NO CHANNELS LEFT BEHIND: UNDERSTANDING DIFFUSION ACCELERATION THROUGH RESIDUAL GATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Training acceleration methods for diffusion transformers, such as REPA and dispersive regularization, are commonly explained by additional supervision, spatial-structure priors, or improved feature geometry. We revisit this view and uncover a striking internal change: the AdaLN-Zero residual gate α, which directly controls how conditioned updates enter the residual stream, is substantially reorganized by these methods. This observation motivates our hypothesis that acceleration arises, at least in part, from reshaping the residual gating pathway rather than solely from the feature structures explicitly imposed by auxiliary objectives. Supporting this view, accelerated models are markedly more sensitive to perturbations of α, and their gate geometry reflects semantic relationships among classes. We further introduce the normalized participation ratio (nPR) to characterize the channel-wise organization of the gates and find that, across REPA teachers, projector architectures, and radial-objective targets, broader gate participation accompanies better generation quality. Guided by this insight, we propose a simple teacher-free radial objective that promotes broader gate participation through feature-norm regularization, without pairwise feature comparisons. On ImageNet 256×256 with SiT-B/2, our method improves FID over dispersive loss while increasing mean gate participation over timesteps and classes, further supporting the link between gate organization and training acceleration. Our results identify residual-gate reorganization as a measurable axis of diffusion training acceleration and suggest a simple direction for further acceleration.

1 INTRODUCTION

Recent advances have substantially accelerated diffusion transformer training by injecting additional auxiliary objectives through representation alignment (REPA) with frozen-teacher supervision (Yu et al., 2025), spatially structured representation alignment (Singh et al., 2026), and feature-geometry regularization (Wang & He, 2025). Despite their differences, these approaches are largely understood through the representational structures explicitly imposed by their objectives. Yet, *are these structures themselves responsible for faster training, or do they reshape a more fundamental mechanism within the transformer?*

We revisit this question through the conditioning pathway of AdaLN-Zero (Peebles & Xie, 2023). In diffusion transformers such as DiT (Peebles & Xie, 2023) and SiT (Ma et al., 2024), timestep and class embeddings generate modulation vectors for every transformer block. Among them, the residual gate $\boldsymbol{\alpha}$[1] plays a particularly direct role: it scales each channel of a branch output before residual addition, thus controlling how strongly each feature channel contributes to the residual stream across all token positions. This suggests a complementary view of diffusion acceleration: auxiliary objectives may not only reshape the features they supervise, but also reorganize *how those features are injected into the residual stream*. We therefore conjecture that residual gating may provide a common internal signature of accelerated diffusion training (Fig. 1).

To this end, we begin with two simple interventions: removing the AdaLN-Zero modulation pathway during training and perturbing the residual gate itself at test time. Gating matters even for the baseline (without acceleration), but strikingly, accelerated models are far more vulnerable to its disruption. This contrast suggests that acceleration does more than improve the features being learned:

[1] We refer to $\boldsymbol{\alpha}$ as a *gate*, as it modulates the embeddings through channel-wise multiplication.

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.

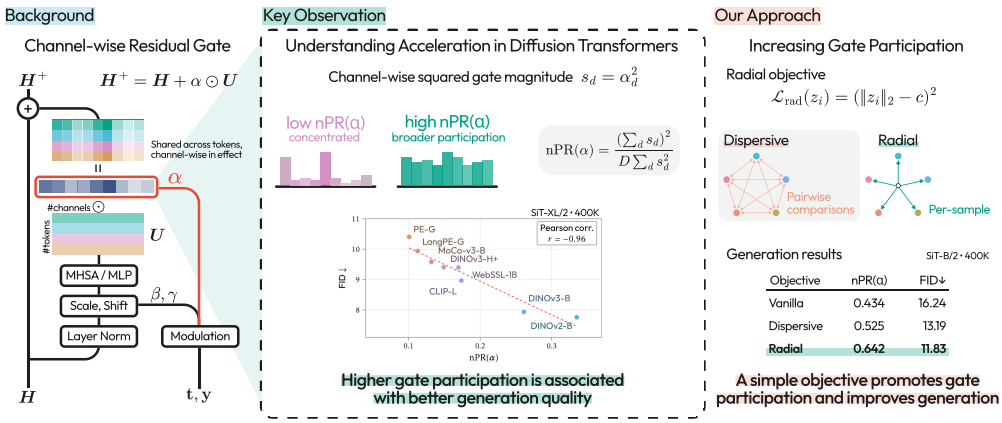


Figure 1: **Overview: from residual gate insights to improved diffusion training.** ***(Left)*** AdaLN-Zero gates scale residual-branch outputs channel-wise, with each gate shared across token positions. ***(Center)*** Our proposed *normalized participation ratio (nPR)* measures how evenly squared gate magnitudes are distributed across channels. Across REPA teachers, higher mean gate participation is associated with lower FID. ***(Right)*** Our proposed *Radial objective* penalizes per-sample feature norms without a pretrained teacher or pairwise feature comparisons.

it fundamentally changes how the model uses them through the gating pathway. Motivated by this observation, we first ask whether residual gates carry interpretable semantic structure. Using WordNet (Fellbaum, 1998), we find that distances between class-conditioned gates increase with semantic distance between classes, indicating that the gating pathway reflects the semantic organization of the generated categories. Next, we quantify gate participation using the *normalized participation ratio* (nPR), a scale-invariant measure of how broadly gate magnitude is distributed across channels that also guides our subsequent method design. Higher gate participation, averaged over timesteps and classes, tracks better generation quality across REPA teachers and, for a fixed teacher, across projector designs, identifying gate participation as a correlate of generation quality within each sweep.

Based on gate participation, we introduce the Radial objective, a simple teacher-free objective that acts only on hidden-feature norms. Unlike representation alignment, it requires no pretrained teacher guidance; unlike dispersive regularization, it requires no pairwise feature comparisons. Instead, the Radial objective penalizes deviations of each sample's flattened feature norm from a target. Despite acting only on feature norms, this objective increases residual-gate participation and improves generation quality. As a result, our method improves both Fréchet inception distance (FID; Heusel et al., 2017) and mean gate participation, supporting residual-gate participation as a candidate mechanism underlying diffusion training acceleration. Quantitatively, on ImageNet 256×256 (Deng et al., 2009) with SiT-B/2 trained for 400K steps, the Radial objective achieves an FID of 11.83, improving over Vanilla[2] (16.24) and Dispersive (13.19), while increasing the mean final-block MLP gate participation from 0.434 in Vanilla to 0.642.

Overall, our results point beyond the representations explicitly targeted by existing acceleration methods. These objectives also reorganize the residual gating pathway, whose channel participation tracks generation quality in controlled sweeps. We argue that this perspective identifies residual-gate organization as a previously overlooked axis of diffusion acceleration and provides a simple route to promote it without either a representation teacher or pairwise feature comparisons.

2 Background

Diffusion transformers process latent image tokens through transformer blocks containing a multihead self-attention (MHSA) branch and an MLP branch (Peebles & Xie, 2023). Each branch adds an update to the residual stream. We study the AdaLN-Zero conditioning used in these blocks.

Let $\boldsymbol{H} \in \mathbb{R}^{N\times D}$ denote N tokens with D channels. The conditioning pathway produces the modulation vectors $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma} \in \mathbb{R}^{D}$ and the output embedding $\boldsymbol{H}^{+}$ is sequentially computed as $\boldsymbol{H}^{+} = \boldsymbol{H} + \boldsymbol{\alpha} \odot \boldsymbol{U}$, where $\boldsymbol{U}$ can be either $\texttt{MHSA}(\widetilde{\boldsymbol{H}})$ or $\texttt{MLP}(\widetilde{\boldsymbol{H}})$, and $\widetilde{\boldsymbol{H}} = (\mathbf{1}+\boldsymbol{\gamma}) \odot \texttt{LayerNorm}(\boldsymbol{H}) + \boldsymbol{\beta}$.

[2]We refer to Vanilla as a baseline SiT (Ma et al., 2024) model.

Here, $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ specify the scale and shift of the normalized branch input; $\boldsymbol{\alpha}$ scales the branch output before residual addition. Accordingly, $\boldsymbol{\alpha}$ may be interpreted as a residual gating vector. The Hadamard product $\odot$ acts channel-wise, with each vector broadcast across tokens; MHSA and MLP branches have separate modulation parameters in each block. Note that $\boldsymbol{\alpha}$ is structurally distinct: Whereas $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ modulate the branch input and are mediated by the subsequent projection layers in MHSA or MLP, $\boldsymbol{\alpha}$ directly gates the branch output before residual addition. We thus conjecture that training acceleration is governed primarily by $\boldsymbol{\alpha}$.

We refer to $\boldsymbol{\alpha}$ as the *channel-wise residual gate*: α_d multiplies channel d at every token position. Its entries are signed scaling coefficients. The gate is obtained directly from the conditioning pathway, without an image forward pass. Because $\boldsymbol{\alpha}$ directly scales the residual branch output, it lets us examine how conditioning controls each channel's contribution to the residual stream. We first test how generation quality responds to changes in the gate's channel-wise magnitude distribution, then examine the structure of gates learned under accelerated training. Prior work has examined AdaLN initialization and parameter updates (Zhu et al., 2026) and the semantic structure of conditioning embeddings (Pham et al., 2026). How the residual gate values organize channel contributions under accelerated training remains underexplored.

