From SRA to Self-Flow: Data Augmentation or Self-Supervision?

Dengyang Jiang[1] Mengmeng Wang[2✉] Harry Yang[1] Jingdong Wang[3†]
[1]The Hong Kong University of Science and Technology [2]Zhejiang University of Technology [3]Baidu Inc.

Abstract

Representation alignment has become an effective way to accelerate diffusion transformer training and improve generation quality. Recent self-alignment methods, such as SRA and Self-Flow, further remove the dependency on external pretrained encoders by constructing alignment within the diffusion model itself. However, the mechanism behind the improvement from SRA to Self-Flow, dual-time scheduling, remains under-examined: Self-Flow attributes its gain to interactions between tokens at different noise levels, where cleaner tokens help infer noisier ones. In this work, we revisit this explanation and ask whether the gain instead comes from data augmentation along the noise dimension. To disentangle these factors, we introduce Attention Separation, which preserves the same dual-timestep input as Self-Flow while blocking attention between tokens assigned to different noise levels. Surprisingly, removing such interaction does not degrade performance and can even improve it, suggesting that the improvement from SRA to Self-Flow mainly comes from data augmentation. Furthermore,We show that Attention Separation itself provides an augmentation effect by splitting a single image into multiple effective training parts to expand the training data. Based on these observations, we combine self-representation alignment with dual-timestep and attention-separation augmentation, and demonstrate the effectiveness of this design on ImageNet. Code: https://github.com/vvvvvjdy/SRA/tree/main/SiT-SRA_DTS_AS

1. Introduction

Enhancing the latent representation capability of Diffusion Transformers (DiTs) [11, 26, 29, 33] during training has been demonstrated to accelerate convergence and improve generation quality [5, 19, 22, 35, 38, 39]. Prior works, such as REPA [39], attempt to achieve this by aligning the internal features of DiTs with those of a frozen, pre-trained image encoder (e.g., DINOv2 [28]). However, this external alignment strategy often falls short in scenarios where a sufficiently powerful encoder is absent, or when scaling up the training data and model size for DiTs [5, 42]. To address this, recent research has pivoted toward representation alignment within the DiT itself [5, 13, 19]. Pioneer work like Self-Representation Alignment (SRA) [19], which aligns latent representations in earlier layers under higher noise conditions with those in deeper layers under lower noise levels of the same model to progressively reinforcing internal representation learning. Subsequently, Self-Flow [5] extends this self-representation alignment paradigm to multi-modal scenarios and larger scales (e.g, Text-to-Image, Text-to-Video, Text-to-Audio), demonstrating that it consistently outperforms the external alignment methods like REPA, and the self-representation alignment baseline SRA.

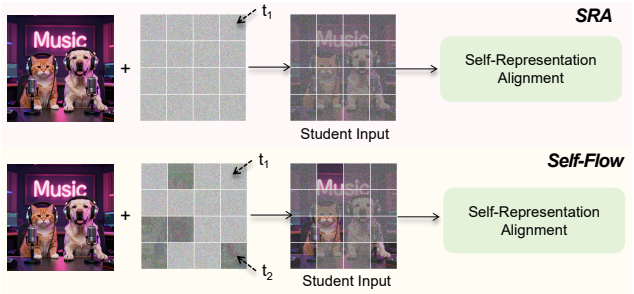

Figure 1. **Difference Between Self-Flow and SRA.** Self-Flow adopts SRA's Self-Representation Alignment method while differs in the student input processing: SRA utilizes a single noise level (t_1) for all student input tokens, whereas Self-Flow employs a dual-timestep scheduler where tokens at two distinct noise levels (t_1 and t_2) coexist in the same image.

Notably, as illustrated in Figure 1, Self-Flow also adopts the Self-Representation Alignment method pioneered by SRA. The key distinction, however, lies in how the input samples are processed for the student. Specifically, Self-Flow introduces a dual-timestep scheduling, where a single input sample to the student model contains patches corrupted by two distinct noise-levels. Consequently, the performance gains achieved by Self-Flow over SRA are primarily attributed to this specific design. In Self-Flow paper, the explanation of the mechanism of this dual-timestep

✉Corresponding author. †Project leader.

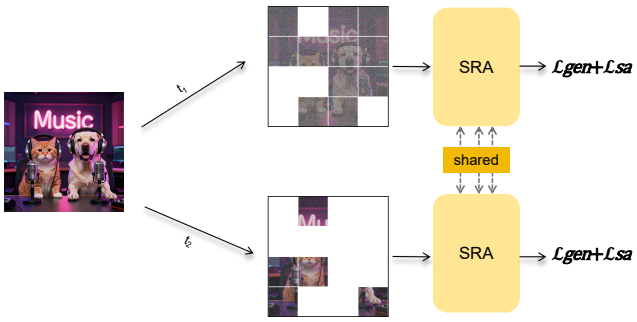

Figure 2. **Attention Separation disentangles self-supervision from augmentation.** Given a sample with dual-timestep scheduling, Attention Separation preserves the heterogeneous-noise input but partitions attention into independent timestep groups, so tokens at t_1 cannot interact with tokens at t_2. This removes token interactions while keeping the noise-state augmentation introduced by dual-timestep scheduling. Meanwhile, Attention Separation can also be interpreted as creating multiple part-conditioned views of one image to expand the training distribution, thereby also acting as a data augmentation.

scheduling is: *"by applying different noise levels to different tokens, the model is encouraged to use cleaner tokens to infer noisy tokens. This drives learning strong representations alongside generative capabilities."* Nevertheless, we question that dose these improvements indeed stem from superior self-supervision achieved by interactions of different noise-level tokens?

