ATTRACT TO ALIGN, REPEL TO DIFFERENTIATE: CONTRASTIVE REPRESENTATION ALIGNMENT FOR DIFFUSION MODELS

Anonymous authors
Paper under double-blind review



Figure 1: **Qualitative comparison of representation alignment strategies.** All samples are generated using a SiT-XL/2 backbone trained for 100K steps without classifier-free guidance. Each column shares identical initial noise and sampling settings. CoRA produces visibly more coherent object structures and sharper local details than the baselines.

ABSTRACT

Representation alignment effectively accelerates diffusion transformer training by pulling hidden states toward frozen vision model features. However, solving independent point-to-point regression makes this alignment purely *attractive*. It lacks the *repulsive* force needed to differentiate spatially distinct patches, causing inefficient structural transfer and feature oversmoothing. We propose **Co**ntrastive **R**epresentation **A**lignment (**CoRA**) to mitigate representation collapse by introducing an intra-image contrastive objective. By explicitly contrasting student patches against distinct teacher patches, CoRA provides the missing repulsive gradient to maximize spatial micro-diversity. To prevent destructive gradient interference, we decouple the alignment and repulsion objectives into parallel pathways. This enables safe scaling of the negative pool size and ensures seamless compatibility with mainstream frameworks, including convolution-based approaches like iREPA. Extensive experiments demonstrate that CoRA significantly accelerates convergence and improves generative quality, adding negligible training overhead and strictly zero inference cost.

1 INTRODUCTION

Diffusion transformers Peebles & Xie (2023); Ma et al. (2024) are the dominant architecture for high-fidelity image generation, but they remain expensive to train: SiT-XL/2 still requires on the order of 7M optimization steps. A recent line of work shows that much of this cost is avoidable when the diffusion model is given access to a strong external visual representation. REPA Yu et al. (2025) aligns the hidden states of an early diffusion block with the patch features of a frozen self-supervised encoder (e.g. DINOv2 Oquab et al. (2023)) via a patch-wise cosine similarity loss, and obtains a 17×

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.

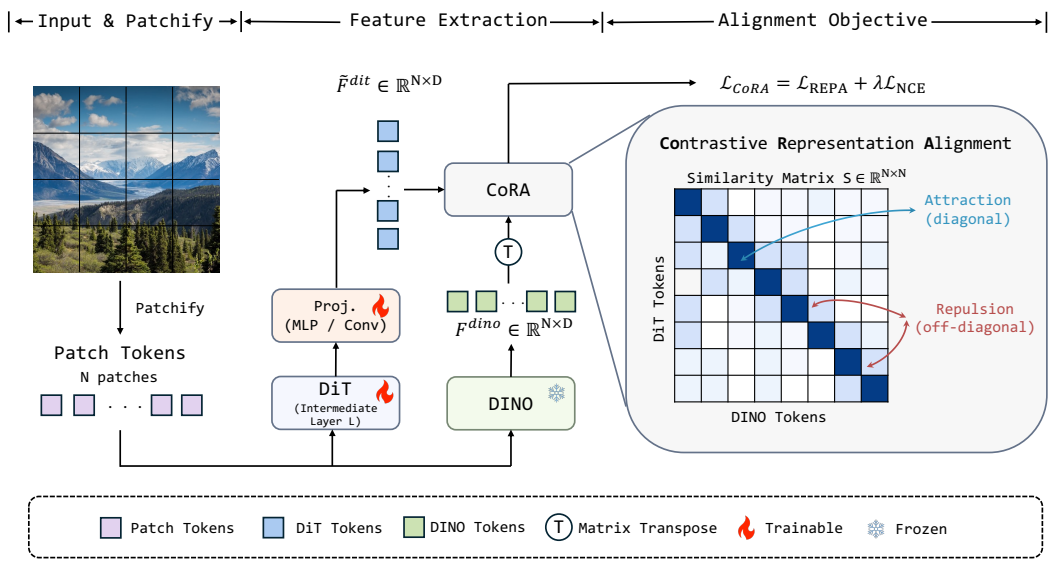


Figure 2: **Contrastive Representation Alignment (CoRA).** We align intermediate features from a trainable diffusion transformer with representations from a frozen teacher model. Beyond the standard cosine attraction between matched patches (diagonal of the token-wise similarity matrix), CoRA adds an explicit contrastive signal between distinct patches of the same image (the off-diagonal entries), encouraging student to distinguish spatially different teacher targets beyond pointwise attraction.

speedup on SiT-XL/2. Subsequent work has refined the alignment target Wu et al. (2026) and the projector that carries the alignment Singh et al. (2026), but the alignment loss itself has remained a sum of pointwise cosine similarities between matched patches.

There is, however, more than one way to align two sets of features. The current loss only *pulls* matched patches together. But the same alignment can be encouraged by also *pushing* unmatched patches apart—telling the diffusion features that the patch at one location should look different from the patch at another. Without such a repulsive force, the loss tells the student what each patch should look like, but not what it should remain distinct from. Two patches that the encoder treats as visually distinct can drift toward similar embeddings in the diffusion features, each close enough to its own target for the loss to be satisfied while the boundaries between them quietly soften. The student captures the teacher's overall *shape* but blurs the *details* that define its micro-diversity. Recent evidence suggests this is exactly the part of the alignment that matters most. Across 27 visual encoders, Singh et al. (2026) find that the encoder's intra-image patch-similarity matrix—the pattern of which patches are similar to which within an image—predicts generation quality far more strongly than its global semantic accuracy. Yet this is precisely the structure that a sum of pointwise cosine similarities transfers only indirectly.

We close this gap by adding a single regularizer to the cosine objective: a token-level contrastive loss between the diffusion features and the encoder features at the alignment layer. Each projected diffusion token is pulled toward its matched encoder token as before and, at the same time, pushed away from the encoder tokens of other patches in a negative pool. Since the encoder is frozen, the negative side is not co-adapted with the diffusion features and the loss is asymmetric by construction. When stacked on top of iREPA Singh et al. (2026), whose convolutional projector deliberately mixes information across patch tokens, we route the contrastive term through a parallel MLP head to avoid interference with the alignment objective; details are in Section 3.4.