Diffusion transformer acceleration and its mechanics. Recent auxiliary losses to accelerate diffusion transformer training fall into two broad categories: teacher-guided (REPA (Yu et al., 2025) and iREPA (Singh et al., 2026)) and teacher-free (dispersive loss (Wang & He, 2025)). REPA and iREPA accelerate training by aligning diffusion features with representations from a frozen vision foundation model through a learnable projector. Dispersive loss is a regularizer that encourages larger pairwise ℓ_2 distances between the hidden features of samples within a batch.

3 Understanding Acceleration in Diffusion Transformers

3.1 Motivation

What lies beyond existing representation-level supervision? Existing explanations for diffusion acceleration largely focus on the representations explicitly targeted by the auxiliary objective: REPA guides hidden features toward pretrained visual representations (Yu et al., 2025), while dispersive objectives reshape their geometry (Wang & He, 2025). We conjecture that these feature-level effects are not the full story. Instead, acceleration may critically depend on how such supervision reorganizes the AdaLN-Zero gating pathway (beyond prior initialization analyses (Zhu et al., 2026)), which directly controls the contribution of conditioned residual updates throughout the network.

Pilot study 1: How does gating affect generation quality during training? We perform a pilot study by training models with in-context conditioning (Peebles & Xie, 2023) in place of AdaLN-Zero, removing its conditional modulation pathway, including the residual gate. The FID improvements from both Dispersive and REPA become substantially smaller under in-context conditioning (Table 1). Relative to Vanilla, REPA's FID change shifts from -18.24 to -13.48, while Dispersive's shifts from -4.98 to $+1.15$. This comparison supports a role for the AdaLN-Zero modulation pathway in the gains from auxiliary objectives; Pilot 2 isolates the residual gate.

Table 1: **FID changes under different conditioning architectures**. Replacing AdaLN-Zero with in-context conditioning (*i.e.*, removing the AdaLN-Zero modulation, including residual gates) substantially weakens the acceleration gains. All models are SiT-B/2 trained for 100K steps; FID (↓) follows Appendix A.3, and Δ is the change from Vanilla with the same conditioning.

Conditioning architecture	Vanilla	REPA	Δ	Dispersive	Δ
AdaLN-Zero (baseline)	45.41	27.17	-18.24 (-40.2%)	40.43	-4.98 (-11.0%)
In-context	47.23	33.76	-13.48 (-28.5%)	48.39	$+1.15$ ($+2.4\%$)

Pilot study 2: Effect of gating at test time. After training, we perturb the gate values of frozen Vanilla, Dispersive, and REPA models during sampling to measure their sensitivity to changes in the learned gates. This intervention evaluates sensitivity without allowing the model weights to adapt to the modified gates. We use SiT-XL/2 checkpoints trained for 400K steps. At each sampling step, we redistribute the squared magnitudes of the MLP gate at the eighth layer (L8), where the REPA objec-

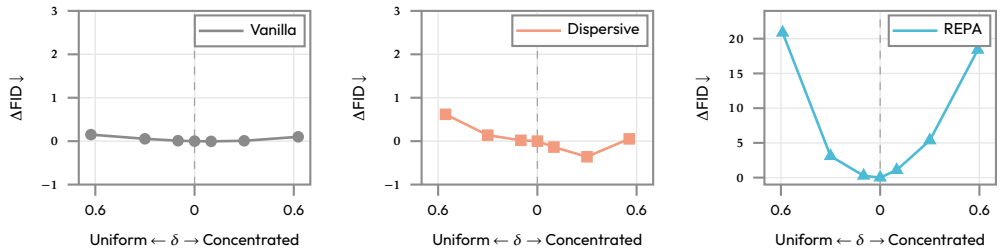


Figure 2: **Does gate redistribution in Vanilla, Dispersive, and REPA affect FID?** The center (*i.e.*, $x = 0$) is the unmodified baseline; moving left makes gate magnitudes more uniform, and moving right makes their distribution more concentrated. Tick labels give relative gate perturbation magnitude δ. All three panels show FID change from each model's own baseline.

tive is applied, making the distribution either more uniform or more concentrated while preserving the gate's signs and ℓ_2 norm. We match the relative perturbation magnitude $\delta = \|\boldsymbol{\alpha}^{\text{new}} - \boldsymbol{\alpha}\|_2 / \|\boldsymbol{\alpha}\|_2$ across these directions and models. At a matched perturbation magnitude of $\delta = 0.30$, Dispersive shows larger absolute FID changes than Vanilla, while REPA responds much more strongly (Fig. 2). Both studies show that generation quality is sensitive to $\boldsymbol{\alpha}$, most strongly in REPA (ΔFID up to $+5.356$ at $\delta = 0.30$). With frozen weights, this test measures sensitivity to departures from the learned gates, not a preference for uniform gates; Section 3.2.2 compares gates learned under different training. Since REPA/iREPA is primarily understood through representation alignment, this raises a question: *Could the gating pathway be an underlying piece of the acceleration?*

3.2 Empirical Insights with Residual Gating

Here, we investigate the role of the residual gate in accelerated diffusion-transformer training. Our pilot study revealed that generation quality in existing training acceleration methods is unexpectedly sensitive to the residual gate $\boldsymbol{\alpha}$, which suggests that the gating pathway may play a more central role than previously appreciated. This leads us to revisit REPA's acceleration from a different perspective. We study the structure of the learned gates to explore this possibility.

3.2.1 How Do Residual Gates Behave in Training Acceleration Methods?

Residual gates reflect semantic distance. We examine how class-conditioned gate distances respond to semantic distance, and whether this response differs across training objectives. The semantic distance is measured as WordNet distance (Fellbaum, 1998), which is the number of edges on a shortest path between class synsets. The partial WordNet noun hierarchy in Fig. 3 highlights the five-edge path from dog to tiger. We compare the class-conditioned MLP gates of Vanilla, REPA, and Dispersive; experimental settings are provided in Appendix B.3. We use $t = 0.5$ as a representative midpoint of the diffusion trajectory; Appendix B.4 reports the same analysis at $t = 0.25$ and

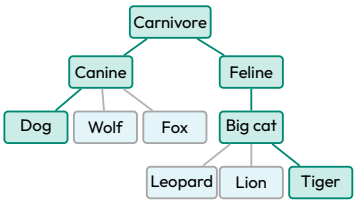


Figure 3: **Illustration of a WordNet hierarchy and path distance.** The highlighted shortest path from dog to tiger spans five hypernym edges.

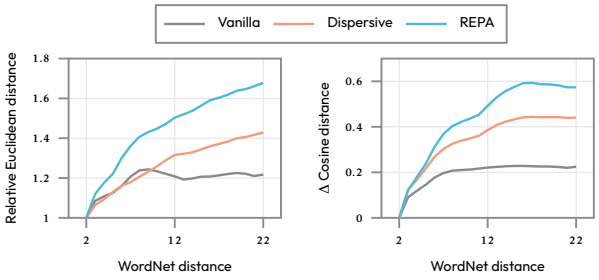


Figure 4: **Change in residual gates versus WordNet distance**. We compare how the gates change in Vanilla, REPA, and Dispersive models at $t = 0.5$. ***(Left)*** Euclidean distance between gate magnitudes $\boldsymbol{a}_i = |\boldsymbol{\alpha}_i|$, divided by its bin mean at WordNet distance 2. ***(Right)*** Cosine distance after centering $\boldsymbol{a}_i$ across classes, shifted by its bin mean at WordNet distance 2 to start at zero.

0.75. We compute distances for all 499,500 unordered class pairs and average them at each WordNet distance. Every curve attains its minimum bin mean at WordNet distance 2. Because overall gate scale differs across models (Fig. 5), we divide each Euclidean curve by this minimum; for the scale-invariant centered cosine distance, we subtract it. Gate distances generally increase with semantic distance in all three models, with the largest increase under REPA (Fig. 4). At WordNet distance 22, Euclidean distances reach $1.678\times$ their starting value for REPA, compared with $1.217\times$ for Vanilla and $1.428\times$ for Dispersive; centered cosine distances increase by 0.574, 0.225, and 0.441, respectively.

The hierarchy analysis shows that class-conditioned gates reflect semantic hierarchy, as the distance between gate vectors increases with semantic distance between classes. Since a larger distance between two gate vectors indicates greater separation in gate space, this trend suggests that REPA induces the strongest semantic separation among class-conditioned gates. Together, these observations motivate quantifying the channel-wise magnitude distributions of these gates and examining their relationship with generation quality across training configurations.

