoder shape the representations learned by diffusion transformers, with a focus on how the encoder learns abstract features. We introduced DDT-RFE, which combines a residual-free encoder with feature aggregation at its output. This design allows successive encoder blocks to transform features without identity shortcuts, while keeping information from different encoder depths accessible to the decoder. With the proposed initialization, DDT-RFE achieves overall improvements over DDT across four visual understanding tasks using fewer encoder blocks. It also achieves a lower FID for image generation on ImageNet under the same training budget. Our ablations further show the importance of both initialization and output feature aggregation. Together, these results suggest that separating feature transformation within the encoder from aggregation at its output can benefit both representation learning and image generation. We hope that our findings encourage the community to reconsider how information flows between blocks and how this affects the learning of abstract representations in diffusion models.

REFERENCES

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1209–1218. IEEE, 2018.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015.

Chaofan Gan, Yuanpeng Tu, Xi Chen, Tieyuan Chen, Yuxi Li, Mehrtash Harandi, and Weiyao Lin. Unleashing diffusion transformers for visual correspondence by modulating massive activations. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=s3MwCBuqav.

Bobby He, James Martens, Guodong Zhang, Aleksandar Botev, Andrew Brock, Samuel L Smith, and Yee Whye Teh. Deep transformers without shortcuts: Modifying self-attention for faithful signal propagation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NPrsUQgMjKK.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Alec Helbling, Tuna Han Salih Meral, Benjamin Hoover, Pinar Yanardag, and Duen Horng Chau. Conceptattention: Diffusion transformers learn highly interpretable features. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=Rc7y9HFC34.

Yiping Ji, Hemanth Saratchandran, Peyman Moghadam, and Simon Lucey. Always skip attention. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23115–23123. IEEE, 2025.

Yiping Ji, James Martens, Jianqiao Zheng, Ziqin Zhou, Peyman Moghadam, Xinyu Zhang, Hemanth Saratchandran, and Simon Lucey. Cutting the skip: Training residual-free transformers. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=iJl3L059s6.

Yiping Ji, Mahalakshmi Sabanayagam, Peyman Moghadam, Hemanth Saratchandran, and Simon Lucey. The quantization benefits of residual-free transformers. *arXiv preprint arXiv:2605.25880*, 2026b.

James Martens, Andy Ballard, Guillaume Desjardins, Grzegorz Swirszcz, Valentin Dalibard, Jascha Sohl-Dickstein, and Samuel S Schoenholz. Rapid training of deep neural networks without skip connections or normalization layers using deep kernel shaping. *arXiv preprint arXiv:2110.01765*, 2021.

Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Spair-71k: A large-scale benchmark for semantic correspondence. *arXiv prepreint arXiv:1908.10543*, 2019.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in neural information processing systems*, 34:12116–12128, 2021.

Oriane Siméoni, Gilles Puy, Huy V Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, Renaud Marlet, and Jean Ponce. Localizing objects with self-supervised transformers and no labels. *arXiv preprint arXiv:2109.14279*, 2021.

Nick Stracke, Stefan Andreas Baumann, Kolja Bauer, Frank Fundel, and Björn Ommer. Cleandift: Diffusion features without noise. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 117–127. IEEE, 2025.

Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=ypOiXjdfnU.

Yuchuan Tian, Zhijun Tu, Hanting Chen, Jie Hu, Chao Xu, and Yunhe Wang. U-dits: Downsample tokens in u-shaped diffusion transformers. *Advances in Neural Information Processing Systems*, 37:51994–52013, 2024.

Asher Trockman and J Zico Kolter. Mimetic initialization of self-attention layers. In *International Conference on Machine Learning*, pp. 34456–34468. PMLR, 2023.

Shuai Wang, Zhi Tian, Weilin Huang, and Limin Wang. Ddt: Decoupled diffusion transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 40633–40642, 2026.

