# RELEVANCE-GUIDED TOKEN SUPERVISION ALLOCATION FOR DIFFUSION REPRESENTATION ALIGNMENT

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Recent work has shown that aligning intermediate diffusion features with pretrained visual representations can improve diffusion model training. However, existing methods typically use teacher representations to define alignment targets while allocating supervision uniformly, regardless of image content. To address this limitation, we propose *Relevance-Guided Token Supervision Allocation* (RGSA), which uses teacher-derived relevance cues to replace uniform aggregation of token-wise alignment losses with image-conditioned weighting. Specifically, RGSA derives proxy relevance cues from frozen teacher representations using patch-feature norms to characterize response strength and patch–CLS cosine similarity to capture local–global semantic consistency. These cues are fused into token relevance scores and normalized within each image to assign relative weights to the corresponding token-wise alignment losses. Experiments on ImageNet-1K show consistent improvements over iREPA across SiT-B/2, SiT-L/2, and SiT-XL/2, with similar gains across diverse visual teachers. Controlled ablations further show that RGSA outperforms random and spatially shuffled weighting in both FID and Recall. Together, these results highlight relevance-guided supervision allocation as an effective design choice for diffusion representation alignment. An anonymous code repository is available at https://anonymous.4open.science/r/RGSA_code-533C/.

## 1 INTRODUCTION

Diffusion models have achieved remarkable success in image generation (Ho et al., 2020; Song et al., 2021), but training Transformer-based diffusion models often requires substantial computational resources and long training schedules (Peebles & Xie, 2023; Ma et al., 2024). REPA (Yu et al., 2025) accelerates training by aligning intermediate diffusion features with clean-image representations from frozen visual encoders such as DINOv2 (Oquab et al., 2023). Its improvements in convergence speed and generation quality highlight the potential of representation alignment for diffusion model training.

Building on REPA, subsequent studies have extended representation alignment across architectures, training settings, and guidance strategies. U-REPA (Tian et al., 2026) adapts alignment to diffusion U-Nets, addressing architectural and spatial-resolution differences, while REPA-E (Leng et al., 2025) enables joint optimization of the VAE and diffusion model. Other work examines alignment schedules. HASTE (Wang et al., 2026c) combines early feature and attention alignment with later termination, while REED (Wang et al., 2026a) introduces flexible representation guidance and a curriculum balancing representation learning and data generation. A related direction focuses on spatial information transfer. UniRA (Chen et al., 2026) and sREPA (Xu et al., 2026) enrich representation alignment through local–global consistency constraints and explicit spatial relational matching, respectively. iREPA (Singh et al., 2026) highlights the importance of teacher spatial structure beyond global semantic performance and enhances its transfer through convolutional projection and spatial normalization, improving convergence across visual encoders and model scales. These advances provide a foundation for studying how diffusion models use teacher-derived spatial knowledge.

Despite these advances in transferring spatial information, the spatial allocation of alignment supervision remains a separate design question. Teacher representations extracted from clean images specify what the student's intermediate features should match (Yu et al., 2025), but do not by them-



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

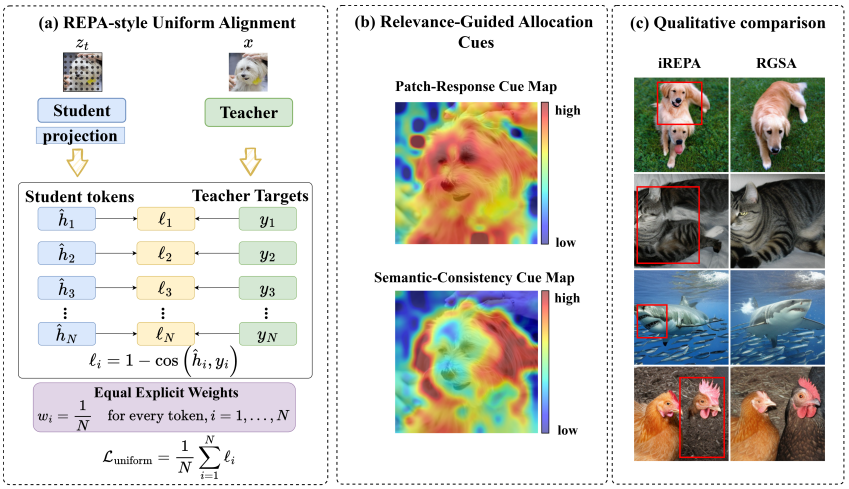



Figure 1: Overview and motivation of RGSA. (a) REPA-style alignment uniformly aggregates token-wise losses with coefficient $1/N$. (b) RGSA derives patch-response and semantic-consistency cues from frozen teacher representations as proxy signals for token relevance. (c) Paired generations from iREPA and RGSA under matched sampling conditions. Red boxes highlight regions for visual comparison. See Appendix E for generation settings and additional examples.

selves determine the relative emphasis placed on different spatial locations. Information at these locations may serve different supervisory roles. For example, in an image of a dog, the foreground may convey object-level semantics, while the background provides scene context; their value for diffusion representation learning may differ. However, REPA and iREPA (Yu et al., 2025; Singh et al., 2026) uniformly aggregate all $N$ token-wise alignment losses with a fixed coefficient of $1/N$, as illustrated in Fig. 1(a). Their explicit aggregation weights therefore remain independent of image content, even though individual losses and gradient contributions may differ.

Teacher representations themselves offer potential cues for this allocation. As shown in Fig. 1(b), patch-response strength and local–global consistency exhibit spatial patterns that vary with image content. These observable differences provide a potential basis for content-aware supervision allocation. This raises our central question: Can teacher-derived spatial relevance cues guide more effective supervision allocation?

To investigate this question, we propose *Relevance-Guided Token Supervision Allocation* (RGSA), which uses relevance cues from a frozen visual teacher to replace uniform aggregation with image-conditioned token-wise weighting. RGSA turns the spatial patterns described above into two proxy relevance cues. A *patch-response cue* uses patch-feature norms to characterize response strength, while a *semantic-consistency cue* uses patch–CLS cosine similarity to capture local–global semantic consistency. The two cues are fused into a nonnegative relevance score for each spatial token.

RGSA normalizes these scores across spatial tokens within each image to obtain supervision-allocation weights that sum to one. The resulting weights modulate the relative contribution of each token-wise alignment loss according to teacher-derived spatial evidence. Built upon iREPA, RGSA preserves the alignment targets, token correspondence, alignment criterion, and model architecture, modifying only the aggregation of the auxiliary alignment objective. It introduces no additional learnable parameters or inference-time computation.

We evaluate RGSA on ImageNet-100 and ImageNet-1K. Under matched training budgets, RGSA achieves lower FID and higher Recall than iREPA across SiT-B/2, SiT-L/2, and SiT-XL/2 on ImageNet-1K, with consistent gains across the evaluated visual teachers. Our analyses show that teacher-guided allocation deviates only mildly from uniform weighting, yet is accompanied by improved generation performance. Controlled ablations on ImageNet-100 further show that random and spatially shuffled weighting do not yield comparable gains. Together, these findings suggest that the benefits depend not merely on introducing non-uniform weights, but also on their correspondence with image content.

We summarize our main contributions as follows:

- We study how teacher supervision should be allocated across spatial tokens in diffusion representation alignment.
- We introduce *Relevance-Guided Token Supervision Allocation* (RGSA), which replaces uniform alignment aggregation with image-conditioned relevance weights from a frozen teacher, without additional learnable parameters or inference-time computation.
- RGSA improves upon iREPA across model capacities and visual teachers on ImageNet-1K. Controlled ablations on ImageNet-100 support the roles of relevance cues, spatial correspondence, and normalized allocation.