The construction of the negative pool turns out to be non-trivial along two axes. On pool *size*, the standard contrastive-learning intuition that more negatives are uniformly helpful does not transfer: convergence improves modestly as the pool grows from a single image to a few neighbouring ones, then degrades, even as it stays above the single-image baseline. On *selection*, a simple spatial window around the query beats teacher-guided false-negative masking Huynh et al. (2022)—a textbook fix in contrastive learning—because the patches the encoder marks as too similar to be repelled are precisely the spatially adjacent, visually similar patches whose distinctness the diffusion model needs to learn.

We refer to the resulting method as **Co**ntrastive **R**epresentation **A**lignment (**CoRA**). On ImageNet-256, CoRA yields substantial improvements over the corresponding REPA and iREPA baselines across a broad range of SiT backbones and training budgets, with the largest gains appearing early in training. The same recipe carries over to pixel-space Li & He (2025) and text-to-image Esser et al. (2024) diffusion transformers without modification. The ablations further show that the gains depend on separating the attractive alignment and repulsive contrastive signals, as well as on selecting informative local negatives.

Contributions.

- We identify a limitation of standard representation alignment: its pointwise objective is purely attractive and provides no explicit signal for distinguishing spatially distinct patches, leaving intra-image patch structure to be transferred only implicitly.
- We propose **CoRA**, a token-level contrastive regularizer on top of the cosine objective. The negative pool does not behave as standard contrastive learning would suggest: a small pool with spatially focused negatives consistently outperforms larger or semantically curated pools.
- Across SiT, JiT, and MM-DiT backbones on class-conditional and text-to-image generation, CoRA improves over the corresponding representation-alignment baselines with negligible training overhead and no inference-time changes.

2 Related Work

Diffusion Models and Scalable Backbones Diffusion models, pioneered by DDPM Ho et al. (2020) and DDIM Song et al. (2020), have become the dominant generative paradigm across modalities Wu et al. (2025); Seedance et al. (2026); Liu et al. (2023). The architectural backbone has shifted from U-Nets Ronneberger et al. (2015) to Diffusion Transformers (DiTs) Peebles & Xie (2023); Ma et al. (2024) for better scalability, complemented by Flow Matching Lipman et al. (2023) for optimized transport paths. Recent works like JiT Li & He (2025) further simplify this by operating in pixel space, emphasizing direct x-prediction. Despite these advances, the high computational cost of training DiTs from scratch remains a significant bottleneck.

Representation alignment for diffusion training. A growing body of work accelerates DiT training by injecting features from a frozen vision encoder into the diffusion model. REPA (Yu et al., 2025) introduced the modern recipe: align the hidden states of an early DiT block to the patch features of a self-supervised encoder (e.g. DINOv2 (Oquab et al., 2023) or CLIP (Radford et al., 2021)) via a patch-wise cosine similarity loss. REG (Wu et al., 2026) extends the idea by entangling low-level latents with a class token so that representation alignment keeps acting at inference time. iREPA (Singh et al., 2026) replaces REPA's MLP projector with a small convolutional projector, motivated by the finding that the spatial structure of the target representation is the strongest single predictor of generation quality across 27 encoders. A separate line modifies the latent space the diffusion model operates in: VA-VAE (Yao et al., 2025) aligns the VAE tokenizer with a vision foundation model, and REPA-E (Leng et al., 2025) unlocks end-to-end joint VAE/DiT training under the alignment objective. CoRA is orthogonal to all of these: it leaves the alignment target, the projector, and the latent space unchanged, and modifies only the loss.

Contrastive and dispersive objectives in diffusion models. Following Jiang et al. (2025); Zheng et al. (2023); Zhu et al. (2024), several recent works couple alignment-style training with a contrastive or dispersive term, but operate at the level of *whole samples*. Contrastive Flow Matching (Stoica et al., 2025) pushes apart the predicted flows of arbitrary sample pairs to sharpen conditional separation. Self-Flow (Chefer et al., 2026) applies heterogeneous noise across tokens to encourage internal representation learning. Dispersive Loss (Wang & He, 2025) adds a self-contained regularizer that spreads the global hidden representations of different images apart. CoRA differs along a single, deliberate axis: the contrastive term is applied at the granularity of *patch tokens within an image*, not whole samples. Sample-level dispersion encourages two different images to occupy different regions of the hidden space; token-level dispersion, as in CoRA, encourages two different patches of the *same* image to remain distinguishable in the diffusion features—a structural property that sample-level methods do not target.

3 METHOD

3.1 PRELIMINARIES

Let $\boldsymbol{x} \in \mathbb{R}^{H\times W\times 3}$ denote a clean image and let f_ϕ be a frozen self-supervised encoder. The encoder produces a grid of T patch tokens $\boldsymbol{y} = f_\phi(\boldsymbol{x}) \in \mathbb{R}^{T\times d_y}$. Let F_θ denote a diffusion transformer, and let $\boldsymbol{h}^{(\ell)} = F_\theta^{(\ell)}(\boldsymbol{x}_t, t, c) \in \mathbb{R}^{T\times d_h}$ be the hidden states after the ℓ-th transformer block on a noisy input $\boldsymbol{x}_t$ at timestep t with class conditioning c. A projector $g_\psi : \mathbb{R}^{d_h} \to \mathbb{R}^{d_y}$ maps the diffusion features into the encoder space.

REPA aligns the projected features to the encoder targets patch-by-patch with a cosine-similarity loss

$$\mathcal{L}_{\text{ATTRACT}}(\theta,\psi) \;=\; -\,\mathbb{E}_{\boldsymbol{x},t}\left[\frac{1}{T}\sum_{i=1}^{T}\text{sim}\big(g_\psi(\boldsymbol{h}_i^{(\ell)}),\, \boldsymbol{y}_i\big)\right], \qquad \text{sim}(\boldsymbol{u},\boldsymbol{v}) = \frac{\boldsymbol{u}^\top \boldsymbol{v}}{\|\boldsymbol{u}\|\,\|\boldsymbol{v}\|}, \tag{1}$$

and the diffusion transformer is trained jointly with $\mathcal{L}_{\text{diff}} + \lambda\,\mathcal{L}_{\text{REPA}}$. REPA Yu et al. (2025) use an MLP for g_ψ while iREPA Singh et al. (2026) replace it with a convolutional projector, arguing that this preserves the encoder's spatial structure during transfer and consistently improves convergence.