3.2.2 Defining Gate Participation and Its Impact

Defining gate participation metric. We quantify gate participation with the participation ratio (PR), which measures how evenly the squared gate magnitudes are distributed across channels. For a nonzero gate vector with squared magnitudes $s_d = \alpha_d^2$, the PR and its normalized form (nPR)[3] are defined as follows:

$$\mathrm{PR}(\boldsymbol{\alpha}) = \frac{\left(\sum_{d=1}^{D} s_d\right)^2}{\sum_{d=1}^{D} s_d^2}, \quad \mathrm{nPR}(\boldsymbol{\alpha}) = \frac{\mathrm{PR}(\boldsymbol{\alpha})}{D}, \quad \boldsymbol{\alpha} \in \mathbb{R}^D. \tag{1}$$

We refer to $\mathrm{nPR}(\boldsymbol{\alpha})$ as gate participation. Higher $\mathrm{nPR}(\boldsymbol{\alpha})$ indicates that squared gate magnitude is spread more evenly across channels; lower $\mathrm{nPR}(\boldsymbol{\alpha})$ indicates concentration in fewer channels. Equal nonzero magnitudes on m of the D channels give $\mathrm{nPR}(\boldsymbol{\alpha}) = m/D$, with a range of $[1/D, 1]$ for nonzero gates. The statistic is invariant to uniform nonzero rescaling. It measures the allocation of gate magnitude across channels, rather than the overall gate scale of the residual update.

We then inspect the learned MLP gate values in Vanilla, REPA, and Dispersive. Fig. 5 pools the signed gate entries across all 1,000 classes at $t = 0.5$, without rescaling them. Their overall scales differ, with standard deviations of 0.049 for Vanilla, 0.285 for Dispersive, and 0.793 for REPA.

Gate participation correlates with generation quality. We next examine whether mean gate participation over timesteps and classes is actually correlated with generation quality. We vary the representation-alignment target to examine the relationship across training configurations. A trainable projector maps the diffusion features to the teacher representation, with alignment applied at corresponding tokens. We conduct class-conditional ImageNet 256×256 generation and report FID.

[3] Pham et al. (2026) also employ nPR to analyze the semantic structure of conditioning embeddings. We instead measure it on the residual gates and compare it across training objectives and against generation quality.

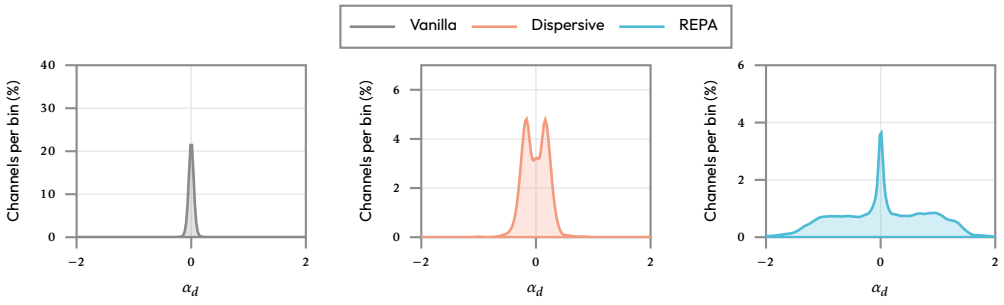
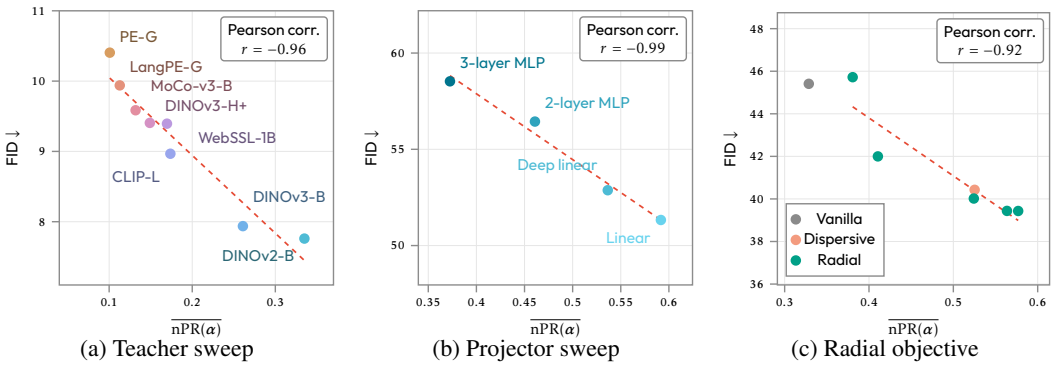


Figure 5: Histograms of signed learned MLP gate entries α_d for Vanilla, REPA, and Dispersive, using the same models as Fig. 4. At $t = 0.5$, we pool all 1,000 classes and 1,152 channels without normalization, using common width-0.025 bins, lightly smoothed with a Gaussian kernel ($\sigma = 1$ bin). All three panels show $[-2, 2]$ with independently scaled linear y-axes. Bin percentages use all entries, including those outside the displayed range.



(a) Teacher sweep (b) Projector sweep (c) Radial objective

Figure 6: **Gate participation and generation quality across teacher, projector, and radial-target sweeps**. *(a)* REPA with eight teachers. *(b)* Projector architectures under a fixed teacher. *(c)* Radial-target sweep at 100K steps, with Vanilla and Dispersive as reference points. All three horizontal axes show $\overline{\mathrm{nPR}(\boldsymbol{\alpha})}$ over 101 uniformly spaced model timesteps in $[0, 1]$ and 1,000 classes. Fits and correlations use the sweep points only, excluding the reference points in *(c)*.

Throughout, we measure gate participation at the block where the auxiliary objective is applied: L8 for SiT-XL/2 REPA, L4 for SiT-B/2 REPA, and L12 for Dispersive and the Radial objective, with Vanilla references measured at the same block. We measure MLP gates at L8 in the released REPA checkpoints (Yu et al., 2025) corresponding to eight teacher encoders. For each checkpoint, we compute $\mathrm{nPR}(\boldsymbol{\alpha}(t, y))$ separately for 1,000 classes and 101 uniformly spaced timesteps, then average the 101,000 values. Across eight REPA teachers, mean MLP gate participation at L8 is strongly negatively correlated with FID ($r = -0.962$; Fig. 6a). The teacher sweep reveals an association between gate participation and generation quality across alignment targets.

Gate participation can track quality with a fixed teacher. We now verify that the association between mean gate participation and generation quality also holds when the teacher is fixed and the projector changes; fixing the teacher removes teacher quality as a common cause of both gate participation and FID. iREPA (Singh et al., 2026) replaces REPA's three-layer MLP projector with a 3×3 convolution, providing a precedent for varying the projection head. Table 2 compares four projector architectures under fixed teacher, target preprocessing, alignment weight, and training budget.

Table 2: **Projector components and results.** We use SiT-B/2 trained for 50K steps. Check marks and crosses indicate present and absent components; $\boldsymbol{W}_i$ denotes affine slots. $\overline{\mathrm{nPR}(\boldsymbol{\alpha})}$ is computed for MLP gates at the fourth layer (L4) over 101 uniformly spaced model timesteps in $[0, 1]$ and 1,000 classes. All four use the same teacher targets and alignment settings (Appendix B.6).

Projector	$\boldsymbol{W}_1$	SiLU	$\boldsymbol{W}_2$	SiLU	$\boldsymbol{W}_3$	$\overline{\mathrm{nPR}(\boldsymbol{\alpha})}$	FID ↓
3-layer MLP	✓	✓	✓	✓	✓	0.373	58.53
2-layer MLP	✓	✓	✗	✗	✓	0.461	56.44
Deep linear	✓	✗	✓	✗	✓	0.536	52.87
Linear (1×1 conv.)	✓	✗	✗	✗	✗	0.592	**51.33**

Reducing the MLP from three affine layers to two and then to one increases mean gate participation from 0.373 to 0.461 and 0.592, while lowering FID from 58.53 to 56.44 and 51.33. Removing the nonlinearities while retaining three affine layers gives FID 52.87. Across the four variants, L4 MLP gate participation averaged over timesteps and classes correlates with FID at $r = -0.991$ (Fig. 6b). Together with the teacher sweep, these results show that the participation-quality association appears both when the alignment target changes and when the projector changes under a fixed teacher. This raises the question of whether encouraging higher gate participation improves generation quality.

4 Increasing Gate Participation for Better Generation Quality

This section studies whether encouraging higher gate participation during training improves generation quality. We introduce the Radial objective, which acts on hidden-feature norms, and examine when its local gradient contribution increases gate participation.

The radial objective. Based on our empirical studies above, we propose a simple training acceleration method, coined Radial objective, which adds a per-sample feature-norm penalty to the velocity prediction loss (Ma et al., 2024). Our insight is that the embedding right before the residual addition directly interacts with the residual gate $\boldsymbol{\alpha}$, which would likely increase gate participation (*i.e.*, $\mathrm{nPR}(\boldsymbol{\alpha})$) and ultimately accelerate training. Given the output $\boldsymbol{Z}_i \in \mathbb{R}^{N\times D}$ of the final block, define $\boldsymbol{z}_i = \mathrm{flatten}(\boldsymbol{Z}_i)$ and optimize the following objective:

$$\mathcal{L} = \mathcal{L}_{\text{velocity}} + \lambda\, \mathbb{E}_i\big[(\|\boldsymbol{z}_i\|_2 - c)^2\big], \tag{2}$$

where the norm is over all tokens and channels of one sample, and c denotes a flattened feature-norm target; $\lambda > 0$ weights the auxiliary penalty. We sweep the radial target over $c \approx k\sqrt{ND}$ with $k \in \{2, 4, 8, 10, 14\}$ and report the results for each target; Appendix C.1 gives the exact values. We use the target with the lowest observed FID ($k = 10$) for the performance comparison. This objective requires neither a teacher nor distances between different samples.