Yangtao Wang, Xi Shen, Yuan Yuan, Yuming Du, Maomao Li, Shell Xu Hu, James L Crowley, and Dominique Vaufreydaz. Tokencut: Segmenting objects in images and videos with self-supervised transformer and normalized cut. *IEEE transactions on pattern analysis and machine intelligence*, 45(12):15790–15801, 2023.

Weilai Xiang, Hongyu Yang, Di Huang, and Yunhong Wang. Denoising diffusion autoencoders are unified self-supervised learners. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15756–15766. IEEE, 2023.

Weilai Xiang, Hongyu Yang, Di Huang, and Yunhong Wang. Ddae++: enhancing diffusion models towards unified generative and discriminative learning. *arXiv preprint arXiv:2505.10999*, 2025.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=DJSZGGZYVi.

Yongsheng Yu, Wei Xiong, Weili Nie, Yichen Sheng, Shiqiu Liu, and Jiebo Luo. Pixeldit: Pixel diffusion transformers for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Xiao Zhang, Ruoxi Jiang, William Gao, Rebecca Willet, and Michael Maire. Residual connections harm generative representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 39669–39679, June 2026.

Hongkai Zheng, Weili Nie, Arash Vahdat, and Anima Anandkumar. Fast training of diffusion models with masked transformers. In *Transactions on Machine Learning Research (TMLR)*, 2024.

Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International journal of computer vision*, 127(3):302–321, 2019.

Xingyu Zhou, Qifan Li, Xiaobin Hu, Hai Chen, and Shuhang Gu. Guiding a diffusion transformer with the internal dynamics of itself. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11536–11545, 2026.

AI USE STATEMENT

In this work, we used generative AI tools to polish the language of the manuscript and improve its clarity and readability, implement and edit Python code for the proposed method, and create or modify figures and tables. We take responsibility for the final content of this work.

A APPENDIX

A.1 GENERATION IMAGE

Figure 3 shows images generated by DDT-RFE-L/2 on ImageNet at 256×256 resolution.



Figure 3: We use a 250-step Euler ODE sampler with cfg=4.

A.2 SELF-ATTENTION AND MLP OPERATIONS

We detail the operations in Eq. (6), following the Self-Attention formulation of Ji et al. (2026a). Let $\mathbf{H} \in \mathbb{R}^{N \times D}$ denote a generic input token matrix. In each encoder block, the Self-Attention and MLP operation take input $\mathbf{H}_{\ell-1}$ and $\tilde{\mathbf{H}}_\ell$, respectively, with normalization and adaptive modulation applied as defined in Eq. (6). For clarity, we omit normalization layers, rotary positional embeddings (RoPE), and bias terms, and suppress the block index ℓ.

Self-attention. Let n_h be the number of heads and $d_h = D/n_h$ the dimension of each head. The query, key, and value matrices for head h are

$$\mathbf{Q}_h = \mathbf{H}\mathbf{W}_{\mathrm{Q},h}^\top, \qquad \mathbf{K}_h = \mathbf{H}\mathbf{W}_{\mathrm{K},h}^\top, \qquad \mathbf{V}_h = \mathbf{H}\mathbf{W}_{\mathrm{V},h}^\top, \tag{14}$$

where $\mathbf{W}_{\mathrm{Q},h}, \mathbf{W}_{\mathrm{K},h}, \mathbf{W}_{\mathrm{V},h} \in \mathbb{R}^{d_h \times D}$. The attention matrix is

$$\mathbf{A}_h = \eta\big(\mathbf{Q}_h \mathbf{K}_h^\top\big), \qquad \eta(\mathbf{S}) = \mathrm{softmax}\left(\frac{\mathbf{S}}{\sqrt{d_h}}\right), \tag{15}$$

The head outputs are concatenated and projected as Self-Attention output:

$$\mathrm{SA}(\mathbf{H}) = \mathrm{Concat}(\mathbf{A}_1\mathbf{V}_1, \ldots, \mathbf{A}_{n_h}\mathbf{V}_{n_h})\,\mathbf{W}_{\mathrm{O}}^\top, \tag{16}$$

where $\mathbf{W}_{\mathrm{O}} \in \mathbb{R}^{D \times D}$.