3.2 WHAT THE COSINE OBJECTIVE LEAVES ON THE TABLE

An attractive-only loss is geometry-blind. Equation 1 treats the T patch alignments as T independent regression problems: it is a sum of *pointwise* attractions between $g_\psi(\boldsymbol{h}_i^{(\ell)})$ and $\boldsymbol{y}_i$. The loss inspects each pair $\big(g_\psi(\boldsymbol{h}_i^{(\ell)}),\, \boldsymbol{y}_i\big)$ in isolation, and never compares $g_\psi(\boldsymbol{h}_i^{(\ell)})$ to $\boldsymbol{y}_j$ for $j \neq i$. As a result, the pairwise structure that distinguishes patch i from patch j—which iREPA Singh et al. (2026) identify as the strongest single predictor of how useful a target encoder will be for generation—is not part of what $\mathcal{L}_{\text{REPA}}$ explicitly optimizes; the diffusion model must reconstruct it as a side-effect of matching each token to its own target.

Implication. This makes a concrete prediction about the alignment loss: an objective that only matches each diffusion feature to its target transfers spatial structure only *implicitly*, and the quality of this transfer is bounded by how distinguishable the encoder's per-patch targets already are along the direction of the matching map. A loss that explicitly contrasts non-matching patches should fill this gap. We develop this contrastive objective next and evaluate its design choices empirically in Section 4.3.

3.3 AN INTRA-IMAGE DISCRIMINATIVE OBJECTIVE

Construction. The simplest instantiation of the new term draws its negatives from a single image. For each query patch index $i \in \{1,\dots,T\}$, we treat the matched encoder token $\boldsymbol{y}_i$ as the positive and the remaining tokens $\{\boldsymbol{y}_j\}_{j\neq i}$ from the same image as negatives. Section 4.3 relaxes this choice and studies how the size and selection of the negative pool affect convergence; the design we settle on is described there.

Loss. Let $\boldsymbol{z}_i$ denote the projected diffusion feature aimed at the contrastive head (we discuss the choice of projector in Section 3.4). Write $\tilde{\boldsymbol{z}}_i$ and $\tilde{\boldsymbol{y}}_i$ for the ℓ_2-normalized versions of $\boldsymbol{z}_i$ and $\boldsymbol{y}_i$, and let $\tau > 0$ be a temperature. The intra-image InfoNCE loss is

$$\mathcal{L}_{\text{NCE}}(\theta,\psi) \;=\; -\,\mathbb{E}_{\boldsymbol{x},t}\left[\frac{1}{T}\sum_{i=1}^{T}\log\frac{\exp\big(\tilde{\boldsymbol{z}}_i^\top \tilde{\boldsymbol{y}}_i/\tau\big)}{\sum_{j=1}^{T}\exp\big(\tilde{\boldsymbol{z}}_i^\top \tilde{\boldsymbol{y}}_j/\tau\big)}\right]. \tag{2}$$

Since the encoder f_ϕ is frozen, the negative-side targets $\boldsymbol{y}_j$ are not co-adapted with $\boldsymbol{z}_i$, and the joint-encoder collapse mode familiar from contrastive self-supervised learning—in which positive and negative branches drift toward the same trivial solution—cannot occur.

3.4 ROUTING THE CONTRASTIVE TERM WHEN THE ALIGNMENT PROJECTOR IS CONVOLUTIONAL

When g_ψ is an MLP (REPA), each $g_\psi(\boldsymbol{h}_i^{(\ell)})$ depends only on its own token, and the contrastive loss in Equation 2 can be applied directly through the same projector. iREPA replaces the MLP

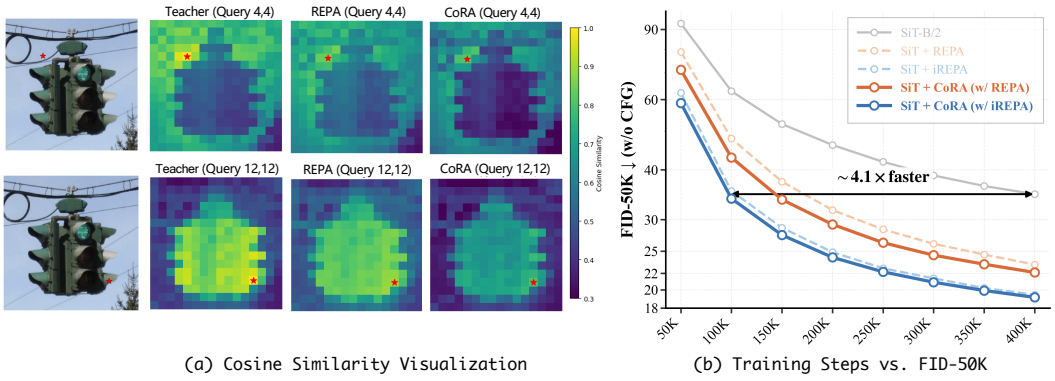


Figure 3: Comparison of cosine-similarity maps and convergence curves. **(a) Cosine-similarity map** for the query patch. Compared to REPA, CoRA provides better spatial contrast and produces attention maps that are more consistent with the teacher's representations. **(b) Convergence curve** of FID-50K across training iterations for SiT-B/2 evaluated without CFG. CoRA achieves significantly faster convergence and superior generation quality compared to both REPA and iREPA baselines.

with a small convolutional projector to better preserve the encoder's spatial structure during transfer; this projector mixes information across neighbouring tokens through its receptive field. Passing the contrastive loss through it would let the projector satisfy Equation 2 simply by emitting per-token features that look distinct after the convolution—irrespective of whether the underlying $\boldsymbol{h}_i^{(\ell)}$ are themselves distinguishable—so the contrastive supervision would never reach the diffusion features.

We therefore route the contrastive loss through a separate position-wise MLP head g_η^{nce} when stacking on iREPA:

$$g_\eta^{\text{nce}} : \mathbb{R}^{d_h} \to \mathbb{R}^{d_y}, \qquad \boldsymbol{z}_i \;=\; g_\eta^{\text{nce}}\big(\boldsymbol{h}_i^{(\ell)}\big), \tag{3}$$

while the alignment loss continues to use iREPA's CNN projector unchanged. Note that g_η^{nce} acts independently on each token, any patch-level discriminability the contrastive term rewards must originate from $\boldsymbol{h}^{(\ell)}$ itself. Both projectors are discarded at inference.