Promoting gate participation with our radial objective. At the current model state, fix a condition (t, y) and analyze its gate $\boldsymbol{\alpha} = \boldsymbol{\alpha}_\theta(t, y)$. For an input at this condition, the final-block output is $\boldsymbol{Z} = \boldsymbol{H} + \boldsymbol{\alpha} \odot \boldsymbol{U} = \boldsymbol{H} + \boldsymbol{U}\,\mathrm{diag}(\boldsymbol{\alpha})$. Below, $\mathbb{E}[\cdot \mid t, y]$ averages over images and noises at the same (t, y); it does not require an actual minibatch to repeat (t, y). To isolate gate-weighted channel norms, we drop the skip–residual cross term from $\|\boldsymbol{H} + \boldsymbol{U}\,\mathrm{diag}(\boldsymbol{\alpha})\|_{\mathrm{F}}^2$ and analyze the norm

$$\widetilde{r}_{\boldsymbol{\alpha}} = \left(\|\boldsymbol{H}\|_{\mathrm{F}}^2 + \sum_{d=1}^{D} \alpha_d^2 \|\boldsymbol{U}_{:,d}\|_2^2\right)^{1/2}, \tag{3}$$

where $\|\boldsymbol{U}_{:,d}\|_2^2$ is the squared pre-gating norm of channel d, summed over tokens. We denote the radial objective evaluated with this norm, $\mathbb{E}[(\widetilde{r}_{\boldsymbol{\alpha}} - c)^2 \mid t, y]$, by $\widetilde{\mathcal{L}}_{\mathrm{rad}}$. We assume *channel-symmetric* pre-gating norm patterns within this conditional analysis, assigning no a priori preference to any residual channel. This assumption puts channels on equal footing before gating (Appendix D).

Proposition 1 (Gate participation under channel symmetry)**.** Under channel symmetry and the regularity conditions in Appendix D, a sufficiently small conditional gate update $\boldsymbol{\alpha}^+ = \boldsymbol{\alpha} - \eta \nabla_{\boldsymbol{\alpha}} \widetilde{\mathcal{L}}_{\mathrm{rad}}(\boldsymbol{\alpha})$, with $\eta > 0$, satisfies

$$\mathrm{nPR}(\boldsymbol{\alpha}^+) \geq \mathrm{nPR}(\boldsymbol{\alpha}). \tag{4}$$

The mechanism follows from the relative gate gradient,

$$-\frac{1}{\alpha_d}\frac{\partial \widetilde{\mathcal{L}}_{\mathrm{rad}}}{\partial \alpha_d} = 2\,\mathbb{E}\left[\left(\frac{c}{\widetilde{r}_{\boldsymbol{\alpha}}} - 1\right)\|\boldsymbol{U}_{:,d}\|_2^2 \,\middle|\, t, y\right], \tag{5}$$

which equals $\Delta\alpha_d/(\eta\alpha_d)$ with $\Delta\alpha_d = \alpha_d^+ - \alpha_d$. $c/\widetilde{r}_{\boldsymbol{\alpha}} - 1$ is specific to each input embedding and shared by all channels of that input; channels differ only through $\|\boldsymbol{U}_{:,d}\|_2^2$ and α_d. Consider two otherwise matched input norm patterns at the same (t, y), differing only by exchanging two pre-gating channel norms. If the larger channel norm is paired with the larger gate, the whole-input norm becomes larger, reducing the shared radial coefficient; swapping the channel norms gives the opposite effect. Under channel symmetry, the two patterns are weighted equally. Averaging over such realizations gives weaker gates no smaller fractional magnitude updates than stronger ones:

$$|\alpha_i| > |\alpha_j| \quad \Longrightarrow \quad \frac{\Delta\alpha_i}{\alpha_i} \leq \frac{\Delta\alpha_j}{\alpha_j}. \tag{6}$$

For sufficiently small, sign-preserving steps, weaker gates can thus grow faster or shrink more slowly, reducing gate concentration. Under the same analysis, positive and negative quadratic norm objectives yield uniform relative gate rescaling, whereas the radial objective introduces input-dependent inverse-norm weighting. This result concerns the conditional gate-space update in the cross-term-excluded analysis, rather than the full parameter update during training.

Higher participation and lower FID under radial training. We search over target norms using SiT-B/2 trained for 100K steps. We calibrate λ over the first five training iterations so that the weighted auxiliary gradient norm is approximately 10% of the velocity-loss gradient norm. We measure MLP gates at the twelfth layer (L12), on which the dispersive loss and the radial objective act.

Table 3: **Generation quality at 400K training steps**, SiT-B/2 at 256×256. We report the results of Vanilla, Dispersive, and the proposed Radial objective, with REPA as a teacher-guided reference. The best teacher-free results are in bold.

Method	FID ↓	sFID ↓	IS ↑	Precision ↑	Recall ↑
teacher-free					
Vanilla (Ma et al., 2024)	16.24	5.43	86.67	0.670	0.557
Dispersive (Wang & He, 2025)	13.19	**5.25**	98.19	0.696	0.553
Radial objective (ours)	**11.83**	5.27	**105.93**	**0.714**	**0.559**
teacher-guided					
REPA (Yu et al., 2025)	7.71	5.40	141.54	0.735	0.577

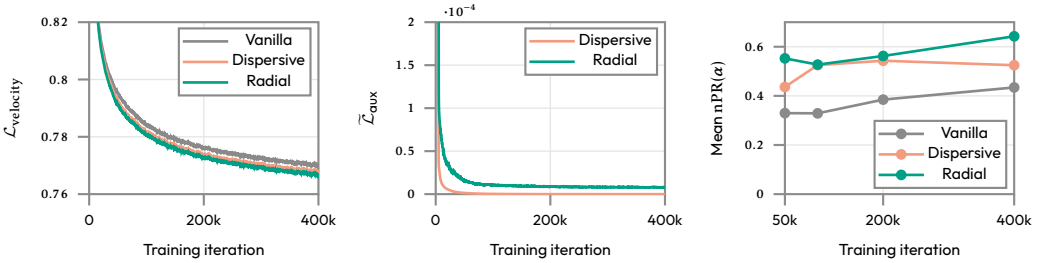


Figure 7: **Training dynamics over 400K steps.** ***(Left)*** Training MSE. ***(Center)*** Normalized auxiliary losses. ***(Right)*** Mean $\mathrm{nPR}(\boldsymbol{\alpha})$ over 101 uniformly spaced timesteps in $[0, 1]$ and 1,000 classes.

The target is selected by the lowest observed FID; Appendix C.1 reports the search values and sensitivity. Across the five settings in Fig. 6c, L12 MLP gate participation averaged over timesteps and classes is negatively correlated with FID ($r = -0.916$). These measurements test whether radial training yields higher participation together with improved generation quality. FID decreases from 45.72 at the lowest target to 39.43 at the selected target ($k = 10$), while mean gate participation increases from 0.381 to 0.577.

Generation quality with and without a teacher. Table 3 compares SiT-B/2 models at 400K training steps. We train Radial objective with the selected target $c = 10\sqrt{ND}$. Its FID of 11.83 improves on Vanilla (16.24) and Dispersive (13.19) while requiring no teacher. Its mean L12 MLP gate participation of 0.642 is likewise the highest, compared with 0.434 for Vanilla and 0.525 for Dispersive.

Gate participation during training. We track gate participation at 50K, 100K, 200K, and 400K training steps for Vanilla, Dispersive, and Radial. At each checkpoint, we evaluate L12 MLP gate participation averaged over timesteps and classes using EMA weights; Fig. 7 (right) connects these measurements without smoothing. Radial maintains higher gate participation than Vanilla at all four checkpoints. Dispersive also exceeds Vanilla, with participation peaking at 200K.

Training-loss visualization. We also report the recorded velocity-prediction MSE and auxiliary training losses. Here, $\mathcal{L}_{\text{aux}}$ denotes the additional regularization term: the dispersive loss for Dispersive and the feature-norm penalty in Eq. (2) for Radial, each measured before weighting by its coefficient. Vanilla has no auxiliary term. For the auxiliary curves in Fig. 7 (center), we plot the normalized gap $\widetilde{\mathcal{L}}_{\text{aux}} = (L_t - L_{\min})/(L_0 - L_{\min})$, where L_t is the raw auxiliary loss at training step t, L_0 is its step-zero value, and $L_{\min} = -\ln 32$ for Dispersive and 0 for Radial.

