COMMUTATOR-GUIDED CALIBRATION OF ATTENTION-HEAD INTERACTIONS IN DIFFUSION TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion Transformers (DiTs) aggregate attention-head outputs additively without explicitly composing their state-dependent residual operators. We diagnose their non-commutativity (NC), or dependence on composition order, using routing-matrix commutators and changes in predicted clean latents under reversed head-operator composition. Their calibrated comparison defines weighted undershoot graphs: Laplacian traces summarize layer-level mismatch, while edge weights identify pairs with weak functional responses relative to their routing-level reference. We use a two-phase procedure: Phase 1 probes a frozen pretrained model to select compact, layer-specific head-pair sets without generation metrics; Phase 2 fine-tunes the model with a one-sided hinge objective applied to these pairs. Across ImageNet and CUB-200-2011, the results suggest that intervention efficacy reflects both calibrated undershoot and layer position: high diagnostic energy alone does not guarantee larger generation gains. Across continuation seeds, simultaneous NC regularization at layers 10 and 24 achieves ImageNet validation FID 6.435 ± 0.161, compared with 8.350 for pretrained DiT-XL/2, and improves Inception Score, and precision over matched denoising-only (no-NC) fine-tuning. On CUB-200-2011, NC achieves FID 5.492 ± 0.117, compared with 8.270 for the domain-adapted baseline, with higher precision. No-NC controls support gains beyond continued denoising training, while pair-budget ablations favour compact interventions over broader regularization. These results support efficacy of selective regularization of attention-head interactions while retaining the additive DiT architecture.

1 INTRODUCTION

Modern foundation models dominate vision tasks by learning rich latent representations that drive state-of-the-art generative approaches such as diffusion models (Peebles & Xie, 2023; Rombach et al., 2022). While these representations enable impressive image synthesis, they lack straightforward geometric interpretation. Yet if latents are to faithfully support image generation under transformations, they must obey fundamental constraints from representation theory, *i.e.* groups acting on the image space (Jin et al., 2024). Many transformations are inherently *non-commutative*: applying transformation A followed by B generally differs from applying B followed by A Chirikjian (2012); Cohen & Welling (2016); Bronstein et al. (2021) (See Figure 1, left panel). Canonical examples include 3D rotations, camera motions, and articulated object transformations. These transformations form non-Abelian groups, where order encodes structure. Moreover, any faithful group representation in the latent space must preserve this non-Abelian structure. Many generative models adopt Euclidean latent parametrizations, with DiTs aggregating attention-head outputs additively regardless of whether the corresponding state-dependent residual operators commute under sequential composition. Recent geometric extensions introduce Riemannian metrics, often derived from covariance or Fisher information matrices, to capture curvature and local sensitivity in the denoising process (Karczewski et al., 2026; Chen & Lipman, 2024). However, metric structure in these methods primarily governs notions of distance, curvature, and shortest-path behaviour. It does not directly model the algebraic structure of ordered operator composition. When transport arises from the interaction of state-dependent transformation operators, the resulting order dependence is an operator-level

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.

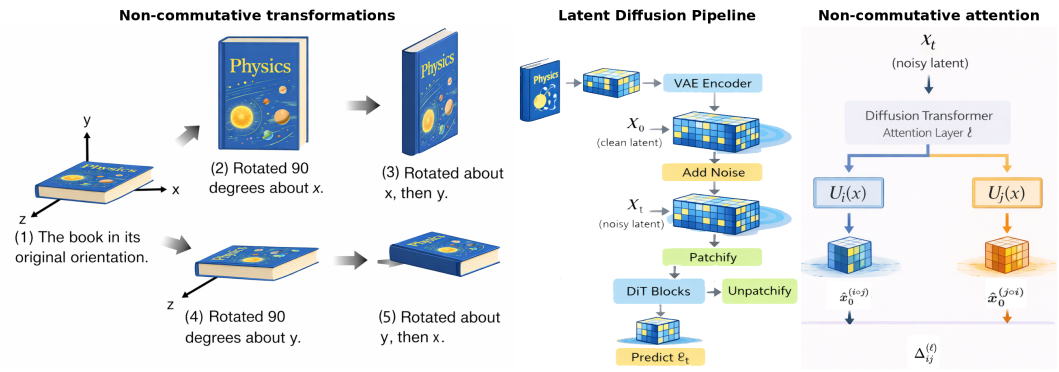


Figure 1: **Non-commutativity in latent diffusion transformers.** *Left:* Non-commutative transformations. *Middle:* Latent diffusion pipeline. *Right:* Non-commutative attention, reversing the order of two head-induced residual operators can change the predicted clean latent, revealing functional order sensitivity $\Delta_{ij}^{(\ell)}$ beyond algebraic routing non-commutativity.

phenomenon rather than a purely metric one. Such interactions are therefore not explicitly represented at the architectural level. This structural mismatch is particularly salient in modern Diffusion Transformers (DiTs) (Peebles & Xie, 2023). While self-attention without positional information is permutation-equivariant along the sequence dimension, recent evidence suggests that attention heads can exhibit non-commutative interactions internally (Ruscio et al., 2025), acting as state-dependent operators whose compositions admit a local Lie-bracket interpretation. Yet, architectural aggregation remains additive, without explicitly composing individual head updates within a block. This highlights a distinction: individual attention heads induce state-dependent residual operators, while the architecture aggregates their outputs linearly. Second-order commutator terms associated with sequential head composition are therefore not explicitly represented in this aggregation rule. We refer to this distinction between additive aggregation and order-dependent operator composition as the *additive paradox*. While prior work has examined additive head aggregation through feature attribution, redundancy, and representational capacity (Leemann et al., 2025; Nguyen et al., 2022), its geometric effect on commutator interactions in diffusion transformers have not, to our knowledge, been explicitly formalized. Order-dependent head interactions admit a geometric interpretation through the composition of state-dependent residual operators and may contribute to generation. We therefore investigate regularizing algebraic routing non-commutativity according to its measured functional consequence. A mismatch arises when reversing head-operator composition produces a weak response at the denoiser output relative to the calibrated routing-level target. Our objective is to penalize this mismatch without requiring all head interactions to commute. Specifically, the calibrated hinge penalizes routing non-commutativity when the constraint is violated and stops penalizing it once the constraint is satisfied, potentially at a nonzero commutator magnitude. In this work, we develop a framework for diagnosing and regularizing attention-head non-commutativity in Diffusion Transformers. We introduce two complementary diagnostics: (1) an *algebraic routing measure* that quantifies commutator magnitude between attention routing matrices; and (2) a *functional order-sensitivity measure* that quantifies changes in predicted clean latents under reversed composition of head-induced residual operators. Their calibrated comparison identifies head pairs whose functional response at the denoiser output is weak relative to the routing-level target, guiding layer-specific pair selection and selective regularization.

In summary, this paper makes three contributions: (a) an operator-based analysis of attention-head order dependence in Diffusion Transformers, with a local Lie-bracket interpretation of state-dependent residual compositions; (b) complementary algebraic and functional diagnostics that characterize the relationship between routing non-commutativity and order sensitivity at the denoiser output; and (c) a layer-wise calibrated hinge regularizer applied to a small, diagnostically selected set of attention-head pairs, enabling short additional fine-tuning that improves generation over matched denoising-only fine-tuning while retaining the additive DiT architecture.

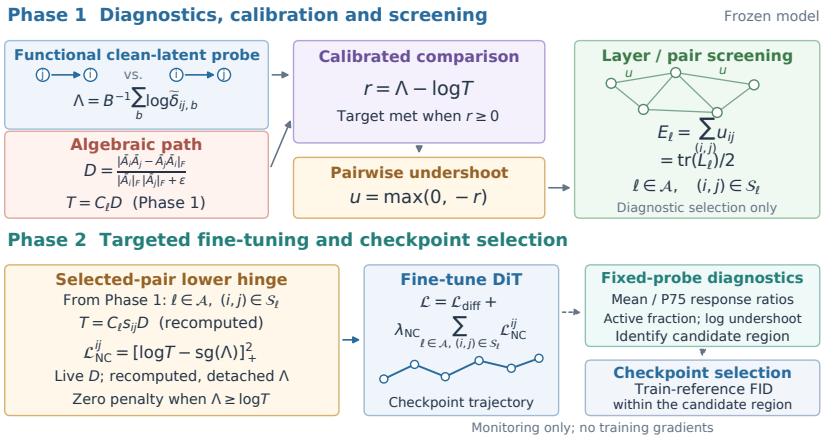


Figure 2: **Two-phase diagnosis, selective fine-tuning, and checkpoint selection.** Phase 1 compares functional order sensitivity with a calibrated routing target through the log-ratio $r_{ij}^{(\ell)}$. Its negative part defines the undershoot weights used for graph-energy layer screening and pair selection. Phase 2 applies the lower-hinge NC loss only to the selected layer/pair sets. The functional response is recomputed but detached, so NC gradients pass through the routing commutator. Fixed-probe diagnostics and train-reference FID are used for monitoring and checkpoint selection, not as training losses. sg denotes stop-gradient.