In this work, we revisit the mechanism behind the gains of dual-timestep scheduling. Rather than attributing the improvement solely to better self-supervision by interactions, *we argue that this design also functions as a form of data augmentation for diffusion training*. Here, data augmentation does not directly alter the semantic content of the clean image[40, 41]; instead, it expands the effective training distribution along the noise dimension. By assigning different noise level to different token subsets, a single clean sample is presented to the model under more diverse noise states, allowing the model to observe more noise-conditioned variants of the same data within training, thus expands the effective training data for the model.

To verify this hypothesis, we introduce Attention Separation, as illustrated in Figure 2. The key idea is to preserve the same dual-timestep input as Self-Flow while removing the interaction between tokens at different noise levels. Specifically, tokens assigned to the same timestep can attend to each other, whereas tokens assigned to different timesteps are blocked from interacting. This creates a controlled setting: if the improvement of Self-Flow mainly comes from cleaner tokens guiding noisier tokens through attention, removing such interaction should degrade performance; if the gain remains, the dual-timestep scheduler is more likely acting as noise-state augmentation.

This observation further leads us to reinterpret Attention Separation itself as a form of data augmentation. When applied under single-timestep training, all tokens share the same noise level. Nevertheless, Attention Separation still improves training, we analyze that with such separation, each token group acts as a partial observation of the original image. These partial views are processed by the same shared-parameter model and optimized with the same denoising and self-alignment objectives in a single iteration as shown in Figure 2. Thus, each image yields multiple effective training samples with different content subsets to expand the effective training distribution.

We evaluate this interpretation through controlled ablations and system-level comparisons, our final training scheme improves over previous self-alignment baselines on most metrics and remains on par with, or better than, the external-encoder baseline on ImageNet.

In summary, our main contributions are as follows:

- We revisit the mechanism behind the improvement from SRA to Self-Flow and show that dual-timestep scheduling is better explained as data augmentation rather than self-supervision.
- We introduce Attention Separation, a controlled operation that blocks interactions between tokens at different noise levels, and further show that it can also serve as a data augmentation.
- We combine dual-timestep scheduling and Attention Separation within self-representation alignment, achieving stronger results than previous self-alignment baselines on most metrics and competitive performance with external-encoder alignment.

2. Related Work

2.1. Representation Alignment for Generation

Improved latent representations of diffusion models can accelerate convergence and enhance generation [5, 19, 22, 35, 36, 38, 39]. One prominent avenue is leveraging discriminative priors from pretrained vision encoders for alignment [22, 32, 38, 39]. REPA [39] pioneered this paradigm by aligning intermediate diffusion features with representations from external visual encoders such as DINOv2 [28]. Building upon this paradigm of external representation alignment, subsequent studies have introduced refinements in alignment mechanisms [32, 38], training strategies [22, 37], etc.. Meanwhile, an alternative research avenue has emerged that dispenses with external encoders entirely, opting instead to perform representation alignment internally within the Diffusion Transformer [5, 13, 19, 35]. SRA [19] pioneered this paradigm by aligning latent representations in earlier layers of with those in deeper layers of the same model. Following work Self-Flow [5] extend this idea to larger-scale settings and more modalities (e.g., text-

to-image, text-to-video, and text-to-audio). And shows that self alignment paradigm consistently outperforms the external alignment counterpart. In this work, we revisit this internal alignment paradigm and study whether the improvement of dual-timestep scheduling comes from stronger self-supervision by interactions or from the augmented noise states introduced during training.

2.2. Visual Self-Supervised Learning

Self-Supervised Learning (SSL) leverages pretext tasks to learn robust representations without manual labels [1, 3, 4, 7, 12, 14, 15, 44]. For example, Masked Autoencoders (MAE) [15] reconstruct masked image patches to provide strong downstream initializations. MoCo [14] introduces a momentum encoder and a queue-based dictionary to learn instance-discriminative representations from augmented image views. DINO [4] instead adopts a self-distillation framework, where a student network learns to match a momentum teacher across different views without using negative samples. Follow-up works such as DINOv2 [28] further extend this idea to large-scale settings, producing strong general-purpose visual representations. In visual generation, Self-Supervised Learning is also applied in the training process like pre-training [5, 19, 43, 45] and post-training [20]. In this work, we further examine this self-supervised interpretation in diffusion training, showing that the benefit of dual-timestep scheduling does not primarily rely on self-supervision, but can be explained by its data augmentation effect.