## 2 PRELIMINARIES

**Diffusion training.** Given an image $x$, a pretrained VAE encodes it into a latent $z$. Following SiT (Ma et al., 2024), we construct

$$z_t = \alpha_t z + \sigma_t \epsilon, \qquad t \in [0,1], \quad \epsilon \sim \mathcal{N}(0, I), \tag{1}$$

where $\alpha_t$ and $\sigma_t$ define the interpolation schedule. The model $v_\theta(z_t, t, y)$, conditioned on class label $y$, minimizes

$$\mathcal{L}_{\text{denoise}} = \mathbb{E}\left[\|v_\theta(z_t, t, y) - u_t\|_2^2\right], \qquad u_t = \dot{\alpha}_t z + \dot{\sigma}_t \epsilon. \tag{2}$$

Dots denote derivatives with respect to $t$, and the expectation is over training samples, timesteps, and noise. Let $H_t^l = \{h_{t,i}^l\}_{i=1}^N$ denote the spatial token grid from the $l$-th Transformer block.

**Token-wise representation alignment.** Following REPA (Yu et al., 2025), we align intermediate student features with frozen teacher representations and adopt iREPA (Singh et al., 2026) as our baseline. For the clean image $x$, the teacher encoder $g$ produces CLS and patch representations:

$$(z_{\text{cls}}^T, Z^T) = g(x), \qquad Y^T = \mathcal{S}(Z^T) = \{y_i^T\}_{i=1}^N, \tag{3}$$

where $\mathcal{S}$ denotes iREPA's spatial normalization. Feature extraction and spatial correspondence are detailed in Appendix A.1. A convolutional projector produces $Q_t^l = p_\phi(H_t^l) = \{q_{t,i}^l\}_{i=1}^N$, giving the token-wise alignment loss

$$\ell_i = 1 - \cos\left(q_{t,i}^l, y_i^T\right). \tag{4}$$

iREPA assigns each term the same explicit weight $1/N$:

$$\mathcal{L}_{\text{uniform}} = \frac{1}{N}\sum_{i=1}^N \ell_i. \tag{5}$$

## 3 RELEVANCE-GUIDED TOKEN SUPERVISION ALLOCATION

### 3.1 FROM UNIFORM AGGREGATION TO TOKEN-WISE SUPERVISION ALLOCATION

Building on the iREPA baseline introduced above, RGSA modifies only the aggregation of token-wise alignment losses, as illustrated in Fig. 2.

iREPA uniformly aggregates token-wise losses with a fixed coefficient $1/N$. Although the losses themselves depend on image content, their aggregation weights do not. We therefore view uniform aggregation as a content-independent supervision-allocation rule and generalize it as

$$\mathcal{L}_{\text{align}}(\mathbf{w}) = \sum_{i=1}^N w_i \ell_i, \qquad w_i \geq 0, \qquad \sum_{i=1}^N w_i = 1, \tag{6}$$

where $w_i$ controls the relative contribution of the $i$-th token-wise alignment term. Uniform iREPA is recovered when $w_i = 1/N$ for all $i$.

To construct $\mathbf{w}$ without introducing an additional allocation network, we use image-dependent statistics from the frozen teacher as proxy relevance cues, as detailed below.

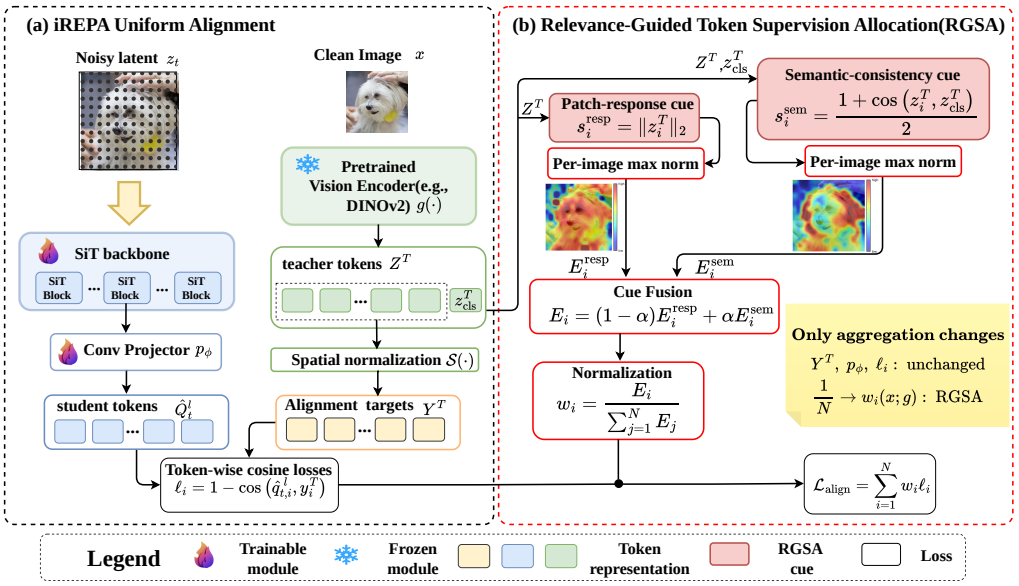



Figure 2: Overview of Relevance-Guided Token Supervision Allocation (RGSA). (a) iREPA uniformly aggregates token-wise alignment losses. (b) RGSA derives patch-response and semantic-consistency cues from frozen teacher representations, fuses them into token relevance scores, and normalizes them into image-conditioned allocation weights. The alignment targets, projector, and token-wise criterion remain unchanged.

### 3.2 Teacher-Derived Relevance Cues

RGSA extracts local feature response and patch–global representational consistency from the frozen teacher representations as two proxy relevance cues. These cues are used for supervision allocation rather than as direct estimates of token importance.

**Patch-response cue.** We use teacher patch-feature norms to characterize spatial response strength, drawing on the use of activation magnitudes in spatial attention transfer (Zagoruyko & Komodakis, 2017). Given teacher patch representations $Z^T = \{z^T_i\}_{i=1}^N$, we define

$$s^{\text{resp}}_i = \left\| z^T_i \right\|_2, \qquad E^{\text{resp}}_i = \frac{s^{\text{resp}}_i}{\max\left(\max_j s^{\text{resp}}_j, \varepsilon\right)}. \tag{7}$$

Here, $\varepsilon > 0$ ensures numerical stability, and per-image max normalization rescales feature magnitudes to $[0, 1]$ before cue fusion.

**Semantic-consistency cue.** To complement the patch-response cue, we use local–global representational consistency as an additional proxy for token relevance. Specifically, we compute the cosine similarity between each teacher patch representation and the CLS representation of the same image. We define

$$c_i = \cos\left(z^T_i, z^T_{\text{cls}}\right), \qquad s^{\text{sem}}_i = \frac{1 + c_i}{2}, \qquad E^{\text{sem}}_i = \frac{s^{\text{sem}}_i}{\max\left(\max_j s^{\text{sem}}_j, \varepsilon\right)}. \tag{8}$$

The affine transformation maps cosine similarity to $[0, 1]$ while preserving its ordering. The CLS representation is used only to construct this relevance cue and does not participate in patch-level alignment.

Here, $s^{\text{resp}}_i$ and $s^{\text{sem}}_i$ denote the raw cue scores, while $E^{\text{resp}}_i$ and $E^{\text{sem}}_i$ are their per-image max-normalized counterparts. Both cues are computed from teacher patch representations $Z^T$ before the iREPA spatial normalization that produces the alignment targets $Y^T$ in Eq. (3).

**Cue fusion.** We combine the normalized cue scores $E^{\text{resp}}_i$ and $E^{\text{sem}}_i$ into a fused relevance score:

$$E_i = (1 - \alpha)E^{\text{resp}}_i + \alpha E^{\text{sem}}_i, \qquad \alpha \in [0, 1], \tag{9}$$

where $\alpha$ is a fixed, non-learnable coefficient, set to 0.1 by default. The cases $\alpha = 0$ and $\alpha = 1$ correspond to patch-response-only and semantic-consistency-only allocation, respectively. We study the effect of $\alpha$ in Appendix D.2.