1.1 Related work

Recent work has begun to extend diffusion models into non-commutative and operator-theoretic settings, including free-probabilistic formulations of denoising diffusion (Das, 2025), operator-order analyses of diffusion-attention interactions (Zhang & Zhou, 2026), and Lie group representations for molecular conformers (Bertolini et al., 2025). In parallel, geometric analyses have emphasized Riemannian in Park et al. (2023) and information-geometric structure (Karczewski et al., 2026) in latent space. Prior work formulates diffusion in free-probabilistic, geometric, or Lie-theoretic terms. We instead study attention-head non-commutativity within Diffusion Transformers, using layer-wise diagnostics and selective hinge regularization to address mismatches between routing non-commutativity and functional order sensitivity.

2 Method

Our approach comprises two phases, summarized in Fig. 2. Phase 1 probes a frozen pretrained DiT to measure routing non-commutativity and functional order sensitivity. Their calibrated comparison identifies candidate intervention layers and compact sets of head pairs. Phase 2 fine-tunes the model using the standard diffusion objective together with a calibrated hinge regularizer applied only to the selected pairs. The additive DiT architecture is retained throughout.

2.1 Diffusion setup and head-induced operators

Let $x \in \mathbb{R}^{N \times d}$ denote the token representation inside a DiT attention block (Peebles & Xie, 2023), where N is the number of spatial tokens and d the channel dimension. For readability, we omit modulation, gating, and bias terms from the notation. The diagnostic probes use the original DiT block implementation. The layer normalization $\tilde{x} = \mathrm{LN}(x)$, queries, keys, and values for head i are:

$$Q_i = \tilde{x}W_Q^{(i)}, \quad K_i = \tilde{x}W_K^{(i)}, \quad V_i = \tilde{x}W_V^{(i)}, \tag{1}$$

where $W_Q^{(i)}, W_K^{(i)}, W_V^{(i)} \in \mathbb{R}^{d \times d_h}$ are learnable projection matrices, d_h denotes the head dimension, and $d = Hd_h$ with H the number of heads. The routing matrix A_i and head output $o_i(\tilde{x})$ are:

$$A_i = \mathrm{softmax}\left(\frac{Q_i K_i^\top}{\sqrt{d_h}}\right) \in \mathbb{R}^{N \times N}, \qquad o_i(\tilde{x}) = A_i V_i\,. \tag{2}$$

The full attention update aggregates heads additively after the output projection as $x \mapsto x + \sum_{i=1}^{H} o_i(\tilde{x})W_O^{(i)}$, where $W_O^{(i)} \in \mathbb{R}^{d_h \times d}$ denotes the corresponding block of

the output projection. Each head, therefore, induces a state-dependent residual operator $U_i(x) = x + o_i(\mathrm{LN}(x))W_O^{(i)}$. While the standard architecture applies normalization once and computes all heads in parallel, this operator decomposition allows us to analyze hypothetical sequential compositions. Even if routing matrices commute algebraically ($A_iA_j = A_jA_i$), the induced residual updates can fail to commute because the second head receives a different normalized state. Thus, state-dependent nonlinear normalization can contribute to non-commutativity. For the local analysis, define: $\mathcal{V}_i(x) = U_i(x) - x$ and $U_i^\varepsilon(x) = x + \varepsilon\mathcal{V}_i(x)$. The normalization difference and small-step composition expansion are:

$$\mathrm{LN}(U_j(x)) \neq \mathrm{LN}(x)\,, \qquad U_i^\varepsilon \circ U_j^\varepsilon - U_j^\varepsilon \circ U_i^\varepsilon = -\varepsilon^2[\mathcal{V}_i, \mathcal{V}_j] + \mathcal{O}(\varepsilon^3)\,, \tag{3}$$

where $[\mathcal{V}_i, \mathcal{V}_j] = D\mathcal{V}_j\,\mathcal{V}_i - D\mathcal{V}_i\,\mathcal{V}_j$. This expansion provides a discrete proxy for Lie-bracket interactions. We refer to the tension between internal non-commutativity and additive aggregation as the *additive paradox*.

Denoising objective. Let $x_0 \sim \mathcal{D}$ denote a clean latent sample. The forward diffusion process samples a timestep $t \in \{0, \ldots, T\}$ and constructs $x_t = \sqrt{\bar{\alpha}_t}\,x_0 + \sqrt{1-\bar{\alpha}_t}\,\epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$ and $\bar{\alpha}_t$ is the cumulative product of noise schedule coefficients Ho et al. (2020). The DiT predicts noise $\hat{\epsilon}_t = f_\theta(x_t, t)$, reconstructs the clean latent $\hat{x}_0 = (x_t - \sqrt{1-\bar{\alpha}_t}\,\hat{\epsilon}_t)/\sqrt{\bar{\alpha}_t}$, and is trained with the diffusion objective $\mathcal{L}_{\text{diff}} = \|\hat{\epsilon}_t - \epsilon\|_2^2$.

2.2 Diagnostics of Non-commutativity

Fig. 2 summarizes the diagnostic-to-intervention pipeline. Phase 1 combines functional order sensitivity, measured under reversed head-operator composition, and normalized routing commutators through a calibrated log-ratio, guiding layer screening and head-pair selection. Phase 2 applies a one-sided hinge objective to the selected pairs. Both are recomputed, but the functional response is detached, so NC gradients propagate through the routing commutator. Fixed-probe diagnostics and train-reference FID support monitoring and checkpoint selection.

2.2.1 Functional order sensitivity

For Phase 1 diagnostics, we use the forward diffusion process defined above with t sampled from the probe interval $[t_{\min}, t_{\max})$. The same interval is used when constructing the fixed-probe NC statistics monitored along the Phase 2 checkpoint trajectory. During ordinary Phase 2 training, diffusion timesteps are sampled over the full training range, but the NC loss is evaluated only on the mini-batch subset whose timesteps fall within the NC probe interval. For a selected layer ℓ and heads (i, j), we compare two sequential single-head residual compositions, $U_{i\circ j} = U_i \circ U_j$ and $U_{j\circ i} = U_j \circ U_i$. Thus, under the standard convention for function composition, $i \circ j$ applies head j first and then head i. Each single-head update is evaluated on the updated hidden state, using the same DiT block modulation and attention gate as in the original model implementation. This gives three noise predictions: the unmodified prediction $\hat{\epsilon}_t$, and the two order-probed predictions $\hat{\epsilon}_t^{(i\circ j)}$ and $\hat{\epsilon}_t^{(j\circ i)}$. We convert each prediction to a reconstructed clean latent, $\hat{x}_0(\hat{\epsilon}) = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\,\hat{\epsilon}}{\sqrt{\bar{\alpha}_t}}$. Motivated by the quadratic small-step scaling of Lie-bracket order effects (Murray et al., 1994), we normalize the reconstructed-latent order difference by the squared mean output perturbation. For sample b, define

$$\delta_{ij,b}^{(\ell)} = \frac{\|\hat{x}_{0,b}^{(i\circ j)} - \hat{x}_{0,b}^{(j\circ i)}\|_2}{\max\Big([\|\hat{x}_{0,b}^{(i\circ j)} - \hat{x}_{0,b}\|_2 + \|\hat{x}_{0,b}^{(j\circ i)} - \hat{x}_{0,b}\|_2]/2, \eta\Big)^2 + \varepsilon} \cdot \dim(\hat{x}_0)^{-1/2}. \tag{4}$$

Here η is a small step-size floor. The denominator normalizes by the squared size of the two single-order perturbations relative to the unmodified prediction, while the floor prevents numerical instability when the perturbation magnitude is very small. Because the same perturbation can have different scale at different diffusion timesteps, we apply an SNR weighting:

$$\widetilde{\delta}_{ij,b}^{(\ell)} = \frac{\min(\mathrm{SNR}(t_b), 10)}{\overline{\mathrm{SNR}}}\,\delta_{ij,b}^{(\ell)}, \qquad \mathrm{SNR}(t) = \frac{\bar{\alpha}_t}{1-\bar{\alpha}_t}. \tag{5}$$

Here $\overline{\mathrm{SNR}}$ is the average SNR over the probe interval. Thus $\widetilde{\delta}_{ij,b}^{(\ell)}$ is the SNR-weighted functional order sensitivity for head pair (i, j) in layer ℓ on sample b. We define the Phase-1 diagnostic as the

mean log observable, $\Lambda_{ij}^{(\ell)} = \frac{1}{B}\sum_{b=1}^{B} \log \widetilde{\delta}_{ij,b}^{(\ell)}$, rather than the log of an arithmetic mean. Phase 2 uses the same mean-log observable inside the NC fine-tuning loss, matching the quantity used for pair selection to the training objective.