2.3. Data Augmentation

Data augmentation enlarges the effective training data by constructing task-preserving variants of existing samples. In visual representation learning, classical strategies include Mixup [41], which interpolates images and labels, Manifold Mixup [34], which performs interpolation in hidden spaces, and CutMix [40], which replaces image regions across samples. In self-supervised learning, augmentations also define different views for representation learning, as in SimCLR [6]. For diffusion-based generation, data augmentation has recently been explored to improve both discriminative and generative training. Diffusion-generated synthetic images can improve ImageNet classification [2, 18], while Degeorge et al. [8] show that competitive text-to-image diffusion models can be trained from ImageNet alone with synthetic long captions and image augmentations such as CutMix and crop-based training. In this work, we focus on understanding the mechanism behind the gains of dual-timestep scheduling [5], and show through Attention Separation that it is better viewed as data augmentation.

3. Preliminary: SRA and Self-Flow

Since both SRA and Self-Flow are built upon the Flow Matching Models [11, 26], we begin by introducing the standard training objective of Flow Matching models. Subsequently, we elaborate on the formulations of SRA and Self-Flow, followed by an analysis of the key distinctions between these two methods.

Flow matching. We consider a conditional flow-matching DiT parameterized by θ. Given a clean sample $x_0 \sim p_{\text{data}}$, condition c, and Gaussian noise $x_1 \sim \mathcal{N}(0, I)$, a noisy sample at timestep $t \in [0, 1]$ is obtained by the linear path [23, 24]:

$$x_t = (1 - t)x_0 + tx_1, \tag{1}$$

where a larger t corresponds to a higher noise level. The model predicts the velocity field along this path and is trained with the standard generation objective:

$$\mathcal{L}_{\text{gen}} = \mathbb{E}_{x_0, x_1, t, c}\left[\|v_\theta(x_t, t, c) - (x_1 - x_0)\|_2^2\right]. \tag{2}$$

SRA. SRA [19] introduces a self-alignment objective without relying on an external representation encoder. Let $h_\theta^m(\cdot)$ denote the feature map from the m-th layer of the student DiT, and let $h_{\bar{\theta}}^n(\cdot)$ denote the feature map from the n-th layer of its EMA teacher, where usually $m \leq n$. For a high-noise timestep t and a lower-noise timestep $s < t$, SRA feeds x_t into the student and x_s into the EMA teacher. The early-layer student representation is then projected by a lightweight head g_ψ and aligned to the stop-gradient teacher representation:

$$\mathcal{L}_{\text{SRA}} = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} d\left(g_\psi(h_\theta^m(x_t, t, c))_i, \text{sg}\left[h_{\bar{\theta}}^n(x_s, s, c)_i\right]\right)\right], \tag{3}$$

where N is the number of tokens, i indexes a token, and $d(\cdot, \cdot)$ is a feature distance such as cosine or ℓ_2 distance. The overall objective is:

$$\mathcal{L} = \mathcal{L}_{\text{gen}} + \lambda\mathcal{L}_{\text{SRA}}. \tag{4}$$

Thus, SRA constructs self-supervision from two asymmetries: the student observes a noisier input than the teacher, and an earlier student layer is encouraged to match a later teacher layer.

Self-Flow. Self-Flow [5] follows the same EMA-based self-alignment principle introduced in SRA, but changes how the student input is constructed. Instead of assigning a single timestep to all tokens, it samples two timesteps t

and s, defines $t_{\text{hi}} = \max(t, s)$ and $t_{\text{lo}} = \min(t, s)$, and builds a token-wise timestep vector $\boldsymbol{\tau} \in [0, 1]^N$:

$$\tau_i = \begin{cases} t_{\text{hi}}, & i \in M, \\ t_{\text{lo}}, & i \notin M, \end{cases} \tag{5}$$

where M is a randomly sampled token mask. The student input is then mixed at the token level,

$$x_{\boldsymbol{\tau},i} = (1 - \tau_i)x_{0,i} + \tau_i x_{1,i}, \tag{6}$$

while the EMA teacher receives the cleaner input $x_{t_{\text{lo}}}$. Its representation objective can be written as:

$$\mathcal{L}_{\text{SF}} = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} d\left(g_\psi(h_\theta^m(x_{\boldsymbol{\tau}}, \boldsymbol{\tau}, c))_i, \text{sg}\left[h_{\bar{\theta}}^n(x_{t_{\text{lo}}}, t_{\text{lo}}, c)_i\right]\right)\right], \tag{7}$$

with g_ψ set to the identity when no projection head is used.

This formulation highlights the essential difference between SRA and Self-Flow: SRA assigns a single noise level to the entire student input. In contrast, Self-Flow introduces dual-timestep scheduling, where tokens with different noise levels coexist within the same student input. Self-Flow attributes its improvement to interactions among tokens at different noise levels: cleaner tokens provide contextual cues for noisier tokens, thereby encouraging stronger self-supervised representation learning. However, this scheduler also changes the training data, the student observes multiple noise levels within one sample instead of a single global timestep, which exposes the model to more diverse noise-level instances and can be viewed as token-wise data augmentation for the denoising task. Therefore, we argue that the gain of Self-Flow over SRA may stem from two entangled factors: interactions for better self-supervision, and heterogeneous-noise training as data augmentation.