### 3.3 Normalized Relevance-Guided Allocation

Given the fused relevance score $E_i$, directly weighting the token-wise alignment losses gives

$$\mathcal{L}_{\text{unnorm}} = \frac{1}{N}\sum_{i=1}^{N} E_i \ell_i = \bar{E}\frac{\sum_{i=1}^{N} E_i \ell_i}{\sum_{i=1}^{N} E_i}, \qquad \bar{E} = \frac{1}{N}\sum_{i=1}^{N} E_i, \tag{10}$$

for $\sum_i E_i > 0$. This formulation couples token-wise redistribution with an image-dependent scalar factor $\bar{E}$. Consequently, when the alignment loss is weighted by $\lambda$, the effective auxiliary-loss scale also varies with the image through $\lambda\bar{E}$.

To isolate relative supervision allocation from this common scale, we normalize the fused relevance scores within each image to obtain allocation weights:

$$w_i = \frac{E_i}{\sum_{j=1}^{N} E_j}, \qquad \mathcal{L}_{\text{align}} = \sum_{i=1}^{N} w_i \ell_i = \frac{\sum_{i=1}^{N} E_i \ell_i}{\sum_{i=1}^{N} E_i}, \qquad \sum_{i=1}^{N} w_i = 1. \tag{11}$$

This normalization removes the common relevance scale while preserving the relative ratios among token scores. When $E_i = c > 0$ for all $i$, it reduces to $w_i = 1/N$, recovering uniform iREPA aggregation.

**Training objective.** The overall training objective combines the denoising loss in Eq. (2) with the alignment loss in Eq. (11):

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{denoise}} + \lambda \mathbb{E}\left[\mathcal{L}_{\text{align}}\right], \tag{12}$$

where the expectation is over training samples, timesteps, and noise, and $\lambda$ controls the alignment strength. The allocation weights adapt to image content through relevance cues derived from frozen teacher representations.

## 4 Experiments

We evaluate RGSA through controlled experiments and investigate the following questions:

- Can relevance-guided token-wise supervision allocation improve upon the uniform aggregation of iREPA across training budgets, model capacities, and pretrained visual teachers? (Tables 2,3, 4 and Fig. 7 )
- Which teacher-derived relevance cues and allocation design choices contribute to the performance of RGSA? (Table 5 and Fig. 4)
- Does RGSA produce token-level alignment outcomes consistent with its relevance-guided supervision allocation? (Fig. 5)

Table 1: SiT model configurations.

| Config. | Layers | Hidden dim. | Heads |
|---|---|---|---|
| B/2 | 12 | 768 | 12 |
| L/2 | 24 | 1024 | 16 |
| XL/2 | 28 | 1152 | 16 |

**Implementation details.** We follow iREPA (Singh et al., 2026) and conduct our main experiments on ImageNet-1K (Deng et al., 2009), using a fixed ImageNet-100 subset for controlled ablations. Images are processed at $256 \times 256$ following ADM (Dhariwal & Nichol, 2021) and encoded into $32 \times 32 \times 4$ latents using the same pretrained Stable Diffusion VAE (Rombach et al., 2022) as iREPA. We evaluate SiT-B/2, SiT-L/2, and SiT-XL/2 (Ma et al., 2024), all with a patch size of 2. Models are trained using AdamW with a learning rate of $1 \times 10^{-4}$ and an effective batch size of 256. ImageNet-100 ablations use 50K steps, while ImageNet-1K experiments use 100K steps by default and 400K steps for the long-horizon SiT-XL/2 comparison. Methods compared within each setting share the same training budget and hyperparameters and are evaluated using their final EMA checkpoints.

**Evaluation and sampling.** We report FID (Heusel et al., 2017), sFID (Nash et al., 2021), IS (Salimans et al., 2016), and generative Precision and Recall (Kynkäänniemi et al., 2019). ImageNet-1K

results use 50,000 generated samples, whereas ImageNet-100 ablations use 5,000 samples for relative comparison among controlled variants. Following iREPA, we use the SDE Euler–Maruyama sampler with 250 function evaluations. Unless otherwise specified, sampling is performed without classifier-free guidance . Full evaluation details are provided in Appendix A.

**Teacher encoders.** We use DINOv2-B/14 (Oquab et al., 2023) as the default frozen teacher and additionally evaluate CLIP ViT-L/14 (Radford et al., 2021) and MoCo v3 ViT-B/16 (Chen et al., 2021). Teacher-specific preprocessing and feature extraction are detailed in Appendix A.

### 4.1 Main Results on ImageNet-1K

We evaluate RGSA on ImageNet-1K across training budgets, pretrained visual teachers, and model sizes, using iREPA as the direct baseline.

**Performance across training budgets.** We compare RGSA with REPA and iREPA on SiT-XL/2 after 100K and 400K training steps (Table 2). At 100K steps, RGSA improves over iREPA under both sampling protocols. Without CFG, it reduces FID from 16.90 to 15.79 while improving sFID and Recall. With CFG, it reduces FID from 5.15 to 4.41 and improves IS, sFID, and Precision while maintaining comparable Recall. At 400K steps, RGSA reduces FID from 7.52 to 7.47 without CFG and from 1.93 to 1.91 with CFG. Additional qualitative comparisons under matched sampling conditions are provided in Appendix E (Fig. 7). These results support the effectiveness of relevance-guided supervision allocation across the evaluated training budgets and sampling protocols.

Table 2: Comparison of REPA, iREPA, and RGSA with and without classifier-free guidance using SiT-XL/2 on ImageNet-1K.

| Method | Steps | w/o CFG | | | | | w/ CFG | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | IS ↑ | FID ↓ | sFID ↓ | Prec. ↑ | Rec. ↑ | IS ↑ | FID ↓ | sFID ↓ | Prec. ↑ | Rec. ↑ |
| REPA (Yu et al., 2025) | 100K | 69.20 | 19.30 | 5.89 | 0.640 | 0.610 | 157.20 | 6.35 | 5.91 | 0.769 | 0.536 |
| iREPA (Singh et al., 2026) | 100K | 77.92 | 16.90 | 6.26 | **0.660** | 0.610 | 179.30 | 5.15 | 6.23 | 0.783 | **0.544** |
| RGSA (Ours) | 100K | **79.55** | **15.79** | **5.64** | 0.655 | **0.624** | **187.07** | **4.41** | **5.76** | **0.796** | 0.540 |
| REPA (Yu et al., 2025) | 400K | 127.40 | 7.76 | 5.06 | 0.700 | **0.660** | 263.00 | 1.98 | 4.60 | 0.799 | **0.610** |
| iREPA (Singh et al., 2026) | 400K | 128.60 | 7.52 | 4.89 | **0.710** | 0.650 | 268.80 | 1.93 | 4.59 | 0.799 | 0.600 |
| RGSA (Ours) | 400K | **129.49** | **7.47** | **4.56** | 0.703 | 0.657 | **275.15** | **1.91** | **4.52** | **0.813** | 0.601 |

**Compatibility across visual teachers.** We evaluate SiT-B/2 trained for 100K steps on ImageNet-1K without classifier-free guidance, using frozen DINOv2-B/14 (Oquab et al., 2023), CLIP ViT-L/14 (Radford et al., 2021), and MoCo v3 ViT-B/16 (Chen et al., 2021) teachers. Under matched configurations with $\alpha = 0.1$ for RGSA, Table 3 shows improvements over iREPA in IS, FID, and sFID across all three teachers, with comparable or better Precision and Recall, supporting compatibility across visual teachers.

**Model scalability.** We evaluate RGSA with SiT-B/2, SiT-L/2, and SiT-XL/2 under the same 100K-step training protocol. As shown in Table 4, RGSA reduces FID over iREPA by approximately 6.1%, 5.0%, and 6.6%, respectively, while also improving IS and Recall across all three capacities. These results show that the gains of RGSA are not specific to a particular model scale.