MLP. We use a two-layer MLP with a GELU activation, applied independently to each token:

$$\mathrm{MLP}(\mathbf{H}) = \mathrm{GELU}\big(\mathbf{H}\mathbf{W}_1^\top\big)\,\mathbf{W}_2^\top, \tag{17}$$

where $\mathbf{W}_1 \in \mathbb{R}^{D_{\mathrm{ff}} \times D}$, $\mathbf{W}_2 \in \mathbb{R}^{D \times D_{\mathrm{ff}}}$, and D_{ff} is the hidden dimension that commonly is set to be $4D$. The GELU activation is applied element-wise.

A.3 Ablation on DDT

To determine the relative importance of the difference model design choices, with a particular focus on the removal of residual connections, we train the DDT baseline with our proposed initialization and linear feature fusion for 80 epochs. Table 6 shows that applying these techniques together degrades generation performance while yielding comparable classification results. Table 7 further shows that this variant performs worse on semantic segmentation and object discovery.

Table 6: Performance on generation and classification tasks.

		Generation		Classification	
		ImageNet 256×256		Best encoder block	
Model	Blocks (En/De)	FID↓	IS↑	top-1	top-5
DDT-L/2	20 / 4	8.98	**112.9**	59.6	82.1
DDT-L/2	12 / 12	**8.82**	108.8	51.0	75.2
DDT-L/2+linear+init	20 / 4	10.90	99.3	**60.4**	**82.8**

Table 7: Performance on Semantic segmentation and Object discovery.

		Semantic segmentation (mIoU)						Object discovery	
		VOC2012		ADE20K		COCO-Stuff		VOC2012	COCO20k
Model	Blocks (En/De)	single	multi	single	multi	single	multi		
DDT-L/2	20 / 4	**50.5**	**54.2**	**25.8**	**27.6**	**25.5**	**26.7**	**42.5**	**31.1**
DDT-L/2	12 / 12	39.0	42.0	20.0	21.3	20.1	21.2	42.1	29.8
DDT-L/2+linear+init	20 / 4	43.1	47.9	23.1	25.8	22.9	24.8	42.1	29.1

A.4 Ablation on DDT-RFE

We ablate the proposed initialization scheme and feature fusion in DDT-RFE-L/2, with quantitative results in Table 8 and feature visualizations in Fig. 4. Removing the proposed initialization increases FID from 7.68 to 9.86 and substantially reduces performance on the visual understanding tasks. In Fig. 4, we can see that most representations have collapsed due to network conditioning issues (Ji et al., 2025). Removing feature fusion only leads to much larger performance drops in both image generation and visual understanding. However, we empirically find that additionally removing the output gates helps recover performance. Nevertheless, this variant still underperforms the full recipe. Specifically, it increases FID by 0.66 and reduces classification accuracy, segmentation mIoU, and correspondence PCK by 0.2 percentage points, 1.1 points, and 4.4 percentage points, respectively. Qualitatively, the model learns some abstraction in the middle blocks, but it still attempts to retain detailed information because no alternative pathways are available.

Table 8: Ablation study on DDT-RFE-L/2.