4. Isolate Effect by Attention Separation

To disentangle these two factors, we design an Attention Separation operation that preserves the dual-timestep noise while selectively removes token interaction from different noise-levels. Specifically, we keep the same dual-timestep scheduling as Self-Flow, which means that the model still observes heterogeneous noise levels within each training sample. The only modification is in the self-attention computation. Let $r_i = \mathbb{1}[i \in M]$ be the group indicator of token i. We construct a binary attention mask:

$$A_{ij}^{\text{sep}} = \mathbb{1}[r_i = r_j] = \mathbb{1}[\tau_i = \tau_j], \tag{8}$$

which allows attention only between tokens of the same noise level. For a self-attention layer with queries, keys, and values (Q, K, V), Attention Separation computes:

$$\text{Attn}^{\text{sep}}(Q, K, V)_i = \sum_{j=1}^{N} \frac{A_{ij}^{\text{sep}} \exp(q_i^\top k_j/\sqrt{d})}{\sum_{l=1}^{N} A_{il}^{\text{sep}} \exp(q_i^\top k_l/\sqrt{d})} v_j. \tag{9}$$

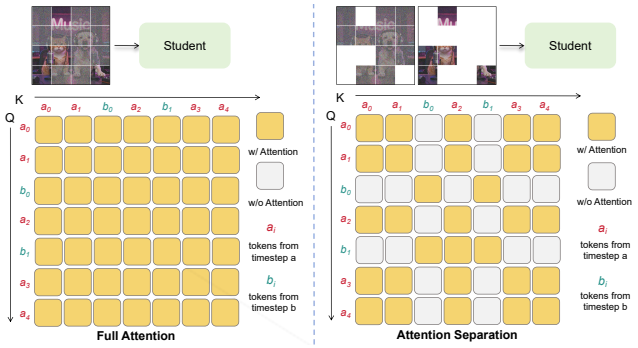

Figure 3. **Attention Separation Visualization.** Given a dual-timestep input, full attention allows all tokens to attend to each other regardless of their assigned timestep. In contrast, Attention Separation applies a block-diagonal attention mask: tokens from the same timestep group can interact, while tokens from different timestep groups are blocked.

Equivalently, tokens from different timestep groups are assigned $-\infty$ attention logits before the softmax. As shown in Figure 3, this turns the full attention matrix into a block-diagonal one: tokens assigned to the same noise level can attend to each other, whereas tokens from different noise levels are prevented from interacting through attention. This controlled setting preserves the heterogeneous-noise training but removes interactions. Therefore, comparing Self-Flow with its attention-separated counterpart allows us to isolate whether the observed gain mainly comes from self-supervision by interaction or from the dual-timestep noise as data augmentation.

5. Data Augmentation Matters More

To answer the question raised above, we conduct controlled ablations on ImageNet 256 × 256 [9] using SiT-B [26] by default. We report FID-10K [16] and IS [31] at different training iterations. Unless otherwise specified, the training and inference hyperparameters settings follow the default choices used in SRA and Self-Flow [5, 19]. Our goal is to disentangle whether the gain mainly comes from token interaction or from heterogeneous-noise data augmentation.

Removing interaction does not weaken dual-timestep training. As Attention Separation blocks attention between tokens assigned to different noise levels while preserving the same heterogeneous noise assignment. If the gain mainly came from cleaner tokens guiding noisier tokens through self-attention, this intervention should degrade performance. Table 1 shows the ablation results of whether isolates the role of token interaction of different noise-levels under dual-timestep scheduling. It can be seen that Attention Separation achieves comparable FID at 100K and improves both FID and IS at later stages, reducing FID from

25.19 to 25.06 and increasing IS from 66.75 to 72.94 at 800K. This result supports the interpretation that the dual-timestep benefit does not primarily rely on the interactions of tokens.

Table 1. **Ablation under dual-timestep scheduling**. Both rows use the same dual-timestep noise assignment; the only difference is whether tokens from different noise levels are allowed to interact through self-attention. The comparable or improved performance indicates that dual-timestep scheduling does not primarily rely on token interaction across noise levels.

Attention	Metric	100K	400K	800K
Full attention	FID ↓	58.16	30.20	25.19
	IS ↑	23.43	54.44	66.75
Attention Separation	FID ↓	58.57	29.89	25.06
	IS ↑	25.20	58.29	72.94

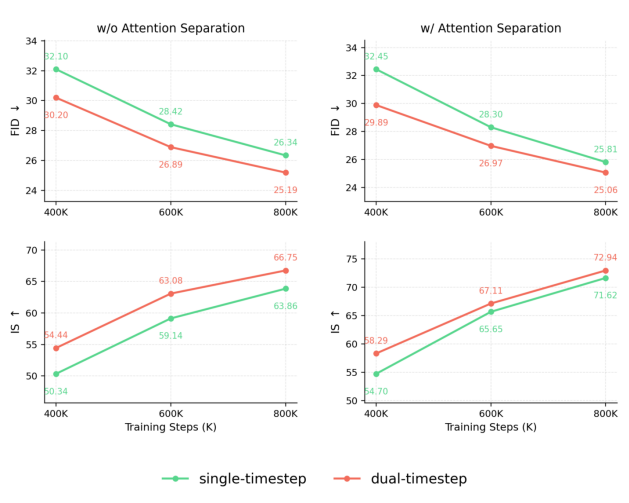

Figure 4. Comparison between single-timestep and dual-timestep training under matched attention settings. **Left**: without Attention Separation. **Right**: with Attention Separation. Dual-timestep training consistently improves over single-timestep training under both full attention and attention-separation, supporting the view that its benefit comes from noise-state augmentation.