Table 3: Generalization across visual teachers using SiT-B/2 trained for 100K steps on ImageNet-1K and evaluated without classifier-free guidance.

| Teacher | Method | IS ↑ | FID ↓ | sFID ↓ | Prec. ↑ | Rec. ↑ |
|---|---|---|---|---|---|---|
| DINOv2-B/14 | iREPA | 34.10 | 43.37 | 7.87 | **0.50** | 0.60 |
| | RGSA | **36.67** | **40.71** | **7.73** | **0.50** | **0.62** |
| CLIP ViT-L/14 | iREPA | 35.76 | 41.30 | 7.59 | 0.49 | **0.61** |
| | RGSA | **36.66** | **40.59** | **7.36** | **0.50** | **0.61** |
| MoCo v3 ViT-B/16 | iREPA | 32.21 | 42.60 | 6.97 | **0.49** | 0.61 |
| | RGSA | **32.35** | **41.90** | **6.74** | **0.49** | **0.62** |

Table 4: Scaling across SiT capacities on ImageNet-1K. All models are trained for 100K steps and evaluated using 50K samples without CFG.

| Model | Method | IS ↑ | FID ↓ | sFID ↓ | Prec. ↑ | Rec. ↑ |
|---|---|---|---|---|---|---|
| SiT-B/2 | REPA | 27.50 | 49.50 | **7.00** | 0.46 | 0.59 |
| | iREPA | 34.10 | 43.37 | 7.87 | **0.50** | 0.60 |
| | RGSA | **36.67** | **40.71** | 7.73 | **0.50** | **0.62** |
| SiT-L/2 | REPA | 55.70 | 24.10 | 6.25 | 0.62 | 0.60 |
| | iREPA | 66.90 | 20.28 | 6.14 | **0.63** | 0.60 |
| | RGSA | **67.80** | **19.26** | **5.75** | **0.63** | **0.62** |
| SiT-XL/2 | REPA | 69.20 | 19.30 | 5.89 | 0.64 | 0.61 |
| | iREPA | 77.92 | 16.90 | 6.26 | **0.66** | 0.61 |
| | RGSA | **79.55** | **15.79** | **5.64** | 0.65 | **0.62** |

Table 5: Controlled ablations of RGSA on ImageNet-100.

| Category | Variant | FID ↓ | Rec. ↑ |
|---|---|---|---|
| Baseline | Uniform (iREPA) | 110.87 | 0.4327 |
| Signal | CLS attention | 116.02 | 0.4256 |
| | Loss-guided | 111.24 | 0.4225 |
| | Response only | **108.65** | 0.4175 |
| | Semantic only | 110.10 | 0.4401 |
| Spatial | Random | 110.72 | 0.4230 |
| | Shuffled | 112.58 | 0.4358 |
| Norm. | Fused w/o norm | 109.60 | 0.4478 |
| | RGSA | 108.92 | **0.4526** |

### 4.2 Controlled Ablations of Supervision Allocation

**Allocation signals and cue composition.** Table 5 compares alternative supervision-allocation signals. CLS attention and detached loss-guided weighting improve neither FID nor Recall over uniform iREPA. The proposed cues exhibit different trade-offs: response-only achieves lower FID, whereas semantic-only achieves higher Recall. Their fusion in RGSA provides a favorable fidelity–coverage trade-off. The full $\alpha$ sweep is reported in Appendix D.2.

**Spatial correspondence and normalization.** Random weighting and spatial shuffling fail to reproduce the joint FID–Recall gains of RGSA, supporting the importance of matching allocation weights to teacher-derived spatial cues. Constructions of the Random, Shuffled, Loss-guided, and CLS-attention controls are detailed in Appendix D.1. Using fused relevance scores without normalization already improves FID and Recall over uniform iREPA. Normalizing these scores further improves both metrics in this setting. Appendix C provides additional analysis of alignment-coefficient sensitivity and gradient behavior.

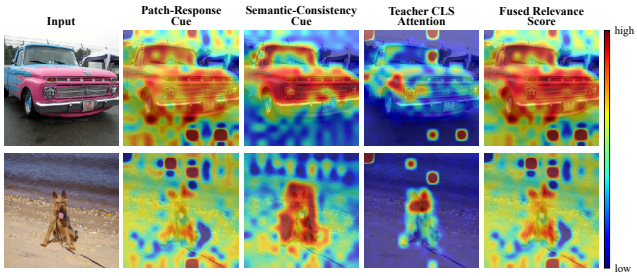



Figure 3: Spatial signals from a frozen DINOv2-B/14 teacher. The fused relevance score combines patch-response and semantic-consistency cues with $\alpha = 0.1$.

### 4.3 Analysis of RGSA

We analyze RGSA from two complementary perspectives. First, we characterize the spatial behavior, non-redundancy, and allocation distribution induced by the teacher-derived relevance cues. Second, we examine how this allocation affects token-level alignment relative to uniform iREPA across diffusion timesteps.

#### 4.3.1 Cue Behavior and Allocation Distribution

**Spatial behavior and cue non-redundancy.** We first examine the spatial behavior of the teacher-derived allocation cues. As shown in Fig. 3, patch-response and semantic-consistency exhibit distinct spatial patterns, while teacher CLS attention is substantially more localized and sparse. At the default $\alpha = 0.1$, the fused relevance score largely follows patch-response while incorporating variation from semantic-consistency, indicating that RGSA does not simply recover teacher attention.

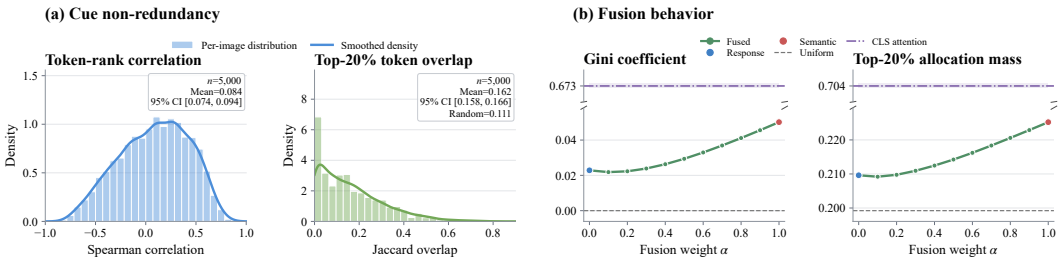



Figure 4: Dataset-level analysis of cue non-redundancy and allocation concentration. (a) Per-image Spearman rank correlation and top-20% Jaccard overlap between patch-response and semantic-consistency; annotations report the mean and 95% bootstrap confidence interval, with the random Jaccard reference shown for comparison. (b) Gini coefficient and top-20% allocation mass across fusion coefficients $\alpha$, with uniform allocation and normalized teacher CLS attention shown as references. Statistics are computed over 5,000 ImageNet-1K images.

We further quantify the agreement between the two RGSA cues over spatial tokens. As shown in Fig. 4(a), their mean Spearman correlation is 0.084, and the top-20% Jaccard overlap is 0.162, only moderately above the 0.111 random reference. These results indicate weak ranking agreement and limited overlap between the top-ranked token sets, suggesting that the two cues provide partially overlapping but distinct spatial rankings.

**Allocation distribution.** As shown in Fig. 4(b), RGSA produces a non-uniform but only mildly concentrated supervision allocation. Across fusion coefficients, the Gini coefficient ranges from approximately 0.02 to 0.05, while the top-20% allocation mass ranges from 0.209 to 0.225. In contrast, the normalized CLS-attention reference is substantially more concentrated, with a Gini coefficient of 0.673 and a top-20% mass of 0.704. These results characterize RGSA as a mild, image-dependent redistribution of supervision rather than a highly concentrated allocation over a small subset of tokens.