2.2.2 Algebraic routing non-commutativity

For a layer ℓ, let $A_i^{(\ell)}$ be the softmax-normalized attention routing matrix for head i. During diagnostic computation the model stores the batch-mean routing operator, and the algebraic routing non-commutativity of heads i and j, denoted as $D_{ij}^{(\ell)}$, is computed as:

$$D_{ij}^{(\ell)} = \frac{\|\bar{A}_i^{(\ell)}\bar{A}_j^{(\ell)} - \bar{A}_j^{(\ell)}\bar{A}_i^{(\ell)}\|_F}{\|\bar{A}_i^{(\ell)}\|_F \|\bar{A}_j^{(\ell)}\|_F + \varepsilon}, \qquad \text{where} \quad \bar{A}_i^{(\ell)} = B^{-1}\sum\nolimits_{b=1}^{B} A_{i,b}^{(\ell)}. \tag{6}$$

This is a routing-level diagnostic. It depends on attention operators from the current hidden states, with value vectors and MLP features contributing only through those states.

2.2.3 Calibration and Phase-2 lower-hinge objective

Phase 1 estimates an initial layer-specific scale C_ℓ and defines the calibrated log-ratio $r_{ij}^{(\ell)} = \Lambda_{ij}^{(\ell)} - \log(C_\ell D_{ij}^{(\ell)})$. Negative values indicate functional order sensitivity below the calibrated routing reference. Following layer screening, we select a small set of head pairs with the most negative ratios within each selected layer. This selection uses Phase 1 diagnostics, without generation metrics. During Phase 2, we recompute the routing commutator $D_{ij}^{(\ell)}$ and mean-log functional observable $\Lambda_{ij}^{(\ell)}$ on the current NC-evaluated mini-batch subset. With an optional positive pair scale $s_{ij}^{(\ell)}$, the target is $T_{ij}^{(\ell)} = C_\ell s_{ij}^{(\ell)} D_{ij}^{(\ell)}$. The underlying lower-hinge objective is

$$\mathcal{L}_{\mathrm{NC}}^{ij} = \left[\log T_{ij}^{(\ell)} - \mathrm{sg}(\Lambda_{ij}^{(\ell)})\right]_+^2, \tag{7}$$

where $[z]_+ = \max(0, z)$ and sg denotes stop-gradient. The functional statistic is recomputed but detached, so the NC gradient passes through the live routing commutator rather than the order-probed predictions. The penalty vanishes when $\Lambda_{ij}^{(\ell)} \geq \log T_{ij}^{(\ell)}$. Calibration and implementation details, together with samplewise aggregation ablations, are provided in Appendices A.6 and A.7.

2.3 Structured Two-Phase Training

Phase 1 probes the pretrained model to screen layers and head pairs using calibrated undershoot:

$$u_{ij}^{(\ell)} = \max(0, -r_{ij}^{(\ell)}), \qquad E_\ell = \sum_{(i,j)} u_{ij}^{(\ell)} = \frac{1}{2}\mathrm{tr}(L_\ell), \tag{8}$$

where L_ℓ is the weighted head-graph Laplacian and the sum counts each probed unordered pair once. The graph energy E_ℓ measures the total calibrated undershoot within a layer. Guided by Fig. 3, we select layer 3 at the dominant early energy peak, layer 10 as an intermediate-depth comparison, and layer 24 at the strongest late-depth peak. Within each selected layer, we retain the top quartile of under-satisfied pairs ranked by $u_{ij}^{(\ell)}$, yielding seven pairs per layer. Phase 2 fine-tunes the model with the NC loss applied only to these selected pairs. Robustness tests show consistent pair prioritization across the evaluated probe batches and routing-aggregation choices (Appendix A.1).

Phase 2 optimization objective. During Phase 2, the model is fine-tuned with the standard diffusion loss plus the selected-pair NC regularizer. Let $\mathcal{A}$ denote the chosen intervention layer set, and let $\mathcal{S}_\ell$ denote the Phase 1-selected head-pair set for layer $\ell \in \mathcal{A}$. The training objective is $\mathcal{L} = \mathcal{L}_{\mathrm{diff}} + \lambda_{\mathrm{NC}} \sum_{\ell\in\mathcal{A}} \sum_{(i,j)\in\mathcal{S}_\ell} \mathcal{L}_{\mathrm{NC}}^{ij}$. The pairwise loss $\mathcal{L}_{\mathrm{NC}}^{ij}$ is the one-sided mean-log lower-hinge loss defined in Section 2.2. Thus Phase 2 does not apply a global NC regularizer over all heads or all layers; it acts only on the selected under-satisfied pairs in the chosen intervention layer set.

2.4 Models and Baselines

We compare three model variants: (i) the standard pretrained Diffusion Transformer (DiT), which serves as the primary baseline without explicit non-commutativity modelling; (ii) our NC-

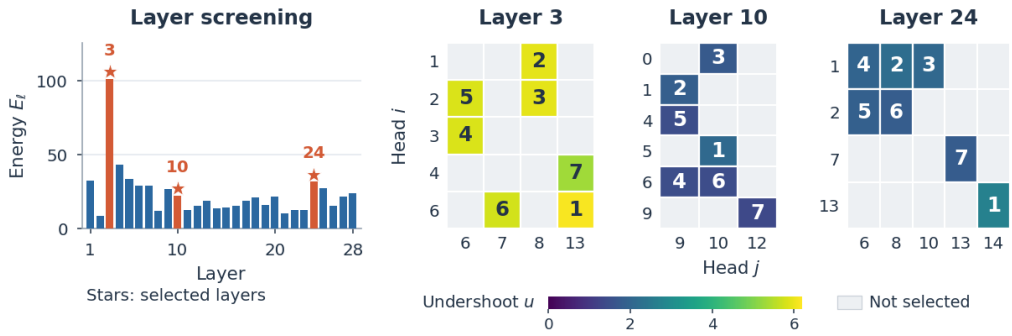


Figure 3: **Phase 1 diagnosis and Phase 2 pair selection.** Left: Phase 1 graph energy across layers, computed as $E_\ell = \sum_{(i,j)} u_{ij}^{(\ell)}$, where $u_{ij}^{(\ell)} = \max(0, -r_{ij}^{(\ell)})$. Stars indicate the selected Phase 2 intervention layers. Right: heatmaps show the seven selected pairs per layer, corresponding to the top quartile of undersatisfied pairs in Layers 3, 10, and 24. Colours show undershoot magnitude on a shared scale; numbers indicate within-layer rank (1 is largest). Phase 2 uses only the selected pairs for the chosen intervention layer or layer set. Rows and columns without selected pairs are omitted; axis labels retain the original zero-indexed head IDs.

Table 1: Ablation comparisons and corresponding results and implementation details.

Ablation	Comparison	Reference
(i) No-NC control	NC fine-tuning versus matched denoising-only continuation.	Table 2
(ii) Pair-set size	Top-10%, 25%, and 50% of pairs ranked by Phase 1 undershoot, with the remaining setup fixed.	Fig. 8
(iii) Intervention location	Layers 3, 10, and 24; neighbouring Layer 25; and joint Layer 10+24 intervention.	Table 2; Fig. 7
(iv) Samplewise aggregation	Original hinge versus samplewise variants with shared or per-image routing targets at Layer 24. Actual NC subset sizes are reported.	Appendix A.7
(v) Hinge versus shrinkage	Fixed-controller comparison on the same seven Layer 24 pairs. The shrinkage coefficient is fixed after matching initial RMS gradient strength.	Appendix A.7; Table 8
(vi) Pair-selection robustness	Ranking correlations and top-quartile overlap across 100 probe batches and alternative routing-aggregation rules.	Appendix A.1; Table 3
(vii) Timestep-window sensitivity	Layer 24 NC window $[50, 250)$ versus $[25, 225)$, $[80, 180)$, and $[75, 275)$; effects on pair selection and aggregate and class-wise FID.	Appendix A.2

augmented DiT, which fine-tunes selected undershooting attention-head interactions using the calibrated NC objective; and (iii) a fine-tuning control with the same training setup but with the NC loss disabled. This control isolates the effect of the NC regularizer from ordinary continued fine-tuning. For each dataset, we evaluate NC interventions both layer-wise, using each selected candidate layer separately, and jointly, using all selected candidate layers together.

Evaluation. We evaluate class-conditional ImageNet 256×256 generation using DiT-XL/2 with DDIM-250 sampling and the EMA VAE. For quantitative evaluation, we report FID-50K, Inception Score (IS), Precision, and Recall, computed from 50k generated samples against the ADM-preprocessed ImageNet validation reference set. For qualitative evaluation, we compare matched-seed generations from the pretrained DiT and the selected NC checkpoints using the same class labels and initial noise. Absolute pixel-difference heatmaps and overlays of the highest-change regions show where layer-wise NC fine-tuning modifies the generated samples, see Fig. 6.