	Generation		Classification	Segmentation	Correspondence
	ImageNet		ImageNet-1k	VOC2012	SPair-71k
	FID↓	IS↑	top-1	mIoU	PCK@0.1
DDT-RFE-L/2	**7.68**	**118.4**	**60.9**	**57.02**	**37.24**
w/o init	9.86	105.7	38.6	30.62	13.89
w/o fusion	16.81	74.2	6.7	3.43	4.20
w/o fusion, w/o gate	8.34	114.8	60.7	55.91	32.87

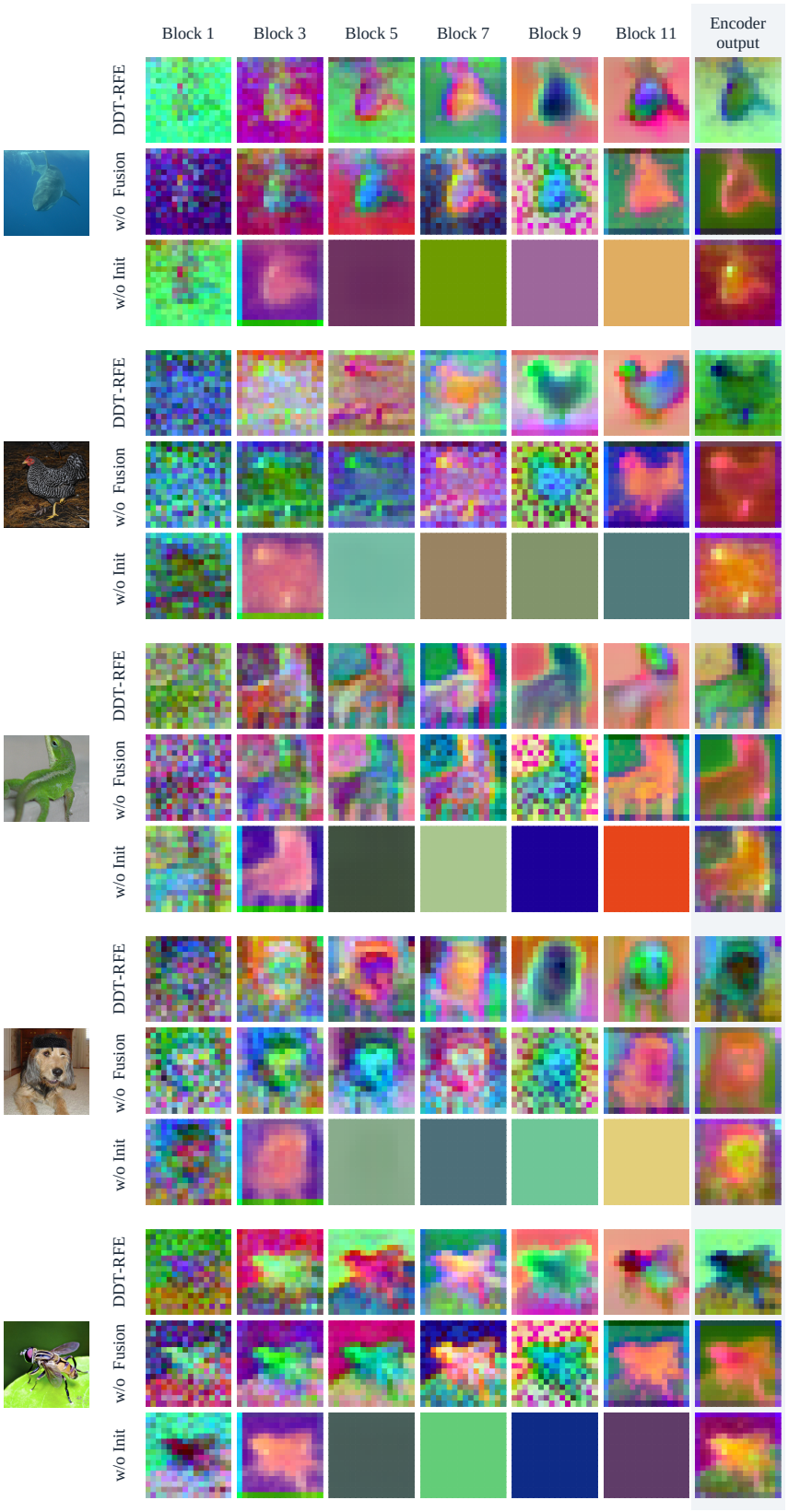


Figure 4: PCA visualization of DDT-RFE and its variants at noise level $t = 0.5$.