The full training objective, in either the REPA or iREPA setting, is

$$\mathcal{L}_{\text{total}} \;=\; \mathcal{L}_{\text{diff}} \;+\; \lambda_{\text{REPA}}\,\mathcal{L}_{\text{REPA}} \;+\; \lambda_{\text{NCE}}\,\mathcal{L}_{\text{NCE}}, \tag{4}$$

where $\mathcal{L}_{\text{diff}}$ and $\mathcal{L}_{\text{REPA}}$ are inherited unchanged from the underlying recipe.

4 Experiments

4.1 Setup

Architectures and target encoder. We follow the protocols of REPA Yu et al. (2025) and iREPA Singh et al. (2026). For the main experiments we use the SiT family Ma et al. (2024) at scales B/2 and XL/2, with the alignment head inherited from the underlying recipe (MLP for REPA, convolutional projector for iREPA). The contrastive head g_η^{nce} is a two-layer position-wise MLP with SiLU activation. Following iREPA, we use DINOv3-B Siméoni et al. (2025) as the target encoder. We additionally report results with DINOv2-B Oquab et al. (2023) in the appendix.

Training. Following Wu et al. (2026), we train with AdamW optimizer Loshchilov & Hutter (2017), batch size 256, and the v-prediction flow-matching loss with a linear path. Unless stated otherwise, $\lambda_{\text{REPA}} = 1.0$, $\lambda_{\text{NCE}} = 0.2$, and InfoNCE temperature $\tau = 0.07$. We retain $\lambda_{\text{REPA}} = 1.0$ to match the default iREPA implementation on which our codebase is built; Section 4.5 verifies that CoRA is robust at $\lambda_{\text{REPA}} = 0.5$. Full hyperparameters and compute details are listed in Appendix B.

Evaluation. Using Obukhov et al. (2020), we report FID Heusel et al. (2017) and Inception Score Salimans et al. (2016) computed on 50,000 generated samples, following the REPA sampling protocol (SDE Euler–Maruyama, 250 steps). For classifier-free guidance Ho & Salimans (2022), we use guidance scale $w = 1.5$ following Ma et al. (2024). We do not use guidance-interval scheduling during evaluation.

Table 1: Quantitative comparison on **SiT-B/2** and **SiT-XL/2** architectures. The best result across all alignment variants for each model scale, CFG setting, training budget, and metric is highlighted in **bold**. † Denotes the usage of the additional space-normalization method. * Denotes that the results are obtained by re-evaluating the pre-trained weights provided by Singh et al. (2026)

Model	w/o CFG				w/ CFG			
	100K Iterations		400K Iterations		100K Iterations		400K Iterations	
	FID ↓	IS ↑	FID ↓	IS ↑	FID ↓	IS ↑	FID ↓	IS ↑
SiT-B/2	62.88	20.89	34.76	42.66	37.91	37.30	11.82	110.23
MLP Based								
+ REPA	47.96	29.02	23.49	64.00	24.48	60.86	5.79	169.8
+ CoRA (Ours)	**42.82**	**34.05**	**21.58**	**68.37**	**18.13**	**77.37**	**5.34**	**174.2**
Conv Based†								
+ iREPA	35.60	43.65	19.59	**75.71**	13.23	107.9	4.30	**202.4**
+ CoRA (Ours)	**34.04**	**43.79**	**19.15**	75.55	**12.26**	**109.1**	**4.27**	200.0
SiT-XL/2	40.99	31.52	16.96	77.16	20.12	67.63	4.50	196.0
MLP Based								
+ REPA*	21.40	63.64	8.10	126.7	7.36	152.1	2.99	286.0
+ CoRA (Ours)	**16.40**	**76.60**	**7.75**	**127.3**	**5.46**	**182.7**	**2.85**	**288.0**
Conv Based†								
+ iREPA*	16.20	78.79	**7.15**	**132.0**	5.21	189.5	3.04	**293.1**
+ CoRA (Ours)	**14.48**	**84.82**	7.49	128.1	**4.67**	**200.3**	**2.93**	291.2

4.2 Main results

Acceleration over REPA. Table 1 reports FID and IS at matched compute against the baseline REPA (MLP-based) architecture. CoRA consistently improves over REPA across different model scales and training-iteration budgets. The gap is notably large at the early end of the training schedule, where the patch-discrimination signal is most informative: at 100K iterations on SiT-B/2, it improves FID from 47.96 to 42.82, a relative reduction of 10.7%. This advantage extends to larger scales, with SiT-XL/2 showing an even more significant relative FID reduction of 23.4% (from 21.40 to 16.40) at 100K iterations. By 400K iterations, CoRA continues to demonstrate superior generation quality, achieving an FID of 21.58 (vs. 23.49) on SiT-B/2 and 7.75 (vs. 8.10) on SiT-XL/2, further accelerating the convergence beyond the standard REPA paradigm.

Stacking on iREPA. Table 1 demonstrates that the contrastive term remains highly effective even when stacked on top of the strong iREPA baseline. On the SiT-B/2 architecture, CoRA improves over iREPA at both early and late training stages, lowering FID by 1.56 points (from 35.60 to 34.04) at 100K iterations, and yielding a further 0.44-point reduction (from 19.59 to 19.15) at 400K iterations. The dual-projector design is essential here: routing the contrastive loss through iREPA's shared CNN projector instead of a parallel MLP head erases the gain (see Table 6).

Headline numbers with CFG. Table 1 reports results with classifier-free guidance. On SiT-XL/2 at 400K iterations, CoRA achieves competitive FIDs of 2.85 (vs. REPA's 2.99) and 2.93 (vs. iREPA's 3.04). Notably, its acceleration benefits are significantly magnified under CFG: at just 100K iterations, the Conv-based CoRA reaches an FID of 4.67. This vastly outperforms the equivalent 100K baselines of REPA (7.36) and iREPA (5.21), and rivals the 400K baseline SiT (4.50) using only 25% of the training compute.

4.3 Designing the Negative Pool

The contrastive term in Equation 2 introduces two critical design dimensions: the *quantity* (pool size) and *quality* (selection strategy) of the negative samples. We ablate both using SiT-B/2 evaluated at 400K iterations without CFG.