5 Discussion and Limitations

Interpreting iREPA and spatial structure. iREPA motivates its convolutional projector and spatial normalization as ways to improve the transfer of spatial structure from teacher representations (Singh et al., 2026). Its spatial diagnostics, including LDS, summarize how similarity between patch tokens varies with their spatial separation. Our fixed-teacher projector sweep in Section 3.2.2 helps distinguish properties that change together when replacing an MLP with a convolution: nonlinear depth, channel transformation, and spatial mixing. At 50K steps, the 1×1 projector reaches FID 51.33, compared with 51.68 for an additional 3×3 convolutional projector trained under the same teacher normalization and alignment settings. The comparable quality of these two heads shows that

neighboring-token mixing inside the projector is not required to attain this performance in the tested configuration. The lower FID of the deep linear head relative to the nonlinear MLP also motivates examining the effect of projector nonlinearities.

Association and intervention. The teacher and projector sweeps show a recurring relationship between gate participation and FID within their respective experimental settings. The post-hoc pilot establishes sensitivity to gate modification, but its uniformizing interventions do not consistently improve generation quality. The WordNet analysis shows class-related structure in the gates; it does not establish a lower bound on the participation needed to represent a semantic hierarchy. The Radial objective provides a simple teacher-free training objective that yields higher participation and lower FID in the tested target sweep. It acts on feature norms and affects the gates indirectly, alongside other model parameters. The local gradient proposition specifies conditions for a positive radial contribution (Appendix D.1); it does not guarantee participation growth under the full optimizer or establish that such growth causes the FID improvement.

6 Related Work

Representation alignment in diffusion transformers. REPA supervises intermediate diffusion features with pretrained visual representations, motivated by the difficulty of learning useful representations from the generative objective alone (Yu et al., 2025). Subsequent work examines the alignment target and projector in iREPA (Singh et al., 2026), and the timing and scope of alignment in HASTE, which combines feature and attention-map supervision before terminating alignment during training (Wang et al., 2025). Our analysis concerns the internal modulation induced by these training choices, focusing on residual-gate participation and geometry.

Spatial structure in diffusion representations. iREPA compares global semantic performance with spatial self-similarity in teacher representations and motivates changes intended to improve the transfer of spatial information (Singh et al., 2026). HASTE also incorporates spatial relationships through attention-map supervision (Wang et al., 2025). Our projector experiments instead examine the role of projection architecture within the tested alignment configurations, without making a broader claim about the necessity of spatial information in diffusion models.

Adaptive modulation and gating in DiTs. DiT uses adaptive normalization and residual scaling through AdaLN-Zero (Peebles & Xie, 2023); SiT adopts this transformer backbone for interpolant-based generative modeling (Ma et al., 2024). Related analyses examine modulation initialization (Zhu et al., 2026) and the concentration of semantic information in conditional embeddings (Pham et al., 2026). We focus specifically on the learned branch-specific residual gate $\boldsymbol{\alpha}$, its allocation across channels, and its geometry across training objectives.

Representation geometry and effective dimensionality. Sparsity measures characterize the concentration of coefficients (Hurley & Rickard, 2008), while intrinsic-dimension estimates have been used to assess self-supervised representations (Mordacq et al., 2026). Dispersive training modifies representation geometry through an auxiliary separation objective (Wang & He, 2025). We apply participation ratio to squared gate coordinates in the model's channel basis. This measures the allocation of gate magnitude across channels, rather than the spectral rank or intrinsic dimension of hidden representations, and serves here as a diagnostic associated with generation quality.

7 Conclusion

We examined diffusion transformer acceleration through class-conditioned residual gates. Accelerated models depend more strongly on these gates: removing the AdaLN-Zero modulation shrinks their gains, and perturbing $\boldsymbol{\alpha}$ degrades REPA the most. The gates also reflect the semantic hierarchy among classes, most strongly under REPA and their gate participation tracks FID across REPA teachers and, for a fixed teacher, across projectors. Building on this, the teacher-free Radial objective raised gate participation and improved FID over Dispersive without pairwise feature comparisons. Because gate participation is computed from the conditioning pathway alone, without an image forward pass, it offers a lightweight quantity to monitor in accelerated diffusion training.

AI USE STATEMENT

In this work, we used generative AI tools to write small pieces of code for implementing our methods, to assist in writing proofs, and to formulate mathematical claims. We have not used generative AI tools to develop theoretical models or conceptual frameworks, to provide critical ingredients for proofs, to propose or refine hypotheses, to design or give feedback on the research methodology or experiments, to interpret results, or to translate text. Synthetic data generation, dataset cleaning and reformatting, and qualitative or thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to search for related work, to edit the paper for readability, and to check its formatting. We emphasize that generative AI did not originate the core methodology or generate the full experimental codebase. We have reviewed all AI-assisted work: the authors derived and verified all mathematical claims and proofs themselves, AI-assisted code was validated through sanity checks and by reproducing published baseline results, and every work found through AI-assisted search was checked against the original paper. We take responsibility for the final content of this work, including text, claims, code, analyses, and results produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide all implementation details and hyperparameters in Appendices A to C.

REFERENCES

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Daniel Li, Piotr Dollár, and Christoph Feichtenhofer. Perception Encoder: The best visual embeddings are not at the output of the network. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

Xinlei Chen, Saining Xie, and Kaiming He. An Empirical Study of Training Self-Supervised Vision Transformers. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

Prafulla Dhariwal and Alex Nichol. Diffusion Models Beat GANs on Image Synthesis. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

J.R. Dormand and P.J. Prince. A family of embedded Runge-Kutta formulae. *Journal of Computational and Applied Mathematics*, 1980.

David Fan, Shengbang Tong, Jiachen Zhu, Koustuv Sinha, Zhuang Liu, Xinlei Chen, Michael Rabbat, Nicolas Ballas, Yann LeCun, Amir Bar, and Saining Xie. Scaling Language-Free Visual Representation Learning. In *IEEE International Conference on Computer Vision (ICCV)*, 2025.

Christiane Fellbaum (ed.). *WordNet: An Electronic Lexical Database*. MIT Press, 1998. ISBN 978-0-262-27255-1.

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. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

Jonathan Ho and Tim Salimans. Classifier-Free Diffusion Guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Niall P. Hurley and Scott T. Rickard. Comparing Measures of Sparsity. *IEEE Transactions on Information Theory*, 2008.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved Precision and Recall Metric for Assessing Generative Models. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

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 (ECCV)*, 2024.

Julie Mordacq, Vicky Kalogeiton, and Steve Oudot. IdEst: Assessing Self-Supervised Learning Representations via Intrinsic Dimension. In *International Conference on Machine Learning (ICML)*, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research (TMLR)*, 2024.

William Peebles and Saining Xie. Scalable Diffusion Models with Transformers. In *IEEE International Conference on Computer Vision (ICCV)*, 2023.

Trung X. Pham, Kang Zhang, Ji Woo Hong, and Chang D. Yoo. A Hidden Semantic Bottleneck in Conditional Embeddings of Diffusion Transformers. In *International Conference on Learning Representations (ICLR)*, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved Techniques for Training GANs. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2016.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. 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 *International Conference on Learning Representations (ICLR)*, 2026.

Runqian Wang and Kaiming He. Diffuse and Disperse: Image Generation with Representation Regularization. *arXiv preprint arXiv:2506.09027*, 2025.

Ziqiao Wang, Wangbo Zhao, Yuhao Zhou, Zekai Li, Zhiyuan Liang, Mingjia Shi, Xuanlei Zhao, Pengfei Zhou, Kaipeng Zhang, Zhangyang Wang, Kai Wang, and Yang You. REPA Works Until It Doesn't: Early-Stopped, Holistic Alignment Supercharges Diffusion Training. In *Conference on Neural Information Processing Systems (NeurIPS)*, 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 (ICLR)*, 2025.

Jie Zhu, Mingyu Ding, Boqiang Duan, Leye Wang, and Jingdong Wang. Unveiling the Secret of AdaLN-Zero in Diffusion Transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2026.

APPENDIX

A COMMON SETUP

A.1 MODELS AND TRAINING

All experiments use class-conditional ImageNet (Deng et al., 2009) generation at 256×256 resolution with SiT (Ma et al., 2024) and AdaLN-Zero, except the in-context conditioning ablation of Pilot study 1 (Appendix B.1). The teacher sweep, gate intervention, hierarchy analysis, and gate histograms use SiT-XL/2, with 28 blocks, hidden width 1,152, and 16 attention heads. The projector and Radial objective experiments use SiT-B/2, with 12 blocks, width 768, and 12 heads. Both architectures use 2×2 latent patches and an MLP expansion ratio of four. Layer numbers are one-based: L1 denotes the first transformer block.

Local SiT-B/2 training. Unless stated otherwise, our SiT-B/2 runs use seed 42, global batch size 256, bfloat16 mixed precision, and AdamW (Loshchilov & Hutter, 2019) with a constant learning rate of 10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and zero weight decay. The label-drop probability is 0.1 and the EMA decay is 0.9999. Our SiT-B/2 Dispersive runs use the best-performing PyTorch configuration of the official implementation (Wang & He, 2025): the InfoNCE-L2 dispersive loss on the final-block output (L12) with coefficient $\lambda = 0.25$ and temperature $\tau = 1$, computed within each per-GPU batch of 32 on eight GPUs. The velocity objective uses a linear interpolation path. Training uses cached ImageNet latents from the ft-MSE VAE, sampled from the stored posterior.