### 4.3.2 Token-Level Alignment Outcomes

We examine whether RGSA yields larger alignment improvements for tokens assigned higher teacher-derived relevance scores. For paired comparison, RGSA and iREPA are evaluated using identical images, diffusion timesteps, noise realizations, and teacher targets.

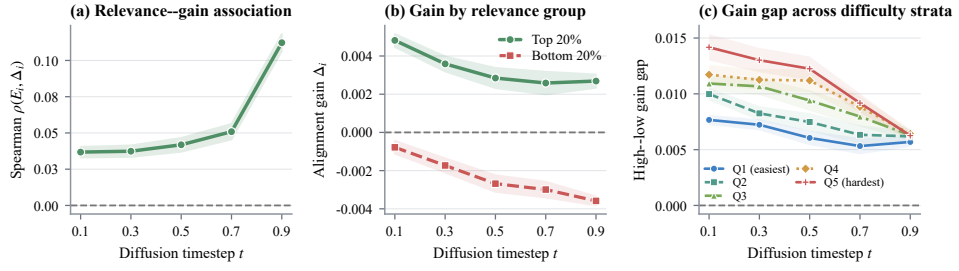



Figure 5: Token-level alignment outcomes after 100K training steps. We define the paired token-level alignment improvement as $\Delta_i = \ell_i^{\mathrm{iREPA}} - \ell_i^{\mathrm{RGSA}}$, where $\Delta_i > 0$ indicates lower alignment error under RGSA for the same token. (a) Per-image Spearman correlation between token relevance $E_i$ and $\Delta_i$. (b) Mean $\Delta_i$ for the top- and bottom-20% relevance tokens. (c) High–low relevance gain gap across five baseline-difficulty strata, defined by quintiles of the iREPA token-wise alignment error (Q1: easiest; Q5: hardest). Shaded regions denote image-level bootstrap 95% confidence intervals over 5,000 ImageNet-1K images.

**Relevance–gain association.** Figure 5(a) shows a modest positive association between token relevance $E_i$ and paired alignment improvement $\Delta_i$ at every evaluated diffusion timestep. Thus, tokens assigned higher relevance scores tend to exhibit larger alignment improvements relative to uniform iREPA.

**Alignment outcomes across relevance groups.** Figure 5(b) compares tokens in the top and bottom 20% of the relevance distribution. The top-20% group has positive mean $\Delta_i$, whereas the bottom-20% group has negative mean $\Delta_i$ across the evaluated timesteps. RGSA therefore improves mean alignment in the high-relevance group while increasing mean alignment error in the low-relevance group relative to iREPA. This pattern is consistent with a relevance-dependent redistribution of alignment improvements.

**Alignment gains across baseline difficulty.** We use the iREPA token-wise alignment error as a descriptive measure of baseline difficulty and partition tokens into five equal-sized strata, from Q1 (easiest) to Q5 (hardest). Within each stratum, we compare mean $\Delta_i$ between high- and low-relevance tokens. As shown in Fig. 5(c), this high–low gain gap is positive across all five strata and all evaluated timesteps. At low-to-intermediate diffusion timesteps, the gap is generally larger in harder strata; the differences across strata narrow at the largest evaluated timestep. These comparisons show that the relevance-dependent gain pattern is observed within each baseline-difficulty stratum.

Together, these results provide descriptive token-level evidence that RGSA preferentially improves alignment for tokens assigned higher teacher-derived relevance scores. This behavior is consistent with the intended supervision allocation.

## 5 Related Work

**Diffusion representation alignment.** REPA (Yu et al., 2025) aligns diffusion features with clean-image representations from a frozen encoder. iREPA (Singh et al., 2026) emphasizes teacher spatial structure, while UniRA (Chen et al., 2026) and sREPA (Xu et al., 2026) introduce consistency and relational constraints, respectively. For guidance scheduling, HASTE (Wang et al., 2026c) combines feature and attention alignment with early termination, while REED (Wang et al., 2026a) balances representation learning and generation through a curriculum. U-REPA (Tian et al., 2026) and REPA-E (Leng et al., 2025) extend alignment to diffusion U-Nets and joint VAE–diffusion training, respectively. At the token level, MaskAlign (Pang et al., 2026) combines pre-mask token mixing and random patch-token masking, uniformly aggregating alignment losses over retained tokens. Our work builds on diffusion representation alignment and focuses on how teacher supervision is allocated across all spatial tokens, using image-conditioned relevance weights in place of uniform aggregation.

**Adaptive supervision in representation transfer.** Spatially or token-adaptive supervision has been explored in representation distillation. ACAM-KD (Lan & Tian, 2025) uses adaptive spatial–channel masks for feature selection, while MoVE-KD (Cao et al., 2025) employs attention-based distillation to weight visual encoders and emphasize informative visual tokens. HAWAII (Wang et al., 2026b) uses token importance scores for fine-grained distillation from multiple teachers. These methods primarily address discriminative or vision–language distillation. Our work explores a similar principle in diffusion representation alignment, using teacher-derived relevance to allocate supervision while preserving iREPA's alignment targets and token-wise loss.

## 6 Conclusion

In this paper, we study how teacher supervision should be allocated across spatial tokens once token-wise correspondence is established in diffusion representation alignment. We introduce *Relevance-Guided Token Supervision Allocation* (RGSA), which derives image-conditioned relevance cues from a frozen visual teacher to redistribute token-wise alignment supervision without additional learnable parameters or inference-time computation. Experiments on ImageNet-1K show consistent improvements over iREPA across model capacities and pretrained visual teachers. Controlled ablations on ImageNet-100 support the roles of teacher-derived relevance cues, spatial correspondence, and normalized allocation. Token-level analyses further show that alignment improvements tend to favor tokens assigned higher teacher-derived relevance scores. These findings highlight supervision allocation as a distinct design dimension of diffusion representation alignment.

## AI USE STATEMENT

Generative AI tools were used to assist with language editing, manuscript drafting, and feedback on experimental presentation and interpretation. All AI-assisted content was reviewed and verified by the authors, who take full responsibility for the final manuscript, experimental design, results, and conclusions.

## REFERENCES

Jiajun Cao, Yuan Zhang, Tao Huang, Ming Lu, Qizhe Zhang, Ruichuan An, Ningning Ma, and Shanghang Zhang. Move-kd: Knowledge distillation for vlms with mixture of visual encoders. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19846–19856. IEEE, 2025.

Hesen Chen, Mengping Yang, Junyan Wang, Zhiyu Tan, and Hao Li. UniRA: Unified representation alignment for diffusion models via local, structural, and global constraints, 2026.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9620–9629. IEEE, 2021.

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*, pp. 248–255. Ieee, 2009.

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

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

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.

Qizhen Lan and Qing Tian. Acam-kd: adaptive and cooperative attention masking for knowledge distillation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3957–3966. IEEE, 2025.

Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. Repa-e: Unlocking vae for end-to-end tuning with latent diffusion transformers. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18262–18272. IEEE, 2025.

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

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

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.

Lianyu Pang, Tianlin Pan, Cheng Da, Changqian Yu, Huan Yang, Kun Gai, Song Guo, and Wenhan Luo. Maskalign: Token-subset representation alignment for efficient diffusion training. *arXiv preprint arXiv:2606.08788*, 2026.

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

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.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

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.

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.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Yuchuan Tian, Hanting Chen, Mengyu Zheng, Yuchen Liang, Chao Xu, and Yunhe Wang. U-repa: Aligning diffusion u-nets to vits. *Advances in Neural Information Processing Systems*, 38:11003–11024, 2026.

Chenyu Wang, Cai Zhou, Sharut Gupta, Johnson Lin, Stefanie Jegelka, Stephen Bates, and Tommi Jaakkola. Learning diffusion models with flexible representation guidance. *Advances in Neural Information Processing Systems*, 38:131176–131222, 2026a.