Dual-timestep scheduling primarily works as data augmentation. Figure 4 compares single-timestep and dual-timestep training under matched attention settings. With full attention, dual-timestep training consistently improves FID from 32.10/28.42/26.34 to 30.20/26.89/25.19 at 400K/600K/800K iterations. More importantly, the same improvement remains under Attention Separation, where interactions between different noise-level tokens are explicitly blocked: dual-timestep training improves FID from 32.45/28.30/25.81 to 29.89/26.97/25.06 and also improves IS across all training stages. This indicates that the gain does not rely on stronger self-supervision induced by token interactions. Instead, dual-timestep scheduling changes the training data seen by the student: each image is decomposed into token subsets observed at different noise levels, exposing the model to more noise-state variants within the same training iteration, thus expand the effective training distribution. Therefore, its effect is better understood as data augmentation to expose the model to more data.

6. Attention Separation Is Also a Data Augmentation

Table 2 compares full attention and Attention Separation under the single-timestep setting. In this case, it do not has any interaction from tokens in different noise level, since all tokens share a timestep. It only partitions the image tokens into several non-interacting groups. However, We observe that Attention Separation still brings clear gains, especially in IS, even when all tokens share the same timestep. To investigate the source of the performance gains, we conducted the following equivalent substitution analysis.

As illustrated in Figure 2 and Figure 3, Attention Separation can be interpreted as converting one training image into multiple part-conditioned training views. Whether the two groups use different timesteps ($t_1 \neq t_2$) or the same timestep ($t_1 = t_2$), the separation mask makes each token group behave like a partial observation of the original image. These partial views are processed by the same model with shared parameters and optimized with the same denoising and self-alignment objectives in one iteration. Equivalently, a single image provides multiple effective training samples, each containing a different subset of the full image. This expands the effective training distribution without introducing external data. Together with the results in Table 2, indicating that Attention Separation itself also acts as data augmentation along the sample view, while dual-timestep scheduling augments the sample along the noise-state dimension.

Table 2. **Effect of Attention Separation under single-timestep training**. Since all tokens share the same timestep, the separation mask only partitions image tokens into non-interacting parts. The gains under single-timestep training show that Attention Separation can be beneficial even without cross-noise tokens, suggesting a augmentation effect.

Attention	Metric	100K	400K	800K
Full attention	FID ↓	62.95	32.10	26.34
	IS ↑	22.26	50.34	63.86
Attention Separation	FID ↓	62.39	32.45	25.81
	IS ↑	24.08	54.70	71.62

Effect of the mask ratio. We further study how the mask ratio affects dual-timestep training with Attention Separation. Let $\alpha = |M|/N$ denote the fraction of tokens assigned to one timestep group; thus, $\alpha = 0.25$ partitions an image into two groups with 25% and 75% tokens. Table 3 reports the ablation results.

Table 3. **Effect of mask ratio**. We compare single-timestep training, dual-timestep training with full attention, and dual-timestep training with Attention Separation in different mask ratio. All results are tested on the model trained with 800K iterations. A mild ratio preserves the augmentation benefit, while larger ratios hurts performance due to a stronger training–inference mismatch.

Time Scheduling	Attention	Mask Ratio	FID ↓	IS ↑
Single	Full	-	26.34	63.86
Dual	Full	0.25	25.19	66.75
	Separation	0.25	25.06	72.94
Dual	Full	0.35	24.87	65.46
	Separation	0.35	27.50	68.12
Dual	Full	0.50	24.39	66.62
	Separation	0.50	38.19	67.20

When full attention is used, changing the mask ratio does not harm dual-timestep training, and the performance remains consistently better than the single-timestep baseline. This is expected under our augmentation interpretation: regardless of the exact partition ratio, the model is still exposed to more noise states within each image, while full-image attention allows every token to access the complete image context. Thus, changing the mask ratio mainly changes the relative amount of tokens in one group, but does not prevent the model from learning with global spatial context. However, the behavior changes once Attention Separation is applied. While $\alpha = 0.25$ achieves the best IS and comparable FID, larger ratios degrade FID substantially, especially at $\alpha = 0.50$. We hypothesize that this degradation comes from a stronger training–inference mismatch induced by overly balanced separation. During training, Attention Separation decomposes each image into two non-interacting token groups, so each attention component can only aggregate information from a partial view of the image. As the mask ratio approaches 0.50, both groups become incomplete views with similar size, and neither group consistently preserves most of the global image context. In contrast, inference uses standard full-image attention, where all tokens interact globally. This gap between part-level training and full-image inference becomes more severe at larger mask ratios, leading to the observed FID degradation.