Quantity: More negatives are not always better. We control the negative pool size by expanding each query's candidate pool across k images within the mini-batch (from $T-1$ up to $kT-1$ negatives, where $T=256$). As shown in Table 2, generation quality peaks at $k=4$ (1,023 negatives)

and then modestly degrades. This contrasts with the conventional contrastive-learning intuition that "more negatives are uniformly helpful" Chen et al. (2020); He et al. (2020). Because our objective is to sharpen patch-level spatial structures rather than learn global instance discrimination, the useful supervision saturates once the pool contains sufficient contextual diversity. Negatives drawn from too many unrelated images simply dilute the local structural signal.

Table 2: **Effect of negative pool size (k).** FID and IS on SiT-B/2 at 400K iterations.

Configuration (k)	# Negs.	FID ↓	IS ↑
CoRA (w/o $\mathcal{L}_{\text{NCE}}$)	-	23.49	64.00
CoRA ($k=1$)	255	23.15	64.00
CoRA ($k=2$)	511	21.53	**68.50**
CoRA ($k=4$)	**1023**	**21.34**	68.00
CoRA ($k=8$)	2047	21.67	68.05
CoRA ($k=16$)	4095	21.82	67.36
CoRA ($k=32$)	8191	22.02	67.39

Quality: Precise selection trumps raw capacity. Beyond raw scaling, we evaluate how to filter the intra-image negatives. Table 3 demonstrates that our spatial windowing strategy (restricting negatives to a local $w \times w$ grid) consistently outperforms a textbook semantic masking baseline (which masks "false negatives" based on teacher similarity Huynh et al. (2022)).

Crucially, the 5×5 spatial window achieves a highly competitive FID of 21.58 using ***only 24 negatives***. Compared to brute-force scaling—which requires over 500 or 1,000 uncurated negatives to reach similar performance (Table 2)—spatial windowing proves that *sample quality decisively outweighs sample quantity*. The semantic baseline fails because the patches the teacher marks as "too similar" are precisely the spatially adjacent, visually similar neighborhoods that CoRA most needs to push apart. Spatial windowing retains these informative hard negatives, whereas teacher-guided masking mistakenly discards them.

Table 3: **Spatial windowing vs. teacher-guided semantic masking.** FID-50K on SiT-B/2 at 400K iterations. Semantic masking removes from the contrastive sum any patch whose encoder similarity to the query exceeds $\rho = 0.7$; spatial windowing instead restricts the sum to a local window. Spatial windowing wins at every comparable pool size.

Configuration	# Negs.	FID ↓	IS ↑
CoRA (w/o $\mathcal{L}_{\text{NCE}}$)	-	23.49	64.00
Semantic ($\rho=0.7$)	-	22.25	66.71
Spatial 3×3	8	21.73	67.08
Spatial 5×5	**24**	**21.58**	**68.37**
Spatial 7×7	48	21.61	67.68

Conclusion. Together, these ablations reveal that standard contrastive learning paradigms—maximizing negative capacity and masking semantic false negatives—are suboptimal for dense generative alignment. By strictly retaining geometrically localized hard negatives, CoRA can match the representation power of massive negative pools using merely a fraction of the samples (e.g., 24 vs. 1,023). This confirms that a small, explicitly curated negative pool is the most efficient and effective way to sharpen patch-level representations without gradient interference.

4.4 Generalization to Pixel-Space and Text-to-Image Generation

Pixel-Space Diffusion. We first apply CoRA to the JiT-B/16 architecture to evaluate its efficacy in pixel-space generation. As reported in Table 4, CoRA consistently enhances generative performance across different training budgets (100 and 200 epochs) and architectural designs. Specifically, within the MLP-based framework at 100 epochs, CoRA improves over the REPA baseline by lowering the FID from 35.67 to 32.62 and increasing the Inception Score (IS) from 44.51 to 48.55. Comparable improvements are observed when stacked on the Conv-based iREPA baseline (reducing FID from 32.56 to 30.74 at 100 epochs), demonstrating that CoRA's repulsive gradient successfully mitigates feature over-smoothing even in the dense, high-dimensional pixel space.

Table 4: **Quantitative comparison on JiT-B/16.** Evaluated across 100ep and 200ep iterations. The best results are highlighted in **bold**.

Model	100ep Iterations		200ep Iterations	
	FID ↓	IS ↑	FID ↓	IS ↑
JiT-B/16 (Baseline)	41.31	37.83	32.12	50.39
MLP-Based				
+ REPA	35.67	44.51	29.20	56.60
+ CoRA (Ours)	**32.62**	**48.55**	**27.51**	**57.97**
Conv-Based†				
+ iREPA	32.56	48.29	27.75	**59.08**
+ CoRA (Ours)	**30.74**	**49.76**	**27.06**	58.12

Text-to-Image Generation. Furthermore, to investigate whether CoRA generalizes beyond class-conditional inputs, we apply CoRA to a text-to-image diffusion-transformer model (MMDiT). We train the model on the MS-COCO dataset and add CoRA as an auxiliary loss with the default alignment settings. Table 5 reports the FID and IS at matched compute (150K iterations) using classifier-free guidance ($w = 2.0$). Adding CoRA noticeably improves over the REPA baseline, reducing the FID from 4.85 to 4.48 and boosting the IS from 34.03 to 34.52. These gains suggest that the patch-level discrimination supplied by the contrastive term remains highly beneficial even when the conditioning signal is complex text rather than a simple class label.

Table 5: **CoRA on text-to-image generation.** Results are reported at 150K training iterations with classifier-free guidance ($w = 2.0$).

Method	#Iters	FID↓	IS↑
MMDiT (Baseline)	150K	6.42	30.70
MMDiT + REPA	150K	4.85	34.03
MMDiT + CoRA (ours)	150K	**4.48**	**34.52**

4.5 Ablations

All ablations use SiT-B/2 trained for 400K iterations, the standard short schedule used by REPA and iREPA for design comparisons. Numbers are FID-50K without CFG.

Table 6: **Ablation on head architectures and alignment objectives.** We investigate the interplay between projection head designs and loss formulations. Adding the repulsive objective to an MLP head substantially improves REPA, whereas applying it directly through a Conv head (iREPA) degrades performance. CoRA decouples the objectives by routing attraction through Conv and repulsion through MLP.