A.2 GATE PARTICIPATION MEASUREMENT

Unless a fixed timestep is stated, we evaluate the conditioning pathway on 101 evenly spaced timesteps in $[0, 1]$, including both endpoints, and all 1,000 ImageNet class labels. For each MLP gate vector $\boldsymbol{\alpha}(t, y) \in \mathbb{R}^D$, we compute $\mathrm{nPR}(\boldsymbol{\alpha}) = (\sum_d \alpha_d^2)^2 / (D \sum_d \alpha_d^4)$ and take the arithmetic mean over the 101,000 conditions, which gives $\overline{\mathrm{nPR}(\boldsymbol{\alpha})}$. We average participation values rather than gate vectors, and do not concatenate MHSA and MLP gates. Gates are computed in FP32 and statistics in FP64. We use the convention $\boldsymbol{x}_t = (1 - t)\boldsymbol{x}_{\mathrm{data}} + t\boldsymbol{x}_{\mathrm{noise}}$ and convert it to each checkpoint's input convention. In the scatter plots, regression lines are ordinary least-squares fits of FID to $\overline{\mathrm{nPR}(\boldsymbol{\alpha})}$, and r is the Pearson correlation across the plotted runs.

A.3 FID EVALUATION

Local FIDs use the ADM evaluator (Dhariwal & Nichol, 2021) and the ImageNet-256 reference statistics, with 50,000 generated images, EMA model weights, the `sd-vae-ft-ema` decoder, CFG 1.5, and sampling seed 42. The sampler is adaptive Dormand–Prince (`dopri5`) (Dormand & Prince, 1980), with absolute tolerance 10^{-6} and relative tolerance 10^{-3}. Besides FID (Heusel et al., 2017), the evaluator reports sFID (Dhariwal & Nichol, 2021), inception score (IS) (Salimans et al., 2016), and precision and recall (Kynkäänniemi et al., 2019). The gate intervention pilot and the teacher sweep use different protocols, described in Appendices B.2 and B.5.

B DETAILS OF THE GATE ANALYSIS

B.1 IN-CONTEXT CONDITIONING PILOT

Table 1 compares six SiT-B/2 runs trained for 100K steps with the setup of Appendix A.1 and evaluated at their 100K EMA checkpoints with the local FID protocol of Appendix A.3. Within each conditioning architecture, the three runs differ only in the auxiliary objective. REPA aligns L4 with DINOv2-B through a three-layer MLP projector with coefficient 0.5, and Dispersive uses coefficient 0.25 at the final block (L12).

In-context conditioning. Following DiT (Peebles & Xie, 2023), we replace AdaLN-Zero with two conditioning tokens, a timestep embedding and a class embedding, prepended to the patch tokens. The blocks are standard pre-LN transformer blocks with affine LayerNorm and no scale, shift, or

gate. The final layer is an affine LayerNorm followed by a linear projection without modulation, and the two conditioning tokens are discarded before unpatchifying. Patch tokens use fixed 2D sine-cosine positional embeddings, and the conditioning tokens receive zero positional embeddings. Auxiliary losses are computed on the patch tokens only.

B.2 Gate Intervention Pilot

The pilot uses SiT-XL/2 EMA checkpoints at 400K steps for Vanilla, Dispersive, and REPA with DINOv2-B (Oquab et al., 2024); the REPA checkpoint is the DINOv2-B checkpoint of the teacher sweep (Appendix B.5). To compare Dispersive and REPA on the same block, the Dispersive checkpoint applies the dispersive loss at L8, where these REPA checkpoints apply alignment, with coefficient 0.5 and $\tau = 1$. At every sampling step, we modify only the realized L8 MLP gate. Writing $p_d = \alpha_d^2/\|\boldsymbol{\alpha}\|_2^2$, the uniform direction uses $p'_d = (1-\epsilon)p_d + \epsilon/D$, while the concentrated direction uses $p'_d = p_d^\kappa / \sum_j p_j^\kappa$ with $\kappa > 1$. We reconstruct $\alpha'_d = \mathrm{sign}(\alpha_d)\|\boldsymbol{\alpha}\|_2\sqrt{p'_d}$ and solve for ϵ or κ separately for each condition to match the target relative perturbation δ. This preserves signs, gate norm, and channel ordering. The dose grid contains $\delta = 0.10, 0.30$, and a model-specific near-maximal feasible dose (0.627 for Vanilla, 0.555 for Dispersive, and 0.590 for REPA).

Each evaluation uses 50,000 generated images, CFG 1.0 (Ho & Salimans, 2022), and 50-step Euler sampling, with sampling seed 42 shared between interventions and their unmodified baselines. Vanilla and Dispersive use the SiT sampling convention and ft-EMA decoder (Rombach et al., 2022), while the REPA checkpoint uses the REPA convention and ft-MSE decoder. Because the pipelines differ, we report FID changes relative to each model's own baseline.

B.3 Gate Histograms and WordNet Distances

The channel histograms and hierarchy analysis use the L8 MLP gates of the SiT-XL/2 EMA checkpoints at 400K steps for Vanilla, Dispersive, and REPA (DINOv2-B), evaluated at $t = 0.5$ for all 1,000 classes.

Gate histograms. We pool the signed gate entries over all $1{,}000 \cdot 1{,}152$ coordinates without centering, squaring, or normalization. We use width-0.025 bins spanning $[-8, 8]$, which cover every entry, and display $[-2, 2]$. Heights are percentages of all coordinates per bin, lightly smoothed with a Gaussian kernel of standard deviation one bin.

WordNet distances. For every class, we take the absolute value of the gate and center these magnitudes across the 1,000 classes separately for each channel. We use WordNet 3.0 shortest hypernym-path lengths over all $\binom{1000}{2} = 499{,}500$ unordered class pairs. The Euclidean distance is computed between gate magnitudes, which class centering does not change, and the cosine distance between centered magnitudes. We average distances over pairs within each WordNet distance bin, retaining bins with at least 500 pairs (distances 2–22). The minimum bin mean of every curve occurs at WordNet distance 2. We divide each Euclidean curve by this minimum, which removes the difference in overall gate scale between models, and subtract it from each centered cosine curve, which is already scale-invariant.

B.4 Gate Structure at Additional Timesteps

The WordNet-distance comparison and the gate histograms in the main text use $t = 0.5$, the mean of the uniform timestep distribution used in training. For the same checkpoints and L8 MLP gates, we additionally evaluate $t = 0.25$ and 0.75. Fig. A shows the same qualitative relationship at both timesteps: gate distances increase with WordNet distance, with REPA showing the largest separation across semantic distances. Fig. B shows that the differences in gate scale between the three models persist at both timesteps.

B.5 Teacher Sweep

We use eight REPA SiT-XL/2 EMA checkpoints at 400K steps released by Singh et al. (2026) on Hugging Face (`REPA-E/iREPA-collections`), with CLIP-L (Radford et al., 2021), DINOv2-B, DINOv3-B and DINOv3-H+ (Siméoni et al., 2025), LangPE-G and PE-G (Bolya et al., 2025),

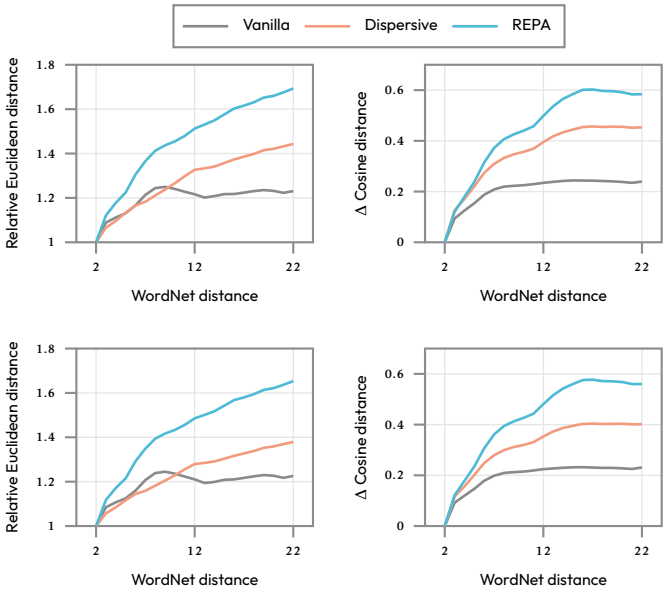


Figure A: WordNet-distance analysis at additional timesteps, using the same checkpoints and distance definitions as Fig. 4. Top row: $t = 0.25$; bottom row: $t = 0.75$. Left: Euclidean distance of gate magnitudes. Right: cosine distance of class-centered gate magnitudes. Euclidean curves are divided by, and cosine curves shifted by, their own bin mean at WordNet distance 2, with the same retained WordNet bins as in the main text.