A.5 More Visualization

In this subsection, we demonstrate additional visualizations of encoder representations for three different models: DDT-L/2 (20en4de), DDT-L/2 (12en12de), and DDT-RFE-L/2 (12en12de) in Figs. 5 to 7

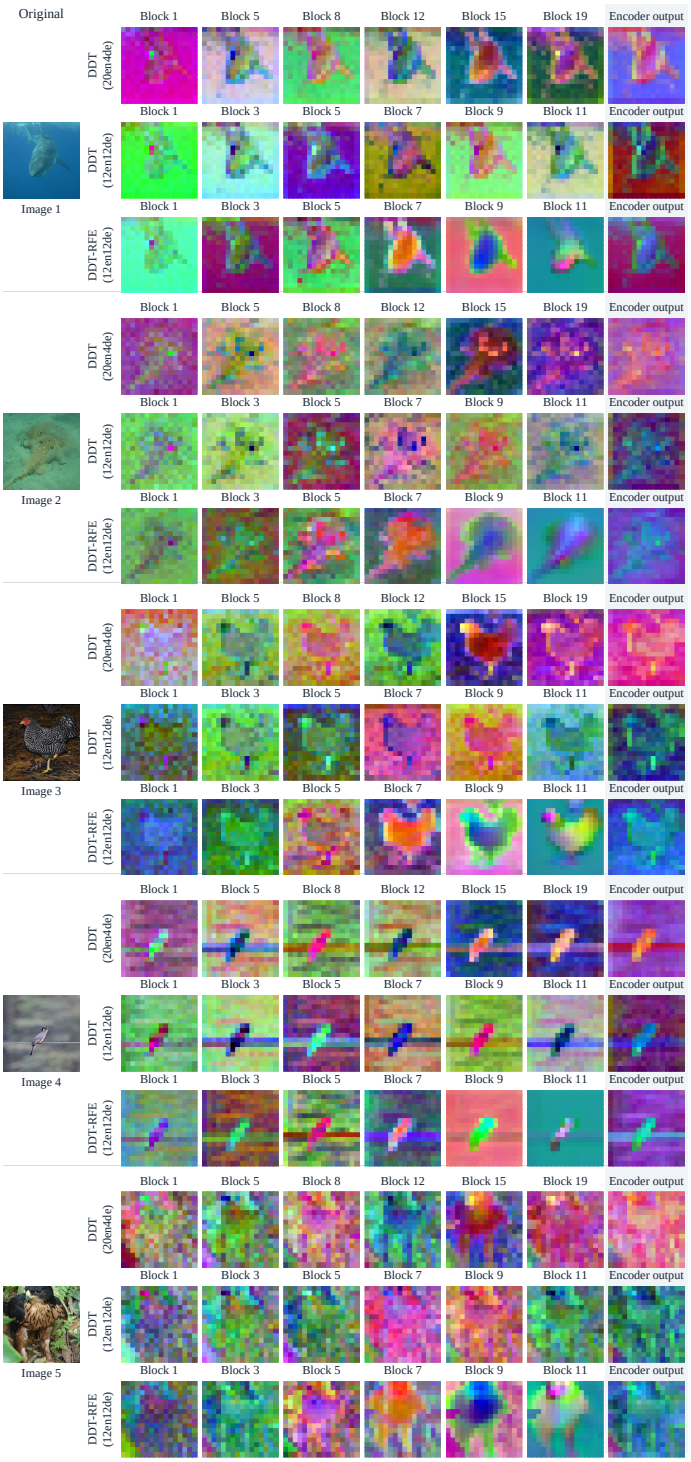


Figure 5: PCA visualization of DDT and DDT-RFE at noise level $t = 0.1$

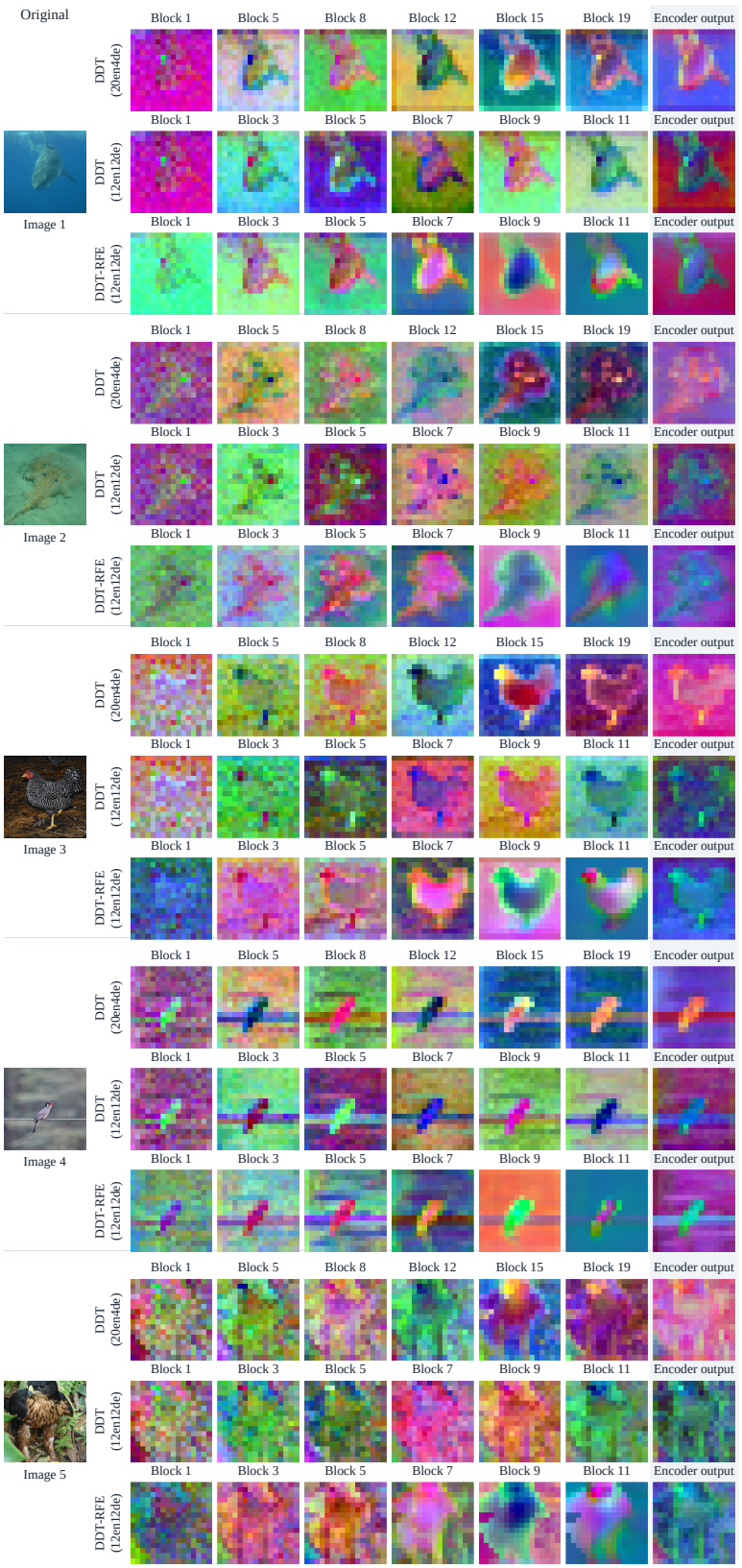