To mitigate the training–inference mismatch at large mask ratios, we further mix single-timestep full-image samples into each training batch. Specifically, for a fraction ρ (in our experiments, we set $\rho = 0.25$) of samples in a mini-batch, we disable the dual-timestep and Attention Separation and assign all tokens the same timestep, i.e., $\tau_i = t$ for all i, while using the standard full-attention. The remaining samples are trained with the original dual-timestep scheduling and Attention Separation. Therefore, each batch contains both separated dual-timestep samples, which preserve the heterogeneous-noise and part-level data augmentation effects, and full-image single-timestep samples, which expose the model to the same global attention pattern used at inference. This mixed setting preserves the augmentation effect from dual-timestep scheduling and Attention Separation for a subset of samples, while also exposing the model to the standard full-image, single-timestep attention pattern used at inference. As shown in Figure 5, this strategy substantially improves FID when the mask ratio is large. At $\alpha = 0.50$, where all-sample Attention Separation gives each attention component only half of the image context, adding full-image samples reduces the 800K FID from 38.19 to 24.15. The gain becomes smaller as the mask ratio decreases. This trend is consistent with our interpretation: when $\alpha = 0.25$, one token group already covers most of the image, so the separated training samples remain relatively close to the full-image inference condition. In this case, replacing part of the batch with vanilla single-timestep samples is less critical and may also weaken the augmentation effect, since those samples no longer receive either dual-timestep noise augmentation or Attention Separation.

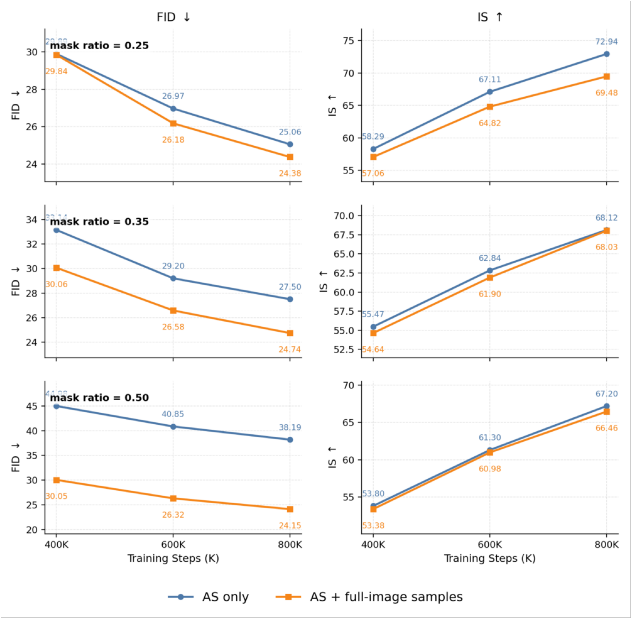

Figure 5. **Effect of adding full-image samples**. We compare Attention Separation applied to all dual-timestep samples with a mixed setting that includes full-image single-timestep samples. Adding full-image samples substantially reduces the mismatch caused by strong separation.

Table 4. **Quantitative results on ImageNet 256×256** with Classifier-free Guidance (CFG) [17]. The **best** and second-best results on each metric are highlighted in bold and underlined.

Model	Training Steps	FID↓	sFID↓	IS↑	Pre.↑	Rec.↑
vanilla diffusion transformers						
DiT-XL/2	7M	2.27	4.60	278.2	**0.83**	0.57
SiT-XL/2	7M	2.06	**4.50**	270.3	0.82	0.59
representation alignment with external encoder						
SiT-XL/2 + REPA	4M	**1.42**	4.70	305.7	0.80	**0.65**
self-representation alignment						
SiT-XL/2 + SRA	4M	1.58	4.65	311.4	0.80	0.63
SiT-XL/2 + Self-Flow	4M	1.47	4.54	305.4	0.80	0.61
SiT-XL/2 + ours	4M	1.44	4.60	**315.3**	0.81	0.63

7. Putting Things Together

The analyses above lead to a unified interpretation of the transition from SRA to Self-Flow. The key component that improves Self-Flow over SRA, dual-timestep scheduling, is not mainly explained by stronger self-supervision as suggested in the Self-Flow paper. By applying Attention Separation, we remove token interactions of different noise-level while preserving the same heterogeneous-noise input, yet the performance does not degrade and can even improve. This indicates that the benefit of dual-timestep scheduling mainly comes from data augmentation that expands the effective training data: the same image is observed under more diverse noise states. We further find that Attention Separation itself also acts as an augmentation mechanism: by splitting one training image into multiple independently optimized token groups under shared model parameters, it increases the number of effective training views derived from the same sample. In this sense, the answer to the question in our title is that **the gain from SRA to Self-Flow is better understood as *data augmentation*, rather than as stronger self-supervision.**

This interpretation naturally leads to our final training scheme. We retain the internal self-alignment objective of SRA, since it provides the representation-learning signal without relying on external encoders. On top of it, we use dual-timestep scheduling to augment each image along the noise-state dimension, and apply Attention Separation to further create part-conditioned training views. Both components are therefore used as augmentation mechanisms within the self-representation alignment framework for training.