Method	**Head Architecture**	**Attraction**	**Repulsion**	**FID ↓**	**IS ↑**
MLP-based Alignment					
Vanilla REPA	MLP	✓	✗	23.49	64.00
REPA + NCE	MLP	✓	✓	**21.58**	**68.37**
Conv-based Alignment					
Vanilla iREPA	Conv	✓	✗	19.59	75.71
iREPA + NCE	Conv	✓	✓	**21.35**	**68.75**
Decoupled Alignment					
Decouple Head	Conv (Attract) + MLP (Attract)	✓	✗	19.64	**75.69**
CoRA(Ours)	**Conv (Attract) + MLP (Repel)**	✓	✓	**19.15**	75.55

Head architecture and the role of repulsion (Table 6). We isolate why CoRA's contrastive term needs a separate projection head when stacked on iREPA. The table reports six configurations on SiT-B/2 at 400K iterations. Three observations emerge.

(i) Repulsion helps on an MLP head but hurts on a CNN head. Adding a repulsive InfoNCE term to vanilla REPA, whose alignment head is an MLP, lowers FID from 23.49 to 21.58. Applying the same term through iREPA's CNN head moves FID in the opposite direction, from 19.59 to 21.35. Repulsion is not universally good or bad: it interacts with the inductive bias of the head it flows through.

(ii) The CNN head's spatial mixing turns repulsion into interference. A CNN projector aggregates information across neighbouring tokens before the loss is applied. When the same head is asked to simultaneously align each token to its target and push it away from its neighbours, the two gradient signals act on overlapping spatial supports, and the conflicting demands degrade the alignment the CNN head was designed to deliver. The result is the regression seen in row 4 of the table.

(iii) Adding a parameter-matched MLP head is not enough; the repulsion has to be routed *through it.* The "Decoupled Head" row uses the same Conv+MLP architecture as CoRA, but applies the attractive loss through both heads (no repulsion). Its FID of 19.64 is essentially indistinguishable from plain iREPA, ruling out the hypothesis that CoRA's gain is a side effect of the extra head's parameters. CoRA, which adds the same MLP head but reserves it for the repulsive term, reaches 19.15. The gain is therefore attributable specifically to the *Conv-attract / MLP-repel* routing, not to the head count.

REPA loss weight (λ_{REPA}). The original REPA study Yu et al. (2025) reports that its performance is robust to the alignment coefficient and saturates once $\lambda_{\text{REPA}} \geq 0.5$: on SiT-XL/2 at 400K iterations, FID changes only from 7.9 at $\lambda_{REPA} = 0.5$ to 7.8 at $\lambda_{REPA} = 1.0$. We therefore retain $\lambda_{\text{REPA}} = 1.0$, the default used by the iREPA codebase from which our implementation is developed, rather than introducing a different alignment scale for CoRA.

Table 7: **Sensitivity to loss weights under their respective alignment backbones.** All results are evaluated at 400K training iterations without CFG. The MLP setting fixes $\lambda_{\text{NCE}} = 0.2$, while the Conv setting fixes $\lambda_{\text{REPA}} = 1.0$.

Settings	Ablated Weight	FID ↓	IS ↑
MLP $\lambda_{\text{NCE}} = 0.2$	$\lambda_{\text{REPA}} = 0.5$	**21.57**	67.33
	$\lambda_{\text{REPA}} = 1.0$	21.58	**68.37**
Conv $\lambda_{\text{REPA}} = 1.0$	$\lambda_{\text{NCE}} = 0.05$	19.49	**75.75**
	$\lambda_{\text{NCE}} = 0.10$	19.30	75.65
	$\lambda_{\text{NCE}} = 0.20$	**19.15**	75.55
	$\lambda_{\text{NCE}} = 0.50$	19.45	73.94
	$\lambda_{\text{NCE}} = 1.00$	20.19	71.93

Table 7 directly verifies this choice for MLP-based CoRA with $\lambda_{\text{NCE}} = 0.2$. At 400K, lowering λ_{REPA} to 0.5 yields essentially unchanged performance: FID is 21.57 versus 21.58 at the default setting, while IS is 67.33 versus 68.37. Thus, CoRA does not depend on an unusually large alignment weight; $\lambda_{\text{REPA}} = 1.0$ is used for consistency with the underlying iREPA recipe.

NCE loss weight (λ). We control the relative strength of the repulsive contrastive objective via the loss weight λ. Table 7 demonstrates the effect of varying $\lambda \in \{0.05, 0.1, 0.2, 0.5, 1.0\}$ within our decoupled architecture at 400K training iterations. As λ increases from 0.05 to 0.2, the FID score steadily improves, reaching an optimal value of 19.15. This confirms our hypothesis that a moderate repulsive force effectively mitigates feature oversmoothing by encouraging spatial distinctiveness among patch tokens.

However, aggressively increasing the weight ($\lambda \geq 0.5$) significantly degrades generation quality, with FID deteriorating to 20.19 at $\lambda = 1.0$. We attribute this to optimization interference: an overly dominant repulsive objective scatters the representations too violently, disrupting the primary attractive alignment and preventing the student from properly anchoring to the teacher's semantic structure. Therefore, we adopt $\lambda = 0.2$ as our default setting to achieve the best balance between semantic alignment and spatial sharpness.

5 Conclusion and Limitations

We revisited the loss of the modern representation-alignment recipe for diffusion-transformer training and identified a structural asymmetry: the standard cosine objective only *pulls* matched patches together and leaves the encoder's pairwise patch structure—recently shown to be the dominant predictor of generation quality—to be transferred only implicitly. We closed this gap with a patch-level contrastive term that contrasts each query patch against the other patch tokens in its negative pool, and showed that two design knobs—the size and the selection criterion of the pool—behave non-trivially: pool size has a non-monotone sweet spot, and spatial selection within the image consistently outperforms semantic selection. The resulting method CoRA can be applied as a drop-in addition to both REPA and iREPA and is removed at inference time. Across the evaluated ImageNet settings, it delivers consistent early-training gains and improves the primary FID metric in a broad range of model-scale and target-encoder configurations. Our findings suggest that the design space of *losses* for representation-aligned generation is far from exhausted: simple changes that target the right gradient—in our case, an explicit repulsive signal over a deliberately chosen negative distribution—can deliver compute-budget gains compounding on architectural and target-encoder advances. We hope this encourages further work on the loss geometry of generation-aligned representation learning.