Checkpoint selection. For the original layer-wise runs, fixed-probe NC diagnostics identify candidate checkpoint regions, and train-reference FID-50K selects checkpoints within those regions. Continuation-seed experiments reuse these fixed endpoints. Validation metrics are computed after selection. The protocol and layer-wise trajectories are provided in Appendix A.3.

Ablation Setup. We study seven aspects of the intervention: ordinary fine-tuning, pair-set size, intervention location, loss aggregation, regularizer choice, robustness of Phase 1 pair selection, and timestep-window sensitivity. See Table 1 for comparison summarizes.

Table 2: ImageNet validation metrics and Phase 2 cost. The pretrained baseline is a single run. No-NC disables the NC objective. Bold marks the better mean within each layer.

Model	Layer	NC pairs	Microsteps	FID ↓	IS ↑	Precision ↑	Recall ↑
DiT-XL/2	–	–	–	8.350	117.02	0.573	0.684
DiT-XL/2 no-NC	3	0	250	**8.398 ± 0.083**	**116.81 ± 0.40**	**0.5709 ± 0.0032**	**0.6862 ± 0.0021**
DiT-XL/2+NC	3	7	250	8.404 ± 0.077	116.77 ± 0.37	0.5708 ± 0.0030	0.6857 ± 0.0023
DiT-XL/2 no-NC	10	0	5500	8.717 ± 0.166	115.64 ± 0.96	0.5658 ± 0.0002	**0.6895 ± 0.0002**
DiT-XL/2+NC	10	7	5500	**8.277 ± 0.011**	**118.10 ± 0.33**	**0.5777 ± 0.0009**	0.6830 ± 0.0007
DiT-XL/2 no-NC	24	0	8000	9.055 ± 0.186	114.33 ± 1.06	0.5597 ± 0.0023	**0.6895 ± 0.0005**
DiT-XL/2+NC	24	7	8000	**6.626 ± 0.188**	**130.53 ± 1.15**	**0.6133 ± 0.0039**	0.6658 ± 0.0017
DiT-XL/2 no-NC	10+24	0	8000	8.982 ± 0.073	114.55 ± 0.62	0.5592 ± 0.0040	**0.6913 ± 0.0020**
DiT-XL/2+NC	10+24	14	8000	**6.435 ± 0.161**	**132.72 ± 0.89**	**0.6302 ± 0.0020**	0.6509 ± 0.0008

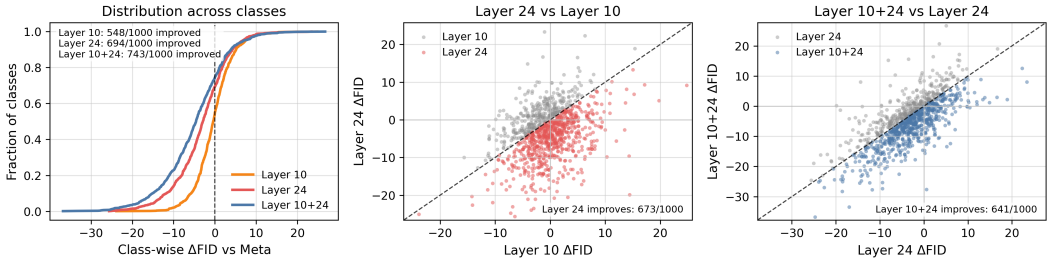


Figure 4: **Class-wise FID changes.** Left: empirical distributions relative to pretrained Meta DiT; negative values indicate improvement. Middle: paired comparison of layer 10 and Layer 24 changes. Right: paired comparison of layer 24 and joint layer 10+24 changes.

3 Results

3.1 Quantitative Results

Fixed-probe NC diagnostics identify candidate checkpoint regions, within which train-reference FID-50K guides checkpoint selection (Figure 9 (last column) in Appendix A.3). After checkpoint selection, Table 2 reports generation metrics against the ADM-preprocessed ImageNet validation reference for the selected checkpoints and matched no-NC controls. The controls use the same fine-tuning setup and checkpoint schedule with $\lambda_{\mathrm{NC}} = 0$. NC improves validation FID over no-NC fine-tuning at layers 10, 24, and joint layers 10+24, with the largest gains at layer 24 and the joint intervention. At layer 3, both approaches remain close to the pretrained model. The improvements at the effective intervention sites therefore extend beyond continued denoising training alone. NC introduces additional computation through order-probed forward evaluations for the selected head pairs. The functional observable is detached; NC gradients propagate through the routing commutator. This overhead is restricted to a compact Phase 1-selected set: seven pairs per single-layer intervention and 14 pairs for the joint layer 10+24 run. The selected layer 10, layer 24, and joint checkpoints require only 5.5k–8k additional training microsteps.

Class-wise effects.

We compute FID separately for each ImageNet class against the validation reference and define $\Delta\mathrm{FID}_c^{(\ell)} = \mathrm{FID}_c^{(\ell)} - \mathrm{FID}_c^{\mathrm{Meta}}$, where negative values indicate improvement over the pretrained Meta DiT. As shown in Fig. 4, layers 10, 24, and 10+24 improve 548, 694, and 743 of 1000 classes, respectively. Layer 24 achieves lower class-wise FID than layer 10 for 673 classes, while the joint intervention outperforms layer 24 for 641 classes, indicating broadly distributed class-level gains.

Grouping classes into coarse semantic categories reveals significant differences in ΔFID for both layer 24 (Kruskal–Wallis $H = 56.22$, $p = 1.09 \times 10^{-7}$) and layer 10+24 ($H = 80.47$, $p = 3.35 \times 10^{-12}$). Group-versus-rest Mann–Whitney tests with Benjamini–Hochberg correction (Benjamini & Hochberg, 1995; Mann & Whitney, 1947) identify the strongest median improvement in wild carnivores for layer 24 (-5.75 FID, $q = 0.0043$) and dogs for Layer 10+24 (-8.74 FID, $q = 3.04 \times 10^{-10}$). Thus, gains are widespread but vary across semantic groups. See an example in in Fig. 5. We

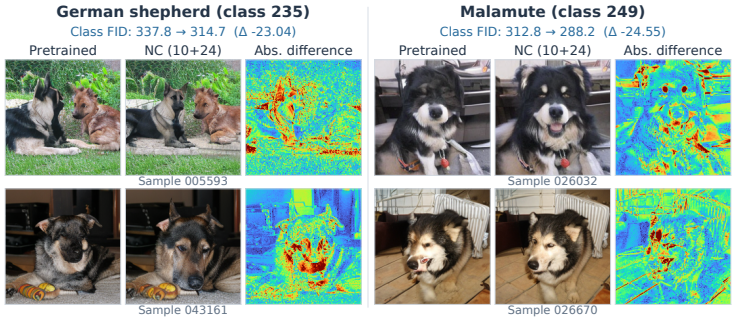


Figure 5: **Matched dog-class examples under joint layer 10+24 intervention.** Each triplet shows the pretrained DiT sample, the NC sample, and an absolute pixel-difference heatmap. FID values summarize each class; heatmaps are normalized separately for each example to reveal spatial changes.

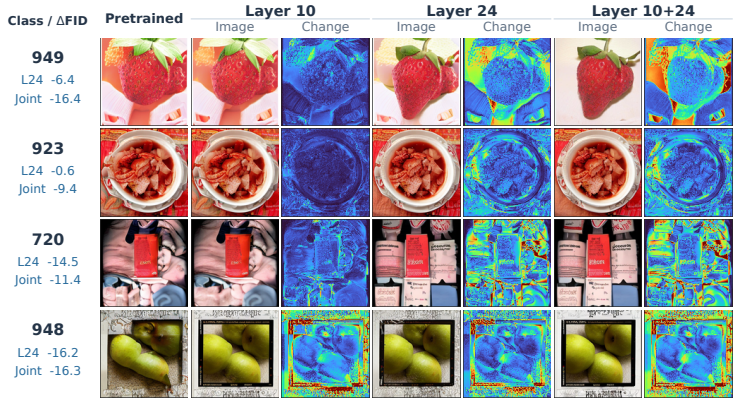


Figure 6: **Layer-dependent visual effects of NC fine-tuning.** Matched class/noise generations from the pretrained DiT and the selected layer 10, layer 24, and joint layer 10+24 checkpoints, decoded through the same EMA VAE. Change maps show absolute pixel differences from the pretrained output, using a shared heatmap scale across all rows and checkpoints. Left: class IDs, class-level ΔFID for layer 24 and the joint intervention.

next examine class-level sensitivity to the timestep window. Although the Phase 1 timestep window does not change the selected layer 24 pair set and has little effect on aggregate validation FID, it does affect which classes benefit. Specifically, we repeat Phase 1 and matched Phase 2 runs using shifted, narrower, and broader windows $[25, 225)$, $[80, 180)$, and $[75, 275)$ in place of the original $[50, 250)$ window. Among the 306 ImageNet classes that do not improve under the original $[50, 250)$ layer 24 run, alternative timestep windows rescue 219–223 classes individually, and 228 classes are rescued by at least one alternative window. Thus, the timestep window does not materially affect the global layer 24 improvement, but it redistributes class-wise gains across classes; details are reported in Appendix A.2.