Yimu Wang, Mozhgan Nasr Azadani, Sean Sedwards, and Krzysztof Czarnecki. Hawaii: Hierarchical visual knowledge transfer for efficient vision-language models. *Advances in Neural Information Processing Systems*, 38:919–943, 2026b.

Ziqiao Wang, Wangbo Zhao, Yuhao Zhou, Zekai Li, Zhiyuan Liang, Mingjia Shi, Xuanlei Zhao, Pengfei Zhou, Kaipeng Zhang, Zhangyang Wang, et al. Repa works until it doesn't: Early-stopped, holistic alignment supercharges diffusion training. *Advances in Neural Information Processing Systems*, 38:136854–136887, 2026c.

Shaodong Xu, Zhendong Wang, Litong Gong, Zexian Li, Wengang Zhou, Tiezheng Ge, and Houqiang Li. Beyond point-wise matching: Structural representation alignment for accelerating diffusion transformers. *arXiv preprint arXiv:2605.16949*, 2026.

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

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *International Conference on Learning Representations*, 2017.

## A Additional Implementation Details

**Data and latent preprocessing.** Images are center-cropped and resized to $256 \times 256$. We use precomputed posterior moments from the frozen Stable Diffusion VAE (Rombach et al., 2022), using the `stabilityai/sd-vae-ft-mse` checkpoint.[1] During training, latents are sampled from the corresponding posterior and normalized using the channel-wise statistics provided by iREPA (Singh et al., 2026).

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

Table 6: Default training configuration.

| Configuration | Value |
|---|---|
| Image resolution | $256 \times 256$ |
| Latent resolution | $32 \times 32 \times 4$ |
| Latent encoder | stabilityai/sd-vae-ft-mse |
| Probability path | Linear |
| Prediction target | Velocity |
| Optimizer | AdamW |
| Learning rate | $1 \times 10^{-4}$ |
| AdamW betas | $(0.9, 0.999)$ |
| Weight decay | 0 |
| Learning-rate schedule | Constant |
| Effective batch size | 256 |
| Mixed precision | FP16 |
| Gradient clipping | 1.0 |
| EMA decay | 0.9999 |
| EMA update frequency | Every optimizer step |
| Class dropout probability | 0.1 |
| Default teacher | DINOv2-B/14 |
| Student alignment depth | 8 |
| Alignment coefficient | 1.0 |
| Spatial-centering coefficient | $\gamma = 0.6$ |
| Fusion coefficient | $\alpha = 0.1$ |
| Random seed | 0 |
| ImageNet-100 budget | 50K steps |
| ImageNet-1K budget | 100K steps |
| Hardware | Up to 3× RTX 3090 (24 GB) |

**Training protocol.** The long-horizon SiT-XL/2 experiment is trained for 400K steps; all other training budgets follow Table 6. Experiments are conducted on a server equipped with three NVIDIA RTX 3090 GPUs. Runtime and memory comparisons use identical hardware configurations.

### A.1 Teacher-Specific Feature Extraction

All teacher inputs are derived from the same deterministic $256 \times 256$ center-cropped image paired with the corresponding cached VAE latent. The DINOv2 (Oquab et al., 2023) and CLIP (Radford et al., 2021) inputs are bicubically resized to $224 \times 224$, whereas MoCo v3 (Chen et al., 2021) directly receives the $256 \times 256$ crop. No additional random cropping or horizontal flipping is applied in the teacher branch. Teacher-specific preprocessing, positional-embedding adaptation, and feature extraction are summarized in Table 7. All teacher encoders remain frozen throughout training.

Table 7: Teacher-specific preprocessing and feature extraction. Input normalization denotes channel-wise pixel standardization using the mean and standard deviation specified for each pretrained teacher. All features are extracted before iREPA spatial normalization.

| Teacher | Input | Patch | Grid | Dim. | Input norm. | Positional embedding | Feature stage |
|---|---|---|---|---|---|---|---|
| DINOv2-B/14 | $224^2$ | 14 | $16^2$ | 768 | ImageNet | Resampled to $16^2$ | Final LayerNorm outputs |
| CLIP ViT-L/14 | $224^2$ | 14 | $16^2$ | 1024 | CLIP | Native $16^2$ | Final Transformer, pre-`ln_post`/projection |
| MoCo v3 ViT-B/16 | $256^2$ | 16 | $16^2$ | 768 | ImageNet | $14^2 \rightarrow 16^2$ | Final LayerNorm outputs |

**DINOv2.** For DINOv2-B/14 (Oquab et al., 2023), the pretrained absolute positional embedding is resampled to match the $16 \times 16$ patch grid. The $256 \times 256$ crop is bicubically resized to $224 \times 224$, yielding a $16 \times 16$ patch grid with patch size 14. We extract `x_norm_clstoken` and `x_norm_patchtokens` after the final LayerNorm.

**CLIP.** For CLIP ViT-L/14 (Radford et al., 2021), the native positional embedding already corresponds to a $16 \times 16$ patch grid at $224 \times 224$ resolution and is used without interpolation. We wrap

the frozen visual tower to expose the complete final Transformer sequence $[\mathrm{CLS}, z_1^T, \ldots, z_N^T]$ before `ln_post` and the visual output projection. The text encoder is not used.

**MoCo v3.** For MoCo v3 ViT-B/16 (Chen et al., 2021), the teacher operates directly on the $256 \times 256$ crop. The pretrained positional embedding is resampled from a $14 \times 14$ patch grid to a $16 \times 16$ grid. We extract the CLS and patch tokens after the final LayerNorm.

**Spatial correspondence and target construction.** The VAE produces a $32 \times 32$ latent grid, which SiT with patch size 2 maps to a $16 \times 16$ student-token grid. Each teacher likewise produces a $16 \times 16$ patch grid. Teacher and student patch tokens are therefore matched directly by spatial index, without feature-level interpolation or pooling after teacher extraction. The convolutional projector preserves the spatial grid resolution while mapping student features to the teacher feature dimension.

The patch-response and semantic-consistency cues are computed from the teacher outputs before iREPA spatial normalization, whereas the spatially normalized patch representations serve as alignment targets. The CLS representation is used only to compute the semantic-consistency cue and is excluded from patch-level cosine alignment.

## B Computational Efficiency

RGSA reuses the frozen teacher representations already extracted by iREPA and adds only patch-response computation, patch–CLS similarity, and token-wise weighted aggregation. These operations incur $\mathcal{O}(ND_t)$ complexity, where $N$ is the number of spatial tokens and $D_t$ is the teacher feature dimension. RGSA involves neither token-to-token pairwise computation nor additional learnable parameters. During generation, all alignment-related components are removed, leaving the inference graph and sampling cost unchanged.

Table 8: Training efficiency on SiT-XL/2 under identical hardware and training settings. Measurements are collected after warm-up using FP16.

| Method | Params (M) | Time/step (s) | Images/s | Memory (GiB) |
|---|---|---|---|---|
| iREPA | 675.70 | 4.9921 | 51.28 | 20.40 |
| RGSA | 675.70 | 4.9857 | 51.35 | 20.41 |

As shown in Table 8, the runtime and throughput differences between iREPA and RGSA are below $0.2\%$, while peak memory increases by only $0.01$ GiB. The small timing difference falls within typical runtime variation and should not be interpreted as a speed improvement. These results indicate that RGSA introduces negligible training overhead under the evaluated setting.

## C Alignment-Coefficient Sensitivity and Gradient Diagnostics

We examine whether RGSA's generation gains persist across alignment coefficients and characterize the accompanying gradient behavior. On ImageNet-100, we train SiT-B/2 for 50K steps with $\lambda \in \{0.5, 0.75, 1.0, 1.25\}$. At each coefficient, RGSA and uniform iREPA use identical training configurations and differ only in token-wise alignment-loss aggregation.