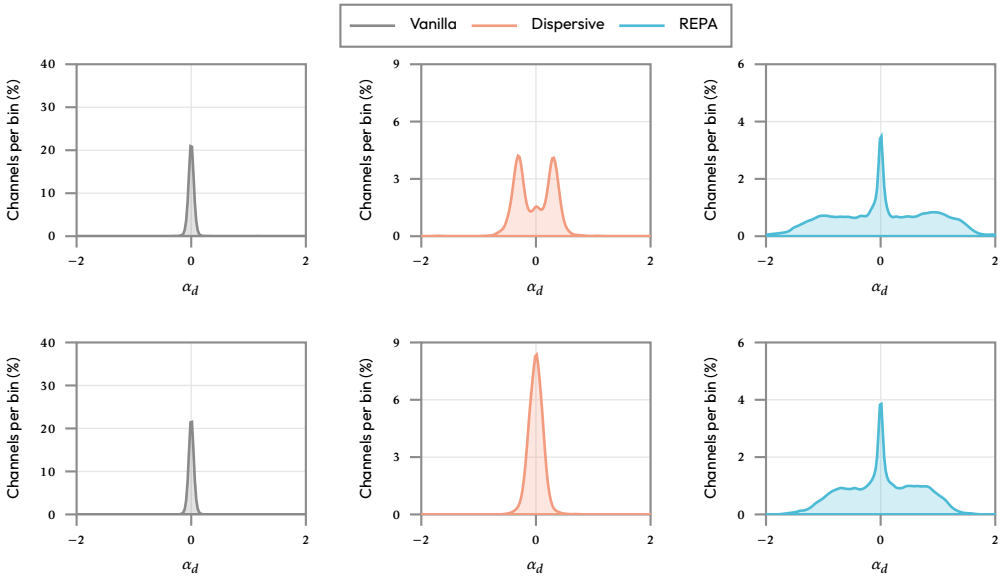


Figure B: Signed gate histograms at additional timesteps, using the same checkpoints, bins, smoothing, and display range as Fig. 5. Top row: $t = 0.25$; bottom row: $t = 0.75$. Columns: Vanilla, Dispersive, and REPA, with per-method y-axis limits shared across the two timesteps.

MoCo-v3-B (Chen et al., 2021), and WebSSL-1B (Fan et al., 2025) teachers. The FIDs are taken from the evaluations released in the same repository at step 400K, which use the ADM evaluator, CFG 1.0, and SDE sampling with 250 steps. We measure $\overline{\mathrm{nPR}(\boldsymbol{\alpha})}$ for the L8 MLP gate of each checkpoint.

B.6 Projector Sweep

All projector variants use SiT-B/2 with DINOv2-B alignment after the fourth block (L4) and alignment coefficient 0.5. Each variant is trained for 50K steps with the setup of Appendix A.1 and evaluated with the local FID protocol.

Projection heads. The three-layer MLP has widths $768 \to 1024 \to 1024 \to 768$, with SiLU between affine layers; the two-layer MLP has widths $768 \to 1024 \to 768$, with one SiLU. The deep linear head uses the same widths as the three-layer MLP without intermediate nonlinearities, and the single linear head is $768 \to 768$. Their parameter counts, including biases, are 2,624,256, 1,574,656, 2,624,256, and 590,592, respectively. All variants normalize teacher targets within each image and channel as $(x - 0.7\mu_N)/(\sigma_N + 10^{-6})$, where μ_N and σ_N are the mean and standard deviation over tokens.

C Details of the Radial Objective

C.1 Target Sweep and Target Choice

The radial penalty acts on the output of the final SiT-B/2 block (L12), flattened to a $256 \cdot 768$ dimensional vector per sample. All sweep points use 100K-step EMA checkpoints trained with the setup of Appendix A.1 and evaluated with the local FID protocol.

Targets and coefficients. We search over $c \in \{875.885, 1751.77, 3503.54, 4500, 6000\}$. For $N = 256$ and $D = 768$, $\sqrt{ND} \approx 443.4$, and these targets correspond to $k \approx \{2, 4, 8, 10, 14\}$ in $c = k\sqrt{ND}$. All five runs use $\lambda = 6.032 \times 10^{-8}$, calibrated over the first five training iterations to make the weighted auxiliary gradient norm approximately 10% of the velocity-loss gradient norm. We select the target with the lowest FID.

Reference points. The Vanilla and Dispersive reference points are 100K EMA checkpoints evaluated with the same protocol; Dispersive uses coefficient 0.25 at the final block.

Sensitivity. Fig. C plots FID and gate participation against the target. The lowest target yields FID 45.72, close to Vanilla's 45.41; the target 3503.54 reaches 40.02, the selected target 4500 reaches 39.43, and the largest target 6000 reaches 39.44. L12 MLP gate participation rises from 0.381 at the lowest target to 0.577 at the selected target and 0.564 at $c = 6000$.

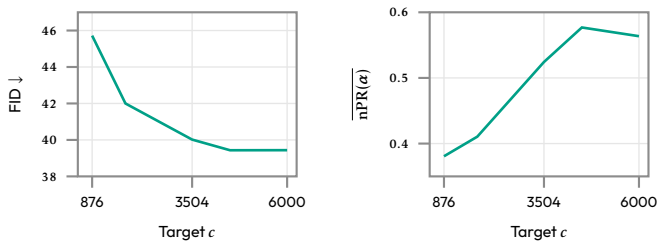


Figure C: Sensitivity to the flattened feature-norm target for SiT-B/2 at 100K steps. Left: ADM FID. Right: $\overline{\mathrm{nPR}(\boldsymbol{\alpha})}$ for L12 MLP gates over 101 uniformly spaced model timesteps in $[0, 1]$ and 1,000 classes.

C.2 Performance Comparison

Table 3 uses SiT-B/2 at 256×256, 400K training steps, the setup of Appendix A.1, and the local FID protocol of Appendix A.3. The three teacher-free rows are the same runs as the training trajectories

in Fig. 7. Dispersive uses coefficient 0.25 at the final block, and the Radial objective uses the target $c = 10\sqrt{ND} \approx 4434$ with coefficient $\lambda = 5 \times 10^{-8}$. REPA aligns L4 with DINOv2-B through a three-layer MLP projector with coefficient 0.5. Mean gate participation is measured at L12 on the same EMA checkpoints.

C.3 Training Dynamics

Fig. 7 uses the three 400K SiT-B/2 runs of Appendix C.2. The loss curves are smoothed with a trailing 5,000-step moving average over every logged iteration. Each raw, unweighted auxiliary loss is normalized as $(L_t - L_{\min})/(L_0 - L_{\min})$, with $L_{\min} = -\ln 32$ for Dispersive and $L_{\min} = 0$ for Radial. The gate participation panel evaluates the L12 MLP gate of the EMA checkpoints at 50K, 100K, 200K, and 400K steps with the protocol of Appendix A.2.

D Conditions and Proof of Proposition 1

Analytical setting. Fix (t, y) at the current model state and let $\boldsymbol{\alpha} = \boldsymbol{\alpha}_\theta(t, y) \in \mathbb{R}^D$. We represent the input realizations at this condition by B feature pairs $\boldsymbol{H}^{(b)}, \boldsymbol{U}^{(b)} \in \mathbb{R}^{N\times D}$, held fixed with respect to the local gate variation, so that $\mathbb{E}[\cdot \mid t, y]$ in the main text is the average over b. These realizations share (t, y) and hence the gate; they need not appear together in a training minibatch. Here $\widetilde{r}_{\boldsymbol{\alpha}}^{(b)}$ denotes Eq. (3) evaluated at $(\boldsymbol{H}^{(b)}, \boldsymbol{U}^{(b)})$. The analyzed loss is

$$\widetilde{\mathcal{L}}_{\mathrm{rad}}(\boldsymbol{\alpha}) = \frac{1}{B}\sum_{b=1}^{B}(\widetilde{r}_{\boldsymbol{\alpha}}^{(b)} - c)^2. \tag{7}$$

Here $c > 0$. At the update point, all α_d are nonzero and all $\widetilde{r}_{\boldsymbol{\alpha}}^{(b)}$ are positive. These conditions ensure that the finite-average loss, $\mathrm{nPR}(\boldsymbol{\alpha})$, and the fractional gate updates are well-defined and smooth locally. A positive auxiliary coefficient merely rescales the step size.

Precise channel-symmetry assumption. For every permutation π of the D channel indices, the two multisets

$$\left\{\left(\|\boldsymbol{H}^{(b)}\|_{\mathrm{F}}^2, \|\boldsymbol{U}_{:,1}^{(b)}\|_2^2, \ldots, \|\boldsymbol{U}_{:,D}^{(b)}\|_2^2\right)\right\}_{b=1}^{B}, \tag{8}$$

$$\left\{\left(\|\boldsymbol{H}^{(b)}\|_{\mathrm{F}}^2, \|\boldsymbol{U}_{:,\pi(1)}^{(b)}\|_2^2, \ldots, \|\boldsymbol{U}_{:,\pi(D)}^{(b)}\|_2^2\right)\right\}_{b=1}^{B} \tag{9}$$

coincide, including multiplicities. The skip norm remains attached to its sample pattern, and the gate is not permuted. This condition concerns squared-norm patterns, not individual feature entries or their signs. It implies identical averaged squared channel norms, but is stronger than equality of these averages alone. Neither Gaussianity nor independence is assumed. Exact symmetry is a property of the analytical model, not a consequence of an architecture lacking hard-coded channel preferences and not a claim about observed mini-batches.