8. System-Level Comparison

8.1. Setup

Implementation details. Unless specified otherwise, our training pipeline closely mirrors the configurations established in privious baselines [5, 19, 26, 29]. Specifically, we employ the AdamW optimizer [25] with a constant learning rate of 1e-4, zero weight decay, and a total batch size of 256, and uniform timestep sampling strategy. Latent representations are extracted utilizing the pre-trained Stable Diffusion VAE [30]. For the model backbone, we adopt the XL/2 SiT, all of which operate with a patch size of 2. For our method, we follow the setups of Self-Flow [5] and SRA [19], where the alignment layer for the student and teacher are 8 and 20, respectively. The teacher is obtained via the Exponential Moving Average (EMA) of the student with a decay of 0.9999, and the coefficient of the alignment loss is set to 0.5. The mask ratio is set to 0.25 as it yields the best performance (ablated in Table 3 and Figure 5). All experiments are conducted on 8 NIVIDA H20 GPUs.

Evaluation metrics. To evaluate generation quality, we report Fréchet Inception Distance (FID [16]), sFID [27], Inception Score (IS [31]), along with precision and recall [21]. To ensure equitable comparisons with existing baselines, we compute these metrics using the official TensorFlow evaluation suite from ADM [10] with 50K generated samples and the standard reference statistics.

Baselines for comparison. We benchmark our method against vanilla DiT and SiT [26, 29] as well as paradigms from both branches of representation alignment: namely, those with and without dependency on external models. Within each category, we benchmark against the representative method. Specifically, we select REPA [39] as the representative for external-model-assisted alignment, and SRA [19] and Self-Flow [5] for self-alignment approaches.

8.2. Results

Our method is competitive with both previous external and self-alignment methods. Table 4 reports ImageNet 256 × 256 results. Compared with the vanilla SiT-XL/2 trained for 7M steps, our method reaches a lower FID using 4M steps, improving FID from 2.06 to 1.44

Table 5. **Quantitative results on ImageNet 512×512** with Classifier-free Guidance (CFG) [17].

Model	Training Steps	FID↓	sFID↓	IS↑	Pre.↑	Rec.↑
vanilla diffusion transformers						
DiT-XL/2	3M	3.04	5.02	240.8	**0.84**	0.54
SiT-XL/2	3M	2.62	4.18	252.2	**0.84**	0.57
representation alignment with external encoder						
SiT-XL/2 + REPA	1M	**2.08**	4.19	274.6	0.83	0.58
self-representation alignment						
SiT-XL/2 + SRA	1M	2.17	4.15	279.3	0.83	**0.59**
SiT-XL/2 + Self-Flow	1M	2.12	**4.10**	280.2	0.83	0.58
SiT-XL/2 + ours	1M	**2.08**	4.12	**282.7**	0.83	0.58

Figure 6. Qualitative results on ImageNet using SiT-XL + ours. We use classifier-free guidance with $w = 4.0$.

and IS from 270.3 to 315.3. Among self-alignment methods, our method improves over SRA and Self-Flow in FID and IS, achieving the best IS and the second-best FID among all compared methods. Although REPA obtains a slightly lower FID with an external pretrained encoder, our method remains comparable while relying only on self-representation alignment inside the diffusion transformer.

The same trend holds at higher resolution. Table 5 shows the results on ImageNet 512×512. Our method matches the best FID of REPA at 2.08, outperforms both SRA and Self-Flow in FID and IS, and achieves the highest IS of 282.7. It also substantially improves over the vanilla SiT-XL/2 baseline trained for 3M steps, reducing FID from 2.62 to 2.08 with only 1M training steps. These results indicate that the augmentation interpretation developed in the controlled studies translates to stronger system-level performance, and that the resulting method remains effective when scaling to higher image resolution.

9. Conclusion

In this work, we revisit the transition from SRA to Self-Flow and study whether the improvement actually comes from. By introducing Attention Separation, we preserve the same heterogeneous-noise input while removing cross-noise token interaction. The resulting performance does not degrade and can even improve, indicating that the benefit of dual-timestep scheduling is better explained as noise-state data augmentation rather than cleaner-to-noisier token interaction alone. We further show that Attention Separation itself provides a part-level augmentation effect by splitting a single image into multiple effective training parts to expand the training data. Based on these findings, we combine dual-timestep scheduling and Attention Separation within the self-representation alignment framework. Experiments on ImageNet 256×256 and 512×512 show that this augmentation-based interpretation leads to a simple and effective training scheme, competitive with both external-encoder alignment and previous self-alignment methods.

References

[1] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15619–15629, 2023. 3

[2] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J. Fleet. Synthetic data from diffusion models improves imagenet classification. *arXiv preprint arXiv:2304.08466*, 2023. 3

[3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022. 3

[4] 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*, 2021. 3

[5] Hila Chefer, Patrick Esser, Dominik Lorenz, Dustin Podell, Vikash Raja, Vinh Tong, Antonio Torralba, and Robin Rombach. Self-supervised flow matching for scalable multimodal synthesis. *arXiv preprint arXiv:2603.06507*, 2026. 1, 2, 3, 4, 7

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607, 2020. 3

[7] Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *International Journal of Computer Vision*, 132:208 – 223, 2022. 3

[8] Lucas Degeorge, Arijit Ghosh, Nicolas Dufour, David Picard, and Vicky Kalogeiton. How far can we go with imagenet for text-to-image generation? *arXiv preprint arXiv:2502.21318*, 2025. 3

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 4

[10] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 7

[11] Patrick Esser, Sumith Kulal, A. Blattmann, Rahim Entezari, Jonas Muller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 1, 3