3.2 Qualitative Results

Figure 6 provides a qualitative view of how NC fine-tuning changes matched generations at different intervention sites. All samples are produced by the latent DiT denoising process and decoded through the same EMA VAE. Because the same class labels and initial latent noise are used across all columns, the difference maps isolate the effect of the checkpoint rather than sampling variation. The heatmaps are computed after VAE decoding, in pixel space, as absolute differences relative to the pretrained Meta output. The layer 10 checkpoint mainly produces localized changes, such as texture, edge, and small boundary adjustments. In contrast, the layer 24 checkpoint changes larger image regions, including object extent, layout, and high-contrast structural areas. The joint layer 10+24 checkpoint preserves this broader structural effect while combining interventions across the middle and late diagnostic regimes. While qualitative, this pattern is consistent with the fixed-probe diagnostics and the quantitative gains reported in Table 2: later and joint NC interventions have greater visual leverage than the layer 10 intervention alone in these matched examples. We also evaluate transfer to CUB-200-2011, where NC fine-tuning improves FID and precision over additional denoising-only fine-tuning. Details are in Appendix A.4 (Table 7).

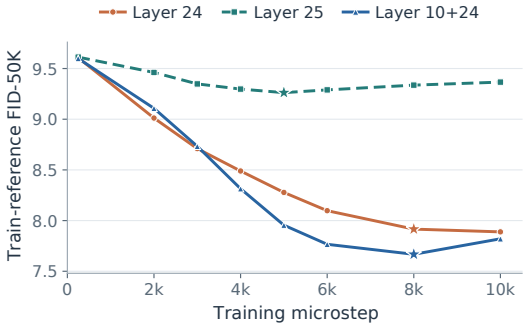


Figure 7: **Intervention location.** Train-reference FID-50K for layers 24, 25, and joint layers 10+24 at shared training steps. Stars mark selected checkpoints.

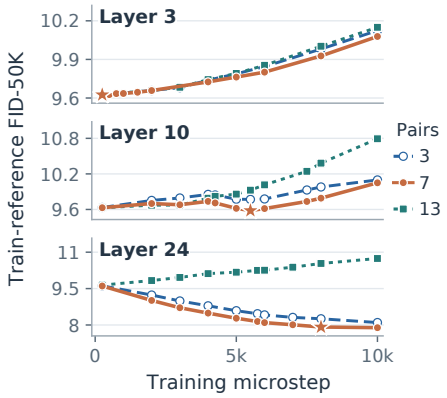


Figure 8: **Pair-set size.** Train-reference FID-50K for 3/7/13 pairs (nominal top 10/25/50%). Stars mark selected seven-pair checkpoints.

3.3 Additional Ablation Results

(1) Pair-set size. Figure 8 compares nominal top-10%, top-25%, and top-50% selections. At layer 24, top-25% performs best among the tested fractions; top-10% also improves, whereas top-50% worsens over training. Layer 10 similarly favours compact selections, while the layer 3 trajectories largely overlap and deteriorate with prolonged fine-tuning. Although the hinge penalizes only currently under-satisfied pairs, adding constraints can still alter optimization through shared attention-block parameters. These results support compact intervention, particularly at layer 24.

(2) Intervention location. Figure 7 shows that the neighbouring layer 25 does not match layer 24, reaching a best train-reference FID-50K of 9.26. Simultaneous NC regularization at layers 10 and 24 reaches train-reference FID-50K of 7.67 at step 8k, improving on the best layer 24 value of 7.89 across the evaluated trajectory. Together with the limited benefit at layer 3 despite its high undershoot energy, these comparisons show that neither high graph energy nor late depth alone identifies the most effective intervention. The joint result suggests that middle- and late-layer interventions can provide complementary benefits. Results for (4) samplewise aggregation and (5) hinge versus plain shrinkage are in Appendix A.7; (6) Phase 1 pair-selection robustness is in Appendix A.1.

4 Conclusion

We introduced a commutator-guided calibration framework for diagnosing and regularizing non-commutative attention-head interactions in DiTs. The framework addresses the additive paradox in DiT attention: although heads are aggregated additively, their state-dependent residual operators can exhibit order-sensitive compositions. By separating algebraic routing incompatibility from functional order sensitivity in the predicted clean latent, the method identifies calibrated undershoot pairs whose functional response is weak relative to their routing-level non-commutativity. Across ImageNet and CUB-200-2011, effective NC adaptation depends on both undershoot magnitude and layer position. Phase 1 should therefore be viewed as a screening diagnostic rather than a standalone oracle: graph energy localizes calibrated mismatch and narrows the intervention space, but generative gains also depend on a layer's leverage over the denoising readout. Accordingly, late-layer interventions can outperform earlier layers with higher diagnostic energy, while complementary layers can benefit from joint adaptation, as demonstrated by the layer 10+24 intervention. Empirically, NC fine-tuning improves ImageNet FID, IS, and precision relative to matched no-NC controls. On CUB, the same procedure improves FID and precision for an already domain-adapted DiT, with a recall–IS trade-off. Although order-probed pair evaluations add per-step cost, Phase 1 restricts this cost to a compact set of selected head pairs and Phase 2 requires only short adaptation runs.This yields a targeted fine-tuning method that retains the original DiT architecture. These results support commutator-guided refinement of attention-head interactions beyond standard denoising training. Future work could also extend to transformer-based flow-matching models with DiT-like residual attention blocks, such as SiT (Ma et al. (2024), by adapting the functional diagnostic to the predicted velocity field.

REFERENCES

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Marco Bertolini, Héctor Le, et al. Diffusion generative modeling on Lie group representations. *Advances in Neural Information Processing Systems*, 2025. NeurIPS Spotlight.

Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

Ricky TQ Chen and Yaron Lipman. Flow matching on general geometries. In *International Conference on Learning Representations*, volume 2024, pp. 47922–47945, 2024.

G.S. Chirikjian. *Stochastic Models, Information Theory, and Lie Groups*. Birkhäuser, 2012.

T. Cohen and M. Welling. Group equivariant convolutional networks. In *ICML*, 2016.

Swagatam Das. A free probabilistic framework for denoising diffusion models: Entropy, transport, and reverse processes. *arXiv preprint arXiv:2510.22778*, 2025.

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

Yinzhu Jin, Aman Shrivastava, and Tom Fletcher. Learning group actions on latent representations. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=HGNTcy4eEp.

Rafal Karczewski, Markus Heinonen, Alison Pouplin, Søren Hauberg, and Vikas K Garg. The spacetime of diffusion models: An information geometry perspective. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=qCsbYJZRA5.

Tobias Leemann, Alina Fastowski, Felix Pfeiffer, and Gjergji Kasneci. Attention mechanisms don't learn additive models: Rethinking feature importance for transformers. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=yawWz4qWkF.

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 *Computer Vision – ECCV 2024*, pp. 23–40. Springer, 2024. doi: 10.1007/978-3-031-72980-5_2.

H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60, 1947. doi: 10.1214/aoms/1177730491.

Richard M. Murray, Zexiang Li, and S. Shankar Sastry. *A Mathematical Introduction to Robotic Manipulation*. CRC Press, 1994. ISBN 9780849379819.

Tan Nguyen, Tam Nguyen, Hai Do, Khai Nguyen, Vishwanath Saragadam, Minh Pham, Khuong Duy Nguyen, Nhat Ho, and Stanley Osher. Improving transformer with an admixture of attention heads. volume 35, pp. 27937–27952, 2022.

Yong-Hyun Park, Mingi Kwon, Jaewoong Choi, Junghyo Jo, and Youngjung Uh. Understanding the latent space of diffusion models through the lens of riemannian geometry. volume 36, pp. 24129–24142, 2023.

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

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Valeria Ruscio, Umberto Nanni, and Fabrizio Silvestri. What are you sinking? a geometric approach on attention sink. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://arxiv.org/abs/2508.02546. Spotlight.

Yukun Zhang and Xueqing Zhou. Where to add PDE diffusion in transformers. *arXiv preprint arXiv:2510.03272*, 2026. Version 3.

Jincheng Zhong, XiangCheng Zhang, Jianmin Wang, and Mingsheng Long. Domain guidance: A simple transfer approach for a pre-trained diffusion model. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=PplM2kDrl3.

A TECHNICAL APPENDICES AND SUPPLEMENTARY MATERIAL

A.1 ROBUSTNESS TO ROUTING AGGREGATION

The graph-energy analysis characterizes how calibrated undershoot is distributed across head pairs and layers. It does not, by itself, establish whether the same pair structure persists across data batches. This distinction is relevant because the functional observable begins at the sample level, whereas the algebraic diagnostic in the main method is computed from batch-mean routing matrices.