**Coefficient sensitivity.** Table 9 shows that RGSA achieves lower FID and higher Recall than uniform iREPA at all four tested coefficients. FID decreases by 1.21–2.92, while Recall increases by 0.0044–0.0199 in absolute terms. Within RGSA, the default $\lambda = 1.0$ yields the lowest FID and sFID and the highest Recall among the tested settings. Thus, although performance varies with $\lambda$, the FID and Recall improvements are not confined to the default coefficient.

**Gradient measurement.** Let $\theta$ denote the SiT backbone parameters and $\vartheta = (\theta, \phi)$ all trainable parameters, including the alignment projector. Using minibatch loss estimates, we measure

$$g_{\mathrm{den}} = \left\| \nabla_\theta \widehat{\mathcal{L}}_{\mathrm{denoise}} \right\|_2, \qquad g_{\mathrm{align}} = \left\| \nabla_\theta \left( \lambda \widehat{\mathcal{L}}_{\mathrm{align}} \right) \right\|_2, \qquad R_{\mathrm{align/den}} = \frac{g_{\mathrm{align}}}{g_{\mathrm{den}} + \varepsilon}. \tag{13}$$

Table 9: Performance on ImageNet-100 using SiT-B/2 trained for 50K steps.

| Method | $\lambda$ | IS ↑ | FID ↓ | sFID ↓ | Precision ↑ | Recall ↑ |
|---|---|---|---|---|---|---|
| iREPA | 0.5 | **12.93** | 110.28 | 45.59 | **0.2672** | 0.4091 |
| RGSA | 0.5 | 12.78 | **109.07** | **44.72** | 0.2646 | **0.4280** |
| iREPA | 0.75 | 12.68 | 112.66 | 44.63 | 0.2534 | 0.4207 |
| RGSA | 0.75 | **13.10** | **109.74** | **44.08** | **0.2560** | **0.4251** |
| iREPA | 1.0 | 12.93 | 110.87 | 44.20 | 0.2480 | 0.4327 |
| RGSA | 1.0 | **12.98** | **108.92** | **43.52** | **0.2492** | **0.4526** |
| iREPA | 1.25 | 12.50 | 111.92 | **47.07** | 0.2666 | 0.4010 |
| RGSA | 1.25 | **12.99** | **110.52** | 47.31 | **0.2718** | **0.4193** |

The two component norms are computed on the same synchronized microbatch using `torch.autograd.grad`, without modifying the gradients used for optimization. We also record the total gradient norm over $\vartheta$ after gradient accumulation and before clipping. Because the component and total norms differ in parameter scope and accumulation stage, they serve as complementary diagnostics rather than an algebraic decomposition of the total norm.

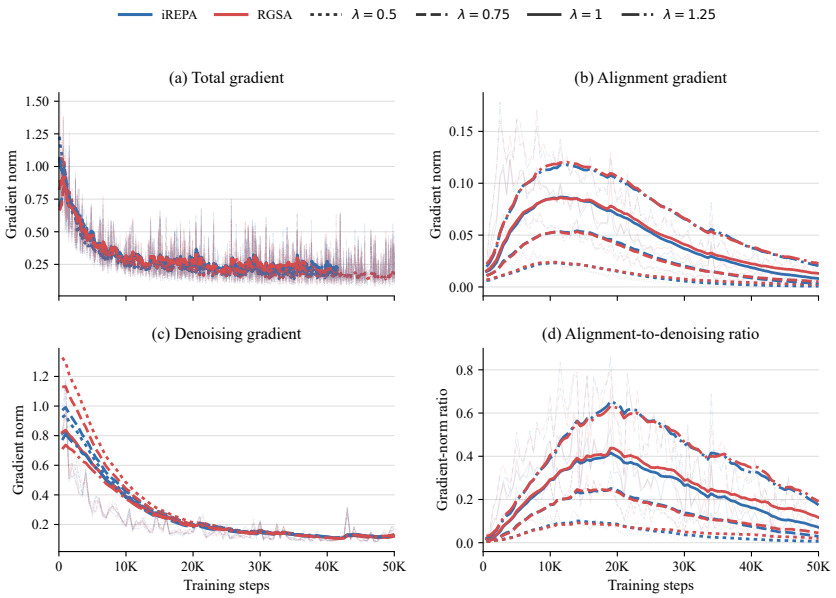



Figure 6: Gradient diagnostics on ImageNet-100 using SiT-B/2 trained for 50K steps with $\lambda \in \{0.5, 0.75, 1.0, 1.25\}$. (a) Pre-clipping total gradient norm over all trainable parameters. (b) Backbone gradient norm from the weighted alignment loss. (c) Backbone gradient norm from the denoising loss. (d) Alignment-to-denoising gradient-norm ratio. Colors distinguish methods, and line styles distinguish coefficients. Faint curves show raw measurements; bold curves show EMA smoothing with a decay of 0.9.

**Gradient behavior.** Figure 6(a,c) shows broadly comparable total-gradient norms between iREPA and RGSA at each tested coefficient, with denoising-gradient norms also approaching similar levels later in training. Alignment-gradient norms and alignment-to-denoising ratios generally increase with $\lambda$ (Fig. 6(b,d)). At matched coefficients, the two methods follow similar overall gradient trajectories, with some differences in the relative strength of alignment gradients.

RGSA achieves lower FID and higher Recall at all four tested alignment coefficients, indicating that its gains are not confined to the default setting. These gradient diagnostics characterize the

optimization behavior accompanying these gains and complement the analysis in Sec. 4.3.2, which shows larger alignment improvements for higher-relevance tokens.

# D Additional Ablation Results

## D.1 Construction of Controlled Ablations

The Random, Shuffled, Loss-guided, and CLS-attention controls retain the same teacher targets and token-wise alignment loss $\ell_i$ as RGSA, and differ only in their non-negative allocation scores $\widetilde{E}_i$. We detail their construction below and report the complete evaluation metrics in Table 10. The allocation weights and alignment loss are

$$w_i = \frac{\widetilde{E}_i}{\max\left(\sum_{j=1}^{N} \widetilde{E}_j, \epsilon\right)}, \qquad \mathcal{L}_{\text{align}} = \sum_{i=1}^{N} w_i \ell_i, \tag{14}$$

where $\epsilon = 10^{-6}$. The weights sum to one whenever $\sum_j \widetilde{E}_j \geq \epsilon$.

**Random.** At each training forward pass, we independently sample $\widetilde{E}_i^{\text{rand}} \sim \mathcal{U}(0,1)$ for every image and spatial token. These scores are independent of image content and spatial position.

**Shuffled.** At each training forward pass, an independent random permutation $\pi$ is sampled for every image. We permute the fused relevance scores from Eq. (9):

$$\widetilde{E}_i^{\text{shuf}} = E_{\pi(i)}. \tag{15}$$

This preserves the per-image multiset of scores while randomizing their correspondence with spatial tokens.

**Loss-guided.** At each training forward pass, we use the current alignment errors:

$$\widetilde{E}_i^{\text{loss}} = \max(\text{stopgrad}(\ell_i), \epsilon)\,. \tag{16}$$

This assigns greater weight to larger errors, subject to the numerical floor, without backpropagating through the allocation rule.

**CLS-attention.** We extract CLS-query attention from the frozen DINOv2 teacher's final Transformer block, with softmax computed over all $N+1$ keys, including the CLS token. After removing the CLS-to-CLS entry, let $a_{h,i}$ denote the attention to patch $i$ in head $h$. We average across all $H$ attention heads to obtain

$$\widetilde{E}_i^{\text{attn}} = \text{stopgrad}\left(\frac{1}{H}\sum_{h=1}^{H} a_{h,i}\right). \tag{17}$$

The resulting scores are normalized across patches within each image using Eq. (14).