Figure 6: PCA visualization of DDT-RFE at noise level $t = 0.3$

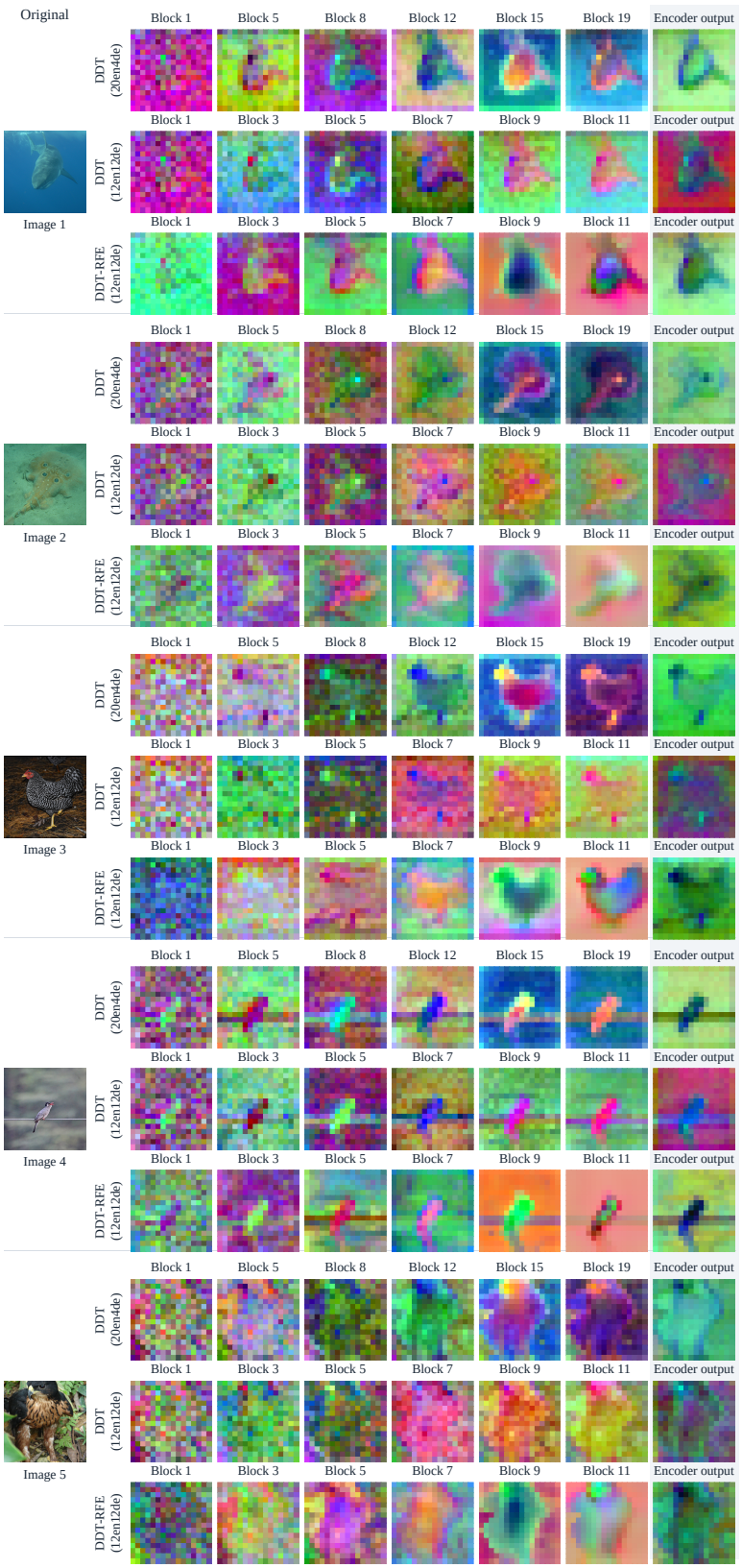


Figure 7: PCA visualization of DDT-RFE at noise level $t = 0.5$