We therefore repeated the diagnostic independently over 100 probe batches for the three selected intervention layers. For each batch k, we computed

$$r_{ij,k}^{(\ell)} = \Lambda_{ij,k}^{(\ell)} - \log\Big(C_\ell D_{ij,k}^{(\ell)}\Big), \qquad E_{\ell,k} = \sum_{(i,j)} \max\Big(0, -r_{ij,k}^{(\ell)}\Big).$$

The candidate pair set was held fixed across batches. We compared each batchwise ranking with the aggregate ranking and measured the Jaccard overlap between their most under-satisfied quartiles. We additionally compared the current algebraic score,

$$D_{ij}^{\text{mean}} = \text{Comm}(\mathbb{E}_b A_{i,b}, \mathbb{E}_b A_{j,b}),$$

with the mean samplewise score,

$$D_{ij}^{\text{sample}} = \mathbb{E}_b[\text{Comm}(A_{i,b}, A_{j,b})].$$

As shown in Table 3, batchwise undershoot rankings closely reproduce the aggregate ranking. Median Spearman correlations exceed 0.99 in all three layers, and the median top-quartile overlap is at least 0.818. The median active-batch fraction is 1.000, indicating that at least half of the evaluated pairs are under-satisfied in every probe batch. The batch-mean and samplewise commutator rankings are also strongly correlated, with ρ between 0.953 and 0.976.

Table 3: Robustness of the Phase 1 diagnostics across 100 probe batches. Batch-rank stability is the median Spearman correlation between a batchwise undershoot ranking and the aggregate ranking. Undershoot overlap is the median Jaccard overlap between their most under-satisfied quartiles. The final correlation compares commutators of batch-mean routing operators with mean samplewise commutators. Active fraction is the median, across evaluated pairs, of the fraction of batches satisfying $r_{ij,k}^{(\ell)} < 0$.

Layer	Batch-rank ρ	Undershoot overlap	Mean vs. samplewise D ρ	Active fraction
3	0.993	1.000	0.976	1.000
10	0.992	0.818	0.953	1.000
24	0.996	1.000	0.968	1.000

A.2 TIMESTEP-WINDOW SENSITIVITY

We further test whether the Phase 1 diagnostic is sensitive to the diffusion timestep window used to estimate NC undershooting. The main experiments use the diagnostic window $[50, 250)$. As sensitivity controls, we repeat Phase 1 with three representative alternative windows, $[25, 225)$, $[80, 180)$, and $[75, 275)$, while keeping the dataset, model, diagnostic statistic, layer, and retained pair fraction fixed.

Table 4 shows that the selected Layer 24 top-25% under-satisfied pair set is unchanged across all windows. All four diagnostics select the same seven head pairs: $(13, 14)$, $(1, 8)$, $(1, 10)$, $(1, 6)$, $(2, 6)$, $(2, 8)$, and $(7, 13)$. Thus, the Layer 24 intervention is not an artifact of a narrowly chosen diagnostic timestep range. Instead, the same under-satisfied head-pair structure is recovered across narrower, shifted, and broader timestep windows.

Table 4: Sensitivity of Phase 1 pair selection to the NC diagnostic timestep window for the Layer 24 intervention. Each row reports the top-25% under-satisfied pairs selected by the Phase 1 undershoot score. The selected pair set is invariant across the original window and all alternative windows.

Phase 1 timestep window	Selected Layer 24 pairs	Pair-set overlap with original
$[50, 250)$ original	$(13, 14), (1, 8), (1, 10), (1, 6), (2, 6), (2, 8), (7, 13)$	$7/7$
$[25, 225)$	$(13, 14), (1, 8), (1, 10), (1, 6), (2, 6), (2, 8), (7, 13)$	$7/7$
$[80, 180)$	$(13, 14), (1, 8), (1, 10), (1, 6), (2, 6), (2, 8), (7, 13)$	$7/7$
$[75, 275)$	$(13, 14), (1, 8), (1, 10), (1, 6), (2, 6), (2, 8), (7, 13)$	$7/7$

We next use these alternative Phase 1 diagnostics to initialize matched Phase 2 training runs, so that downstream performance is evaluated using the same timestep window that produced the selected pairs. As shown in Table 5, validation FID remains stable across the alternative windows. The shifted $[25, 225)$ window gives the lowest FID, while the original $[50, 250)$, narrower $[80, 180)$, and broader $[75, 275)$ windows all remain within a small range. This indicates that the Layer 24 improvement is not tied to a single diagnostic window.

Table 5: Sensitivity of downstream Phase 2 performance to the Phase 1 diagnostic timestep window for the Layer 24 intervention. All runs use the same Phase 2 training setup, NC objective, Layer 24 intervention site, and retained top-25% pair budget. We report validation FID at the same checkpoint used for the main Layer 24 result.

Phase 1 timestep window	Checkpoint	Validation FID ↓
$[50, 250)$ original	0008000	6.777
$[25, 225)$	0008000	**6.675**
$[80, 180)$	0008000	6.804
$[75, 275)$	0008000	6.814

We also examine whether the classes that do not improve under the original Layer 24 window can be rescued by alternative timestep windows. Let a class be "non-improved" if its class-wise ΔFID under the original $[50, 250)$ Layer 24 run is non-negative. This gives 306 classes. For each alternative window, we then recompute class-wise FID on this fixed set of classes and count how often the alternative run improves over Meta, and how often it improves over the original Layer 24 run.

Table 6: Class-level timestep-window sensitivity for Layer 24. The analysis is restricted to the 306 ImageNet classes that do not improve over Meta under the original $[50, 250)$ Layer 24 run. "Rescued" counts classes whose alternative window gives negative ΔFID relative to Meta.

Phase 1 timestep window	Rescued vs Meta	Better than original Layer 24
$[25, 225)$	$223/306$	$228/306$
$[80, 180)$	$219/306$	$228/306$
$[75, 275)$	$222/306$	$230/306$
At least one alternative	$228/306$	$234/306$

The final row reports the union over the three alternative windows: a class is counted if it is rescued, or improves over the original Layer 24 run, under at least one alternative timestep window. Overall, the timestep-window sensitivity analysis shows a two-level effect. At the global level, the diagnostic window has little impact: all alternative windows select the same Layer 24 top-25% pair set, and validation FID remains close across runs. At the class level, however, the gains are partially redistributed. Alternative windows rescue 228 of the 306 originally non-improved classes, indicating that the timestep range can affect which classes benefit most even when the global Layer 24 effect remains robust.

A.3 Checkpoint selection from NC progress and train-reference FID

Selection protocol. For each candidate intervention layer, Phase 2 produces a checkpoint trajectory. On fixed probes, we monitor the mean and 75th percentile (P75) of the calibrated ratio,

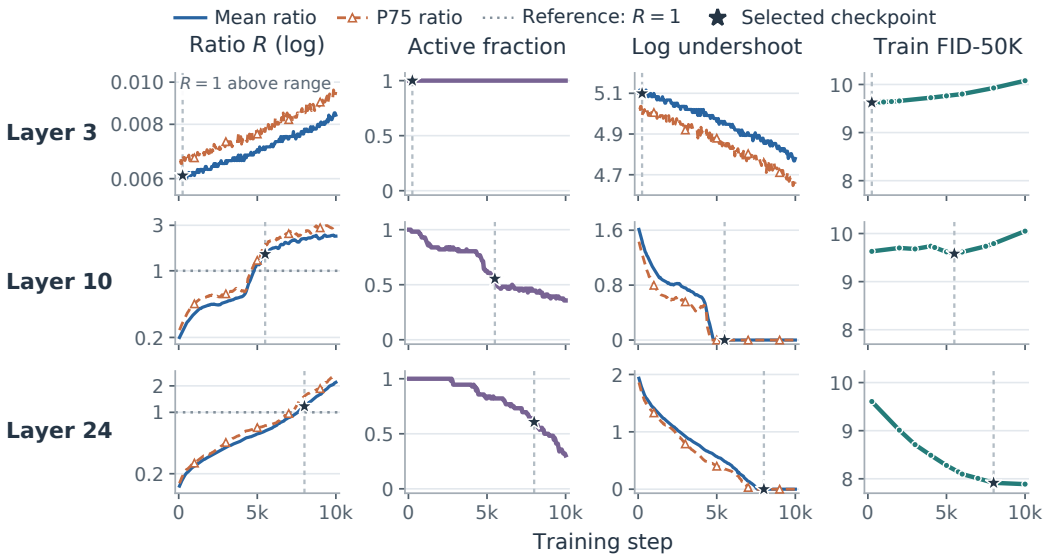


Figure 9: **Layer-dependent checkpoint-selection dynamics.** Rows correspond to Layers 3, 10, and 24. Columns show fixed-probe calibrated ratio summaries, undershoot-active fraction, log-undershoot, and train-reference FID-50K. Stars mark selected checkpoints. Ratio and log-undershoot panels use layer-specific axis ranges; the reference $R = 1$ lies above the displayed range for Layer 3. These trajectories show the original selection runs, rather than averages over continuation seeds.

the undershoot-active fraction, and the log-undershoot. These diagnostics jointly identify candidate checkpoint regions. We then compare FID-50K against the ImageNet training reference within those regions. After checkpoint selection, generation quality is evaluated against the ADM-preprocessed ImageNet validation reference. Validation metrics are not used for checkpoint selection.