[12] Jean-Bastien Grill, Florian Strub, Florent Altch'e, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 3

[13] Yasaman Haghighi, Bastien van Delft, Mariam Hassan, and Alexandre Alahi. Layersync: Self-aligning intermediate layers. *arXiv preprint arXiv:2510.12581*, 2025. 1, 2

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 3

[15] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 3

[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 4, 7

[17] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 7, 8

[18] Dengyang Jiang, Haoyu Wang, Lei Zhang, Wei Wei, Guang Dai, Mengmeng Wang, Jingdong Wang, and Yanning Zhang. Low-biased general annotated dataset generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25113–25123, 2025. 3

[19] Dengyang Jiang, Mengmeng Wang, Liuzhuozheng Li, Lei Zhang, Haoyu Wang, Wei Wei, Guang Dai, Yanning Zhang, and Jingdong Wang. No other representation component is needed: Diffusion transformers can provide representation guidance by themselves. *arXiv preprint arXiv:2505.02831*, 2025. 1, 2, 3, 4, 7

[20] Dengyang Jiang, Xin Jin, Dongyang Liu, Zanyi Wang, Mingzhe Zheng, Ruoyi Du, Xiangpeng Yang, Qilong Wu, Zhen Li, Peng Gao, Harry Yang, and Steven Hoi. D-opsd: On-policy self-distillation for continuously tuning step-distilled diffusion models. *arXiv preprint arXiv:2605.05204*, 2026. 3

[21] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32, 2019. 7

[22] Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. Repa-e: Unlocking vae for end-to-end tuning of latent diffusion transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18262–18272, 2025. 1, 2

[23] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 3

[24] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023. 3

[25] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 7

[26] Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pages 23–40. Springer, 2024. 1, 3, 4, 7

[27] Charlie Nash, Jacob Menick, Sander Dieleman, and Peter W Battaglia. Generating images with sparse representations. *arXiv preprint arXiv:2103.03841*, 2021. 7

[28] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russ Howes, Po-Yao (Bernie) Huang, Shang-Wen Li, Ishan Misra, Michael G. Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1, 2, 3

[29] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 1, 7

[30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 7

[31] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 4, 7

[32] Jaskirat Singh, Xingjian Leng, Zongze Wu, Liang Zheng, Richard Zhang, Eli Shechtman, and Saining Xie. What matters for representation alignment: Global information or spatial structure? *arXiv preprint arXiv:2512.10794*, 2025. 2

[33] Z-Image Team, Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Xin Jin, Liangchen Li, Zhen Li, Zhong-Yu Li, David Liu, Dongyang Liu, Junhan Shi, Qilong Wu, Fengyi Yu, Chi Zhang, Shifeng Zhang, and Shilin Zhou. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025. 1

[34] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, Aaron Courville, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning*, pages 6438–6447, 2019. 3

[35] Mengmeng Wang, Dengyang Jiang, Liuzhuozheng Li, Yucheng Lin, Guojiang Shen, Xiangjie Kong, Yong Liu, Guang Dai, and Jingdong Wang. Sra 2: Variational autoencoder self-representation alignment for efficient diffusion training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 32978–32987, 2026. 1, 2

[36] Shuai Wang, Zhi Tian, Weilin Huang, and Limin Wang. Ddt: Decoupled diffusion transformer. *arXiv preprint arXiv:2504.05741*, 2025. 2

[37] Ziqiao Wang, Wangbo Zhao, Yuhao Zhou, Zekai Li, Zhiyuan Liang, Mingjia Shi, Xuanlei Zhao, Pengfei Zhou, Kaipeng Zhang, Zhangyang Wang, Kai Wang, and Yang You. Repa works until it doesn't: Early-stopped, holistic alignment supercharges diffusion training. *arXiv preprint arXiv:2505.16792*, 2025. 2

[38] Ge Wu, Shen Zhang, Ruijing Shi, Shanghua Gao, Zhenyuan Chen, Lei Wang, Zhaowei Chen, Hongcheng Gao, Yao Tang, Jian Yang, Ming-Ming Cheng, and Xiang Li. Representation entanglement for generation: Training diffusion transformers is much easier than you think. *arXiv preprint arXiv:2507.01467*, 2025. 1, 2

[39] 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 *International Conference on Learning Representations*, 2025. 1, 2, 7

[40] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, 2019. 2, 3

[41] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. 2, 3

[42] Yifu Zhang, Hao Yang, Yuqi Zhang, Yifei Hu, Fengda Zhu, Chuang Lin, Xiaofeng Mei, Yi Jiang, Bingyue Peng, and Zehuan Yuan. Waver: Wave your way to lifelike video generation. *arXiv preprint arXiv:2508.15761*, 2025. 1

[43] Hongkai Zheng, Weili Nie, Arash Vahdat, and Anima Anandkumar. Fast training of diffusion models with masked transformers. *arXiv preprint arXiv:2306.09305*, 2023. 3

[44] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. In *International Conference on Learning Representations*, 2022. 3

[45] Rui Zhu, Yingwei Pan, Yehao Li, Ting Yao, Zhenglong Sun, Tao Mei, and Chang Wen Chen. Sd-dit: Unleashing the power of self-supervised discrimination in diffusion transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8435–8445, 2024. 3