Relative gate update. Differentiating Eq. (7) gives

$$-\frac{\partial\widetilde{\mathcal{L}}_{\mathrm{rad}}}{\partial\alpha_d} = \frac{2\alpha_d}{B}\sum_b\left(\frac{c}{\widetilde{r}_{\boldsymbol{\alpha}}^{(b)}} - 1\right)\|\boldsymbol{U}_{:,d}^{(b)}\|_2^2, \tag{10}$$

which gives Eq. (5). Symmetry makes $B^{-1}\sum_b \|\boldsymbol{U}_{:,d}^{(b)}\|_2^2$ independent of d.

Ordering of the inverse-norm-weighted responses. For $i \neq j$, define the norm with the two channel norms interchanged:

$$\widetilde{r}_{\boldsymbol{\alpha}}^{(b;i\leftrightarrow j)} = \Big(\|\boldsymbol{H}^{(b)}\|_{\mathrm{F}}^2 + \alpha_i^2\|\boldsymbol{U}_{:,j}^{(b)}\|_2^2 + \alpha_j^2\|\boldsymbol{U}_{:,i}^{(b)}\|_2^2 + \sum_{k\neq i,j}\alpha_k^2\|\boldsymbol{U}_{:,k}^{(b)}\|_2^2\Big)^{1/2}. \tag{11}$$

The channel-symmetry assumption gives

$$\begin{aligned}
&\frac{1}{B}\sum_b \frac{\|\boldsymbol{U}_{:,i}^{(b)}\|_2^2}{\widetilde{r}_{\boldsymbol{\alpha}}^{(b)}} - \frac{1}{B}\sum_b \frac{\|\boldsymbol{U}_{:,j}^{(b)}\|_2^2}{\widetilde{r}_{\boldsymbol{\alpha}}^{(b)}} \\
&= \frac{1}{2B}\sum_b (\|\boldsymbol{U}_{:,i}^{(b)}\|_2^2 - \|\boldsymbol{U}_{:,j}^{(b)}\|_2^2)\left(\frac{1}{\widetilde{r}_{\boldsymbol{\alpha}}^{(b)}} - \frac{1}{\widetilde{r}_{\boldsymbol{\alpha}}^{(b;i\leftrightarrow j)}}\right) \\
&= -\frac{\alpha_i^2 - \alpha_j^2}{2B}\sum_b \frac{(\|\boldsymbol{U}_{:,i}^{(b)}\|_2^2 - \|\boldsymbol{U}_{:,j}^{(b)}\|_2^2)^2}{\widetilde{r}_{\boldsymbol{\alpha}}^{(b)}\,\widetilde{r}_{\boldsymbol{\alpha}}^{(b;i\leftrightarrow j)}(\widetilde{r}_{\boldsymbol{\alpha}}^{(b)} + \widetilde{r}_{\boldsymbol{\alpha}}^{(b;i\leftrightarrow j)})}.
\end{aligned} \tag{12}$$

The summands in the final line are nonnegative. Hence larger squared gate magnitudes have no larger inverse-norm-weighted responses, proving Eq. (6). Equality of two squared gate magnitudes also implies equality of their weighted responses. For unequal squared gate magnitudes, the ordering is strict if their squared channel norms differ in at least one sample.

One-step non-decrease of $\mathrm{nPR}(\boldsymbol{\alpha})$**.** For the proof only, define

$$\boldsymbol{v} = -\nabla_{\boldsymbol{\alpha}}\widetilde{\mathcal{L}}_{\mathrm{rad}}(\boldsymbol{\alpha}), \qquad \rho_d = \frac{v_d}{\alpha_d}, \qquad p_d = \frac{\alpha_d^2}{\sum_j \alpha_j^2}. \tag{13}$$

All $p_d > 0$, and the prior nPR definition gives $\mathrm{nPR}(\boldsymbol{\alpha}) = 1/(D\sum_d p_d^2)$. Differentiation in direction $\boldsymbol{v}$ yields

$$\begin{aligned}
\langle \nabla_{\boldsymbol{\alpha}} \log \mathrm{nPR}(\boldsymbol{\alpha}), \boldsymbol{v}\rangle &= 4\left(\sum_d p_d\rho_d - \frac{\sum_d p_d^2\rho_d}{\sum_d p_d^2}\right) \\
&= -\frac{4}{\sum_d p_d^2}\sum_{i<j} p_i p_j (p_i - p_j)(\rho_i - \rho_j) \geq 0.
\end{aligned} \tag{14}$$

Each summand has the stated sign by Eq. (6). If the directional derivative is strictly positive, Taylor expansion gives

$$\begin{aligned}
&\mathrm{nPR}(\boldsymbol{\alpha} + \eta\boldsymbol{v}) - \mathrm{nPR}(\boldsymbol{\alpha}) \\
&\qquad = \eta\,\langle \nabla_{\boldsymbol{\alpha}}\mathrm{nPR}(\boldsymbol{\alpha}), \boldsymbol{v}\rangle + O(\eta^2) > 0
\end{aligned} \tag{15}$$

for every sufficiently small positive η.

It remains to justify finite-step non-decrease when the derivative is zero; a nonnegative first derivative alone would not suffice. In that case, every summand in Eq. (14) is zero. Thus, $p_i \neq p_j$ implies $\rho_i = \rho_j$. If $p_i = p_j$, the equal-gate case of Eq. (12) also implies $\rho_i = \rho_j$. Consequently all fractional rates have a common value ρ; hence $\boldsymbol{v} = \rho\boldsymbol{\alpha}$ and the step is exactly $\boldsymbol{\alpha}^+ = (1+\eta\rho)\boldsymbol{\alpha}$. This preserves $\mathrm{nPR}(\boldsymbol{\alpha})$ whenever the scale is nonzero. Together the two cases establish the finite-step inequality in Proposition 1.

The increase is strict if there exist i, j, b such that $\alpha_i^2 \neq \alpha_j^2$ and $\|\boldsymbol{U}_{:,i}^{(b)}\|_2^2 \neq \|\boldsymbol{U}_{:,j}^{(b)}\|_2^2$. The argument concerns sufficiently small steps, not arbitrary learning rates. Choosing the step small enough also preserves gate signs.

Quadratic comparison. For the same cross-term-excluded norm,

$$\begin{aligned}
\mathcal{Q}_{\pm}(\boldsymbol{\alpha}) &= \pm\frac{1}{B}\sum_b (\widetilde{r}_{\boldsymbol{\alpha}}^{(b)})^2 \\
&= \pm\left[\frac{1}{B}\sum_b \|\boldsymbol{H}^{(b)}\|_{\mathrm{F}}^2 + \sum_d \alpha_d^2 \frac{1}{B}\sum_b \|\boldsymbol{U}_{:,d}^{(b)}\|_2^2\right].
\end{aligned} \tag{16}$$

For this calculation only, let $\bar{q}$ denote the common average $B^{-1}\sum_b \|\boldsymbol{U}_{:,d}^{(b)}\|_2^2$. A gradient step gives $\boldsymbol{\alpha}^+ = (1 \mp 2\eta\bar{q})\boldsymbol{\alpha}$ and therefore preserves $\mathrm{nPR}(\boldsymbol{\alpha})$ whenever the scale factor is nonzero. □

D.1 Scope of the Analysis

The actual residual norm satisfies

$$\begin{aligned}&\|\boldsymbol{H}^{(b)} + \boldsymbol{U}^{(b)}\,\mathrm{diag}(\boldsymbol{\alpha})\|_{\mathrm{F}}^2 \\ &\quad = (\widetilde{r}_{\boldsymbol{\alpha}}^{(b)})^2 + 2\sum_d \alpha_d \langle \boldsymbol{H}_{:,d}^{(b)}, \boldsymbol{U}_{:,d}^{(b)} \rangle.\end{aligned} \tag{17}$$

The analysis intentionally omits the last term. Orthogonality such as $\boldsymbol{h}^\top \boldsymbol{u} = 0$, or $\langle \boldsymbol{H}, \boldsymbol{U} \rangle_{\mathrm{F}} = 0$, does not imply that this gate-weighted cross term vanishes. Nor does a small cross-term value at one gate imply a small derivative with respect to that gate. The proposition applies to the cross-term-excluded norm, not to the full residual norm without additional assumptions.

In actual training the gate is parameterized; other features, task gradients, and optimizer state can also change. The proposition concerns one local Euclidean gate update and does not imply monotonic $\mathrm{nPR}(\boldsymbol{\alpha})$ growth under those full dynamics or a causal relationship between $\mathrm{nPR}(\boldsymbol{\alpha})$ and FID. The quadratic comparison is likewise restricted to the same analysis.