Limitations and Future Work. While CoRA demonstrates strong performance across the evaluated settings, several limitations remain. Our text-to-image experiments remain limited in scale, and broader evaluation on larger models and datasets is needed. We also plan to validate CoRA with a wider range of vision encoders. Finally, although CoRA can be adapted to the CNN-based iREPA framework, a more streamlined compatibility mechanism for convolutional projectors remains an important direction for future architectural refinement.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish the writing of the paper and to support retrieval and discovery of related work. We have not used generative AI tools for research ideation or execution, drafting sections of the paper, generating synthetic datasets, or proving mathematical claims; the remaining disclosure categories are not applicable to this work.

Additionally, we used generative AI tools to identify potentially outdated bibliographic metadata and potential spelling errors. The authors manually verified each suggested bibliographic update against the corresponding official publication record, including whether an arXiv entry should be replaced by its published version, and reviewed and corrected each accepted spelling change. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies representation alignment for improving the training efficiency and generation quality of diffusion models. Our experiments use standard benchmark datasets and pretrained visual encoders, and do not involve the collection of new data or human-subject studies. As with other generative image models, the resulting methods may inherit biases from their training data and could be misused to create misleading or harmful content. We encourage users and downstream deployers to follow the applicable dataset licenses and to evaluate such risks before deployment.

REPRODUCIBILITY STATEMENT

The method is specified in Section 3, including the alignment objective, negative-pool construction, and decoupled projection heads. Section 4.1 describes the training and evaluation protocols, and the main experiments and ablations report results across backbones, training budgets, and target encoders. The appendix provides the detailed hyperparameters and additional ablations, including the loss-weight sensitivity analysis in Table 7, to support independent reproduction of the reported results.

REFERENCES

Hila Chefer, Patrick Esser, Dominik Lorenz, Dustin Podell, Vikash Raja, Vinh Tong, Antonio Torralba, and Robin Rombach. Self-supervised flow matching for scalable multi-modal synthesis. *arXiv preprint arXiv:2603.06507*, 2026.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

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

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*, pp. 9729–9738, 2020.

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.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Tri Huynh, Simon Kornblith, Matthew R Walter, Michael Maire, and Maryam Khademi. Boosting contrastive self-supervised learning with false negative cancellation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2785–2795, 2022.

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.

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*, pp. 18262–18272, 2025.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. *arXiv preprint arXiv:2511.13720*, 2025.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*, 2023.

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

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

Anton Obukhov, Maximilian Seitzer, Po-Wei Wu, Semen Zhydenko, Jonathan Kyl, and Elvis Yu-Jing Lin. High-fidelity performance metrics for generative models in pytorch, 2020. URL `https://github.com/toshas/torch-fidelity`. Version: 0.4.0, DOI: 10.5281/zenodo.3786539.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

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

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

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

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.

Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, et al. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

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? In *The Fourteenth International Conference on Learning Representations*, 2026.

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

George Stoica, Vivek Ramanujan, Xiang Fan, Ali Farhadi, Ranjay Krishna, and Judy Hoffman. Contrastive flow matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1185–1194, 2025.

Runqian Wang and Kaiming He. Diffuse and disperse: Image generation with representation regularization. *arXiv preprint arXiv:2506.09027*, 2025.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Ge Wu, Shen Zhang, Ruijing Shi, Shanghua Gao, Zhenyuan Chen, Lei Wang, Zhaowei Chen, Hongcheng Gao, Yao Tang, Ming-Ming Cheng, et al. Representation entanglement for generation: Training diffusion transformers is much easier than you think. *Advances in Neural Information Processing Systems*, 38:7714–7743, 2026.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15703–15712, 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 *International Conference on Learning Representations*, 2025.

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

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*, pp. 8435–8445, 2024.

A More Extension Experiments

Global feature alignment strategy. In addition to patch-level spatial alignment, transferring the holistic semantic representation from the frozen teacher (e.g., DINOv3) to the student network is crucial for generation quality. We conduct an ablation study to investigate the optimal strategy for global feature alignment. Specifically, we explore two primary design choices:

1. **Target Representation**: We compare using the Global Average Pooling (GAP) of the teacher's patch tokens against using the teacher's explicit `[CLS]` token.
2. **Alignment Objective**: We evaluate a simple Cosine Similarity loss (attraction only) versus an InfoNCE loss (contrastive learning with in-batch negatives).

Based on these choices, we implement and compare four variants:

- **GlobalCos**: Maximizes the cosine similarity between the student's GAP feature and the teacher's GAP feature.
- **GlobalCos-CLS**: Maximizes the cosine similarity between the student's GAP feature and the teacher's `[CLS]` token.
- **GlobalNCE**: Employs a contrastive objective (InfoNCE) to align the student's GAP feature with the teacher's GAP feature, using other images in the batch as negative samples.
- **GlobalNCE-CLS**: Employs the InfoNCE objective to align the student's GAP feature with the teacher's `[CLS]` token.

Table 8: **Ablation on global feature alignment strategies.** We compare different teacher target representations and alignment objectives. "GAP" denotes the global average pooling of patch tokens, and "CLS" denotes the class token from the teacher. "Negatives" indicates whether in-batch negative samples are used.

Method Variant	Teacher Target	Objective	Negatives	FID ↓	IS ↑
Baseline-REPA (No global loss)	-	-	-	23.49	64.00
GlobalCos	GAP	Cosine	✗	23.66	63.84
GlobalNCE	GAP	InfoNCE	✓	23.08	66.96
GlobalCos-CLS	CLS	Cosine	✗	23.23	64.85
GlobalNCE-CLS	CLS	InfoNCE	✓	23.07	65.70

Ablation on Spatial Topology Matching. Beyond standard point-wise feature alignment, we explore whether explicitly regularizing the internal spatial structure of the student's representation provides additional generative benefits. To investigate this, we introduce a **Gram Matching Loss**. Specifically, we compute the token-to-token cosine similarity matrix (effectively a Gram matrix representing spatial self-correlations) for both the normalized student and teacher features, and penalize their Mean Squared Error (MSE). This objective forces the student to mimic the teacher's holistic spatial topology and attention-like correlation patterns.

We evaluate the impact of this explicit structural constraint across different projection architectures: the position-agnostic MLP head (Standard REPA) and the spatially-aware Convolutional head (iREPA). As reported in Table 9, Gram Matching mildly improves the MLP-based REPA baseline, reducing FID from 23.49 to 22.82 and increasing IS from 64.00 to 65.60. In contrast, it degrades the Conv-based iREPA baseline, increasing FID from 19.59 to 20.35 and reducing IS from 75.71 to 73.41. These results suggest that explicit global topology matching is less compatible with the spatial inductive bias of the convolutional projector.