Table 10: Complete metrics for the controlled ablations on ImageNet-100.

| Category | Variant | IS ↑ | FID ↓ | sFID ↓ | Prec. ↑ | Rec. ↑ |
|---|---|---|---|---|---|---|
| Baseline | Uniform (iREPA) | 12.934 | 110.87 | 44.20 | 0.2480 | 0.4327 |
| Allocation signal | Patch-response only | 13.010 | **108.65** | 43.78 | 0.2474 | 0.4175 |
| | Semantic-consistency only | **13.070** | 110.10 | **43.29** | 0.2430 | 0.4401 |
| | CLS-attention | 12.311 | 116.02 | 45.40 | 0.2324 | 0.4256 |
| | Detached loss-guided allocation | 12.607 | 111.24 | 44.08 | 0.2316 | 0.4225 |
| Spatial control | Random | 12.950 | 110.72 | 43.80 | 0.2480 | 0.4230 |
| | Shuffled | 12.480 | 112.58 | 43.39 | 0.2443 | 0.4358 |
| Normalization | Fused w/o norm | 12.893 | 109.60 | 43.44 | **0.2492** | 0.4478 |
| | RGSA | 12.984 | 108.92 | 43.52 | **0.2492** | **0.4526** |

Table 11: Effect of $\alpha$ on ImageNet-100.

| $\alpha$ | FID ↓ | Rec. ↑ |
|---|---|---|
| 0.0 | **108.65** | 0.4175 |
| 0.1 | 108.92 | **0.4526** |
| 0.2 | 110.29 | 0.4444 |
| 0.3 | 111.40 | 0.4458 |
| 0.5 | 112.20 | 0.4443 |
| 0.7 | 109.38 | 0.4459 |
| 1.0 | 110.10 | 0.4401 |

### D.2 Sensitivity to the Fusion Coefficient

We provide the complete sweep of the fusion coefficient $\alpha$ on ImageNet-100. All experiments use SiT-B/2 trained for 50K steps and are evaluated with 5,000 generated samples under the same protocol as the controlled ablations in Sec. 4.2. The endpoints $\alpha = 0$ and $\alpha = 1$ correspond to patch-response-only and semantic-consistency-only allocation, respectively.

As shown in Table 11, patch-response-only allocation ($\alpha = 0$) achieves the lowest FID, while adding a small semantic-consistency contribution improves Recall. In particular, $\alpha = 0.1$ preserves an FID close to the patch-response endpoint while achieving the highest Recall. Larger values of $\alpha$ do not provide further improvements in this setting. We therefore use $\alpha = 0.1$ as the default fusion coefficient throughout the main experiments.

## E Additional Qualitative Results

Both the main-text and additional qualitative comparisons use SiT-XL/2 trained for 400K steps with DINOv2-B/14 as the frozen visual teacher and a classifier-free guidance scale of 4.0. Each pair uses the same class condition, initial latent, and sampler noise realization.

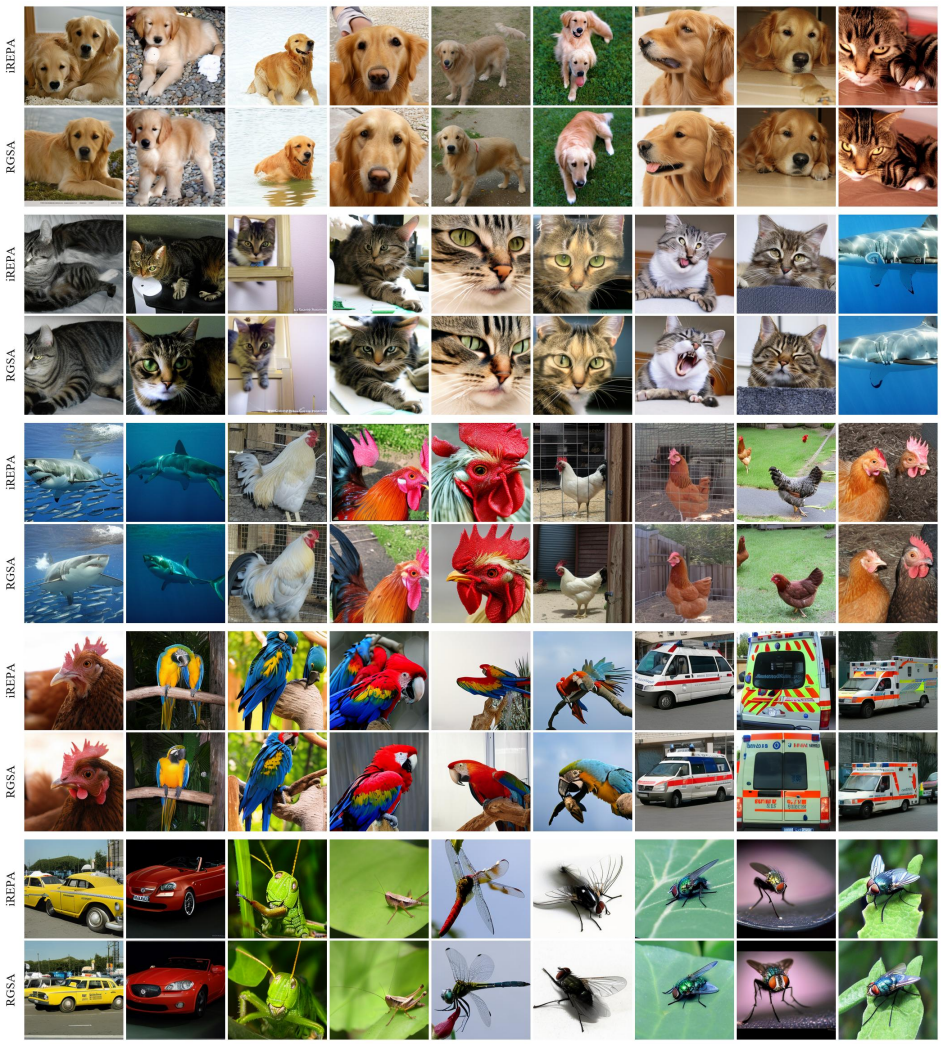



Figure 7: Additional paired generations from iREPA and RGSA under matched sampling conditions.

## F Robustness to Sampling Randomness

We assess sensitivity to sampling randomness using SiT-B/2 checkpoints trained on ImageNet-100 for 50K steps. For each method, we fix its trained checkpoint and generate 5,000 images for each of five sampling seeds, $\{0, 1, 2, 3, 4\}$. All runs use 250-step SDE sampling without classifier-free guidance and are evaluated against the same fixed reference set. We report the mean and sample standard deviation across sampling seeds.

Table 12: Repeated sampling evaluation using SiT-B/2 checkpoints trained on ImageNet-100 for 50K steps. Results are reported as mean $\pm$ sample standard deviation over five sampling seeds, with 5,000 generated images per seed and one fixed trained checkpoint per method.

| Method | FID $\downarrow$ | Recall $\uparrow$ |
|---|---|---|
| Uniform (iREPA) | $110.89 \pm 0.45$ | $0.4279 \pm 0.0042$ |
| Random | $110.42 \pm 0.33$ | $0.4273 \pm 0.0062$ |
| Shuffled | $112.56 \pm 0.74$ | $0.4297 \pm 0.0097$ |
| **RGSA** | $\mathbf{108.98 \pm 0.52}$ | $\mathbf{0.4493 \pm 0.0092}$ |

RGSA achieves lower FID and higher Recall than uniform iREPA, Random, and Shuffled at each of the five evaluated sampling seeds. Compared with uniform iREPA, RGSA reduces the mean FID by $1.91$ and increases the mean Recall by $0.0214$. Random slightly improves the mean FID but does not improve the mean Recall, whereas Shuffled worsens the mean FID. These results show that RGSA's joint improvement in FID and Recall persists across the evaluated sampling seeds under the fixed evaluation protocol.

This experiment measures the variability introduced by sampling randomness under fixed trained checkpoints and a fixed reference set. The results therefore indicate that the observed performance gains remain consistent across different sampling seeds.