The ratio reference is $R = 1$. Crossing this reference in an aggregate summary does not establish that every selected pair satisfies the lower-hinge constraint. In Figure 9, the displayed log-undershoot is computed as $\max(0, -\log R)$ separately from the mean and P75 ratio summaries; it is a monitoring statistic rather than the optimized hinge loss.

Layer 3. The selected-pair ratios remain well below the calibrated reference, while the undershoot-active fraction stays near one. Train-reference FID is lowest at checkpoint 250 and deteriorates with continued fine-tuning, reaching its worst evaluated value at checkpoint 10000. We therefore select checkpoint 250 despite persistent undershoot; selection does not require all NC constraints to be satisfied.

Layer 10. Around checkpoint 5000, the P75 ratio crosses the reference value of one and the plotted log-undershoot summary approaches zero. These trends identify a candidate region for checkpoint comparison. Among nearby evaluated checkpoints, checkpoint 5500 achieves the lowest train-reference FID-50K and is selected.

Layer 24. The diagnostics show a broader interval of sustained progress: selected-pair ratios approach the calibrated reference, the undershoot-active fraction decreases, and the log-undershoot summary plateaus. Train-reference FID improves along the trajectory, and checkpoint 8000 is selected in this regime. Although checkpoint 10000 has slightly lower train-reference FID, selection reflects the joint assessment of NC progress and generation quality rather than the global FID minimum over the entire trajectory.

Continuation-seed experiments. The additional continuation runs reuse the fixed checkpoint endpoints: 250 microsteps for Layer 3, 5500 for Layer 10, and 8000 for Layer 24 and the joint Layer 10+24 intervention. These endpoints are not reselected using the validation results of individual seeds.

A.4 CUB TRANSFER: DIAGNOSTICS, CHECKPOINT SELECTION, AND RESULTS

Transfer setup. CUB-200-2011 contains 200 bird classes, with 5,994 training images and 5,794 test images. Following Zhong et al. (2025), we replace the ImageNet class embedding of pretrained DiT-XL/2 with a randomly initialized 200-class embedding and fine-tune using the standard diffusion objective at 256×256 resolution. The checkpoint at 24k steps serves as the shared CUB-adapted baseline for the subsequent NC and no-NC experiments.

CUB-specific Phase 1 diagnostics. We recompute the diagnostics on the adapted baseline and select CUB-specific under-satisfied head pairs. Figure 10 shows the distribution of calibrated undershoot across layers. Layer 5 captures the strongest early undershoot, while Layer 20 provides a later intervention site. These comparisons assess whether the diagnostic and selective regularization remain useful after adaptation to a fine-grained domain.

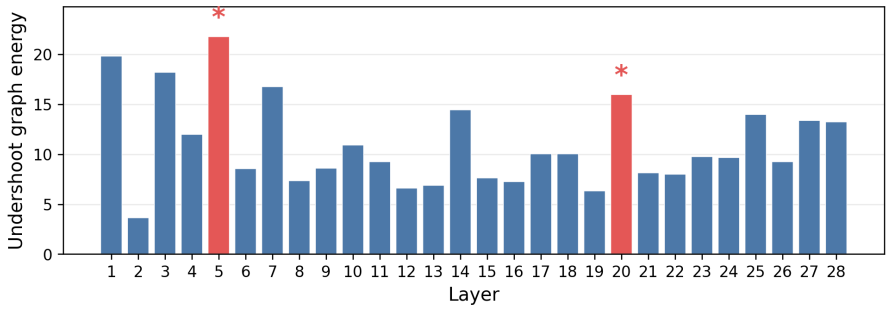


Figure 10: **CUB Phase 1 undershoot graph energy.** Graph energy summarizes the total calibrated undershoot mass in each layer. Stars indicate the CUB intervention layers used in Phase 2.

Phase 2 trajectories and selected checkpoints. We track fixed-probe NC diagnostics and CUB train-reference FID-50K during additional fine-tuning. Figure 11 shows the original checkpoint-selection trajectories. The selected checkpoints are 33k for Layer 5 and 31k for Layer 20, corresponding to 9k and 7k additional steps, respectively. The additional continuation seed uses the same checkpoint endpoints, adapted baseline, and Phase 1 pair selection. Final generation metrics are computed against the CUB test reference after checkpoint selection.

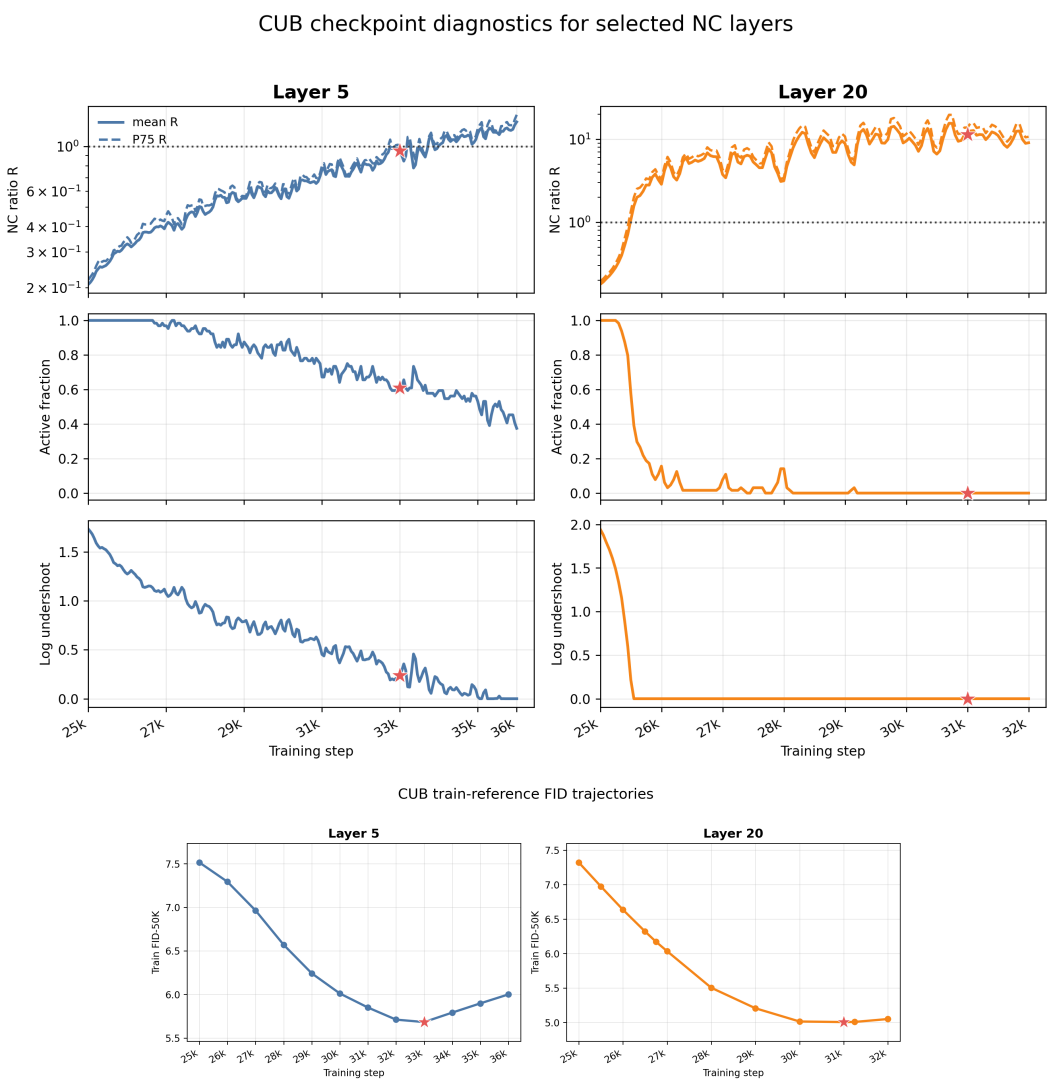


Figure 11: **CUB checkpoint-selection dynamics.** Top: fixed-probe calibrated ratio summaries, undershoot-active fraction, and log-undershoot during Phase 2 fine-tuning at Layers 5 and 20. The dotted line marks the reference value $R = 1$; stars indicate the selected checkpoints. Bottom: FID-50K against the CUB training reference along the same trajectories. Final test-reference metrics are reported separately in Table 7.

Transfer results. Table 7 reports mean±sample SD over two continuation seeds. NC reduces mean FID from 7.871 to 6.265 at Layer 5 and from 6.964 to 5.492 at Layer 20. Precision also increases, while no-NC retains higher IS and recall.