Ablation on Pair-Sample Alignment Strategies. Beyond single-image alignment, we further explore how to leverage paired samples (e.g., augmented views or semantic pairs) to enhance representation learning. Assuming a mini-batch of size $2N$ structured as N anchors and N corresponding positive pairs, we propose and investigate two complementary objectives:

Table 9: **Ablation on Spatial Topology Matching.** We evaluate the effect of adding the Gram Matching Loss (which aligns pairwise spatial correlations) on top of the standard attractive alignment. We compare its impact across different projection head architectures (MLP vs. Conv).

Method	Head Architecture	Point-wise Align	Gram Matching	FID ↓	IS ↑
MLP-based Architecture					
Standard REPA	MLP	✓	✗	23.49	64.00
REPA + Gram	MLP	✓	✓	22.82	65.60
Conv-based Architecture					
iREPA	Conv	✓	✗	19.59	75.71
iREPA + Gram	Conv	✓	✓	20.35	73.41

- **Pair-Sample Repel (Micro-Spatial Level):** To prevent localized feature collapse and maintain fine-grained spatial discriminability, we group the patch tokens of the anchor and its pair into an extended sequence of length $2L$. A token-level contrastive loss is then applied across this combined space. By enforcing the diagonal as the sole positive targets, this objective explicitly forces the student to repel distinct spatial patches not only within the same image but also across its semantically similar pair.
- **Pair-Sample Attract (Macro-Semantic Level):** While spatial structures may misalign between paired views, their holistic semantics should remain identical. To enforce this, we apply Global Average Pooling (GAP) to extract spatial-agnostic representations. We then compute a dense batch-level similarity matrix and employ a multi-positive Supervised Contrastive (SupCon) objective. This explicitly pulls the student's anchor towards both the teacher's anchor and the teacher's paired view, while repelling all other unrelated instances in the batch.

As shown in Table 10, independently applying either strategy brings marginal improvements.

Table 10: **Ablation on Pair-Sample Alignment.** We investigate the impact of utilizing paired data (e.g., augmented views) during alignment. "Micro-Repel" denotes token-level spatial contrast across pairs, while "Macro-Attract" denotes global GAP-based supervised contrastive learning. We assess these complementary objectives independently.

Method	**Micro-Repel** (Token-level)	**Macro-Attract** (Global-level)	FID ↓	IS ↑
Baseline (Standard REPA)	✗	✗	23.49	64.00
+ Pair Repel	✓	✗	22.60	65.75
+ Pair Attract	✗	✓	22.92	65.38

Compatibility with Different Visual Encoders. To ensure our method is not overfitted to DINOv3, we apply CoRA using DINOv2 as the target representation. As shown in Table 11, CoRA successfully improves the primary generative metric (FID) over the attractive-only REPA baseline when aligned to DINOv2 features. This demonstrates that the "attract to align, repel to differentiate" principle is universal and generalizes well across different self-supervised vision paradigms.

Table 11: **Quantitative comparison using DINOv2.** Generative performance on SiT-B/2 evaluated at 400K training iterations without classifier-free guidance (w/o CFG). Using DINOv2-ViT-B as the target encoder, CoRA consistently improves the primary FID metric over the standard REPA baseline.

Method	Target Encoder	# Iters	FID ↓	IS ↑
REPA	DINOv2	400K	22.75	64.92
CoRA (Ours)	DINOv2	400K	**22.04**	**66.95**

B Training Details and Hyperparameters

Hardware and Environment. All models are trained using PyTorch with Automatic Mixed Precision (AMP, specifically `fp16`) and TF32 enabled to accelerate training and reduce memory

footprint. We also utilize PyTorch's `compile` feature for further optimization. Our experiments are conducted on a compute node equipped with 8 NVIDIA A800 (80GB) GPUs.

Optimization and Schedule. Following the standard protocols for diffusion models and flow-matching architectures, we utilize the AdamW optimizer Loshchilov & Hutter (2017) with a constant learning rate of 1×10^{-4}, $\epsilon = 10^{-8}$, and no weight decay. Gradient clipping is applied with a maximum norm of 1.0. For classifier-free guidance (CFG), we randomly drop the conditioning with a probability of $p_{\text{uncond}} = 0.1$ during training. We maintain an Exponential Moving Average (EMA) of the model weights with a decay rate of 0.9999 (updated every step) to stabilize training and improve sample quality during evaluation.

Alignment and Contrastive Hyperparameters. As introduced in the main text, our representation alignment objective integrates both REPA and InfoNCE losses. The target encoder is set to DINOv3-ViT-B16 by default. We retain the projection head of the underlying alignment recipe: an MLP for REPA and a convolutional projector for iREPA. We set the REPA loss weight to $\lambda_{\text{REPA}} = 1.0$ to match the default iREPA recipe; the sensitivity study in Table 7 shows that CoRA remains robust at 0.5. The contrastive NCE loss weight is $\lambda_{\text{NCE}} = 0.2$. We adopt a standard temperature parameter $\tau = 0.07$ for the InfoNCE loss to adequately scale the cosine similarities and control the sharpness of the contrastive distribution. A comprehensive summary of all hyperparameters is detailed in Table 12.

Table 12: Detailed training hyperparameters for our models.

Hyperparameter	**Value**
Optimization Setup	
Optimizer	AdamW
Learning Rate	1×10^{-4}
Weight Decay	0
Optimizer Betas (β_1, β_2)	$(0.9, 0.999)$
Adam Epsilon (ϵ)	1×10^{-8}
Max Gradient Norm	1.0
Batch Size	256
Max Training Steps	$400,000$
EMA Decay	0.9999
Precision	`fp16` (with TF32)
PyTorch Compile	True
Architecture & Flow-Matching	
Generation Target	v-prediction
Flow-Matching Path	Linear
Fused Attention	True
CFG Dropout Probability	0.1
Target Encoder	DINOv3-ViT-B16
Representation Alignment	
REPA weight (λ_{REPA})	1.0
NCE weight (λ_{NCE})	0.2
NCE temperature (τ)	0.07
Data & Augmentation	
Image Resolution	256×256
Random Horizontal Flip	0.5
Center Crop	True