Layer 20 achieves lower final FID with fewer additional steps (7k versus 9k). Together with the stronger early undershoot at Layer 5, this indicates that undershoot magnitude alone does not determine the intervention outcome.

Table 7: CUB-200-2011 test metrics. Fine-tuned rows report mean±sample SD over continuation seeds 0 and 1; the baseline is a single run. Evaluation uses 50k DDIM-250 samples and the EMA VAE against the CUB test reference. No-NC disables the NC objective. Bold marks the better mean within each layer.

Method	Layer	FID ↓	IS ↑	Precision ↑	Recall ↑
Baseline	–	8.270	6.581	0.448	0.758
No-NC	5	7.871 ± 0.005	$\mathbf{6.542 \pm 0.005}$	0.4749 ± 0.0003	$\mathbf{0.7462 \pm 0.0011}$
NC	5	$\mathbf{6.265 \pm 0.199}$	6.331 ± 0.055	$\mathbf{0.5082 \pm 0.0054}$	0.7163 ± 0.0009
No-NC	20	6.964 ± 0.106	$\mathbf{6.465 \pm 0.029}$	0.49266 ± 0.00003	$\mathbf{0.7345 \pm 0.0079}$
NC	20	$\mathbf{5.492 \pm 0.117}$	6.242 ± 0.049	$\mathbf{0.5265 \pm 0.0051}$	0.7123 ± 0.0005

A.5 Implementation Details

All experiments use DiT-XL/2 at 256×256 resolution with the EMA VAE. For ImageNet, we start from the public ImageNet-pretrained DiT-XL/2 checkpoint. For CUB-200-2011, we first initialize from the same ImageNet-pretrained checkpoint, replace the class embedding with a 200-class embedding, and adapt the model to CUB using the standard diffusion objective.

Phase 1 diagnostics. ImageNet Phase 1 is run on the ImageNet training split with the pretrained DiT-XL/2 checkpoint frozen. We probe all 28 transformer layers using 100 diagnostic batches, 40 sampled head pairs per layer, global batch size 32, and the timestep window $[50, 250)$. The layer calibration uses the 75th percentile diagnostic scale, with fixed noise during probing.

For CUB, Phase 1 is run after the CUB baseline adaptation checkpoint at step 0024000. We again probe all layers over the timestep window $[50, 250)$, using 50 diagnostic batches, 8 sampled head pairs per layer, global batch size 32, seed 0, and the EMA VAE. Phase 1 is used only for diagnostic screening and pair selection; no generation metrics are used during this stage.

Phase 2 fine-tuning. ImageNet Phase 2 starts from the same pretrained DiT-XL/2 checkpoint for all NC and no-NC runs. We use the lower-hinge log-ratio NC objective with live backpropagated $D_{ij}^{(\ell)}$, timestep-aware weighting with a cosine schedule in $[0.25, 1.0]$, fixed-probe monitoring every 50 steps, probe batch size 8, and probe seed 1234. The selected single-layer runs use 7 Phase 1-selected head pairs; the joint Layer 10+24 run uses 14 pairs. ImageNet Phase 2 uses partial fine-tuning of the selected NC blocks together with class embedding, normalization parameters, and biases. The effective batch size is 32, implemented as global batch size 2 with 16 accumulation steps. Layers 10, 24, and 10+24 use learning rate 10^{-4} and $\lambda_{\mathrm{NC}} = 10^{-2}$; Layer 3 uses the more conservative setting 10^{-5} for both the learning rate and λ_{NC}. The no-NC controls keep the same schedules and set all NC loss weights to zero.

CUB Phase 2 starts from the CUB-adapted baseline checkpoint at step 0024000 and uses the same selected-pair NC objective. Layer 5 uses code layer 4 with 8 selected pairs, learning rate 10^{-5}, and $\lambda_{\mathrm{NC}} = 10^{-2}$. Layer 20 uses code layer 19 with 8 selected pairs, learning rate 3×10^{-5}, and $\lambda_{\mathrm{NC}} = 10^{-2}$. Both use global batch size 2, 16 accumulation steps, effective batch size 32, checkpointing every 250 steps, and the timestep window $[50, 250)$. The matched no-NC controls use the same continuation schedules but set $\lambda_{\mathrm{NC}} = 0$.

A.6 Calibration and selective regularization details

Diagnostic screening. Phase 1 reduces the search over layers and head pairs to a small set of candidate interventions. Layers are screened using fixed layer-level diagnostics, and pairs are ranked within each layer by calibrated undershoot. No generation metrics enter this selection. Graph energy identifies where undershoot is concentrated; it neither guarantees an optimal intervention nor implies a monotonic relationship with final FID. Generation metrics are used subsequently to assess the fine-tuned checkpoints.

Pair calibration. The positive scale $s_{ij}^{(\ell)}$ adjusts pair-specific reference levels without guaranteeing equal loss or gradient contributions. In online-calibrated runs, it is updated from an exponential moving average of the observed response-to-reference ratio. Updates operate in log space, are centered within each layer according to the configured mode, and are clipped to a fixed positive range. Without pair calibration, $s_{ij}^{(\ell)} = 1$. Phase 1 supplies the initial layer calibration; subsequent calibration updates depend on the run configuration.

Implementation conventions. The selected pairs are supplied explicitly to the Phase 2 launch scripts after diagnostic screening. Implementation layer indices are zero-based; paper layer numbers are obtained by adding one. The batch size B in the functional observable counts images in the NC-evaluated subset, rather than the accumulated diffusion batch. Numerical rescaling, floors, and clipping follow the implementation details specified separately.

Gradient flow and aggregation variants. Although the functional observable is detached, shared parameter updates can change it indirectly. The NC loss does not directly backpropagate through the order-probed clean-latent predictions. We additionally evaluate samplewise variants that apply the hinge before averaging over images under matched training conditions (Appendix A.7).

A.7 NC LOSS IMPLEMENTATION AND SAMPLEWISE ABLATION

Numerical implementation. The main text presents the underlying lower-hinge objective. The implementation additionally uses positive numerical floors, a common layerwise rescaling, and clipping of the positive log-gap. Let $\epsilon_{\mathrm{num}} = 10^{-8}$ and let $\zeta_\ell \geq \epsilon_{\mathrm{num}}$ be the detached scale obtained from the configured quantile of the SNR-weighted functional responses, pooled over candidate pairs and NC-evaluated images before online pair selection.

For brevity, suppress the pair and layer indices and define the stabilized log response and log target as

$$\ell_{\delta,b} = \log\left(\max\left\{\frac{\max\{\mathrm{sg}(\widetilde{\delta}_b), \epsilon_{\mathrm{num}}\}}{\zeta_\ell}, \epsilon_{\mathrm{num}}\right\}\right),$$

$$\ell_T(T) = \log\left(\frac{\max\{T, \epsilon_{\mathrm{num}}\}}{\zeta_\ell} + \epsilon_{\mathrm{num}}\right).$$

Here $\widetilde{\delta}_b$ is the normalized, SNR-weighted functional response, not the raw order-difference norm. The capped hinge is

$$\rho(a) = \left[\min\{6, \max\{0, a\}\}\right]^2.$$

The cap bounds the pairwise penalty at 36; above the cap, the pairwise penalty is saturated.

Original and samplewise objectives. We compare three reductions of the same pairwise objective. The original method averages the log responses before applying the hinge:

$$\mathcal{L}_{\mathrm{batch}}^{ij} = \rho\left(\ell_T(T) - \frac{1}{B}\sum_{b=1}^{B}\ell_{\delta,b}\right).$$

The first ablation applies the hinge separately to each image while retaining the original batch-level routing target:

$$\mathcal{L}_{\mathrm{sample,shared}}^{ij} = \frac{1}{B}\sum_{b=1}^{B}\rho(\ell_T(T) - \ell_{\delta,b}).$$

For both objectives, $T = C_\ell s_{ij}^{(\ell)} D_{ij}^{(\ell)}$, where $D_{ij}^{(\ell)}$ is computed from the batch-mean routing matrices.

The second ablation also replaces the batch-level routing score with the score computed from each image's routing matrices:

$$T_b = C_\ell s_{ij}^{(\ell)} D_{ij,b}^{(\ell)}, \qquad \mathcal{L}_{\mathrm{sample,individual}}^{ij} = \frac{1}{B}\sum_{b=1}^{B}\rho(\ell_T(T_b) - \ell_{\delta,b}).$$

Table 8: Fixed-controller Layer 24 comparison (mean±sample SD over three continuation seeds). Both arms use 8,000 microsteps (500 optimizer updates); each final checkpoint is evaluated with 50k DDIM-250 samples against the ImageNet validation reference.

Objective	FID ↓	IS ↑	Precision ↑	Recall ↑
Fixed-controller hinge	6.850 ± 0.109	129.63 ± 1.13	0.619 ± 0.002	0.657 ± 0.002
Plain shrinkage	6.774 ± 0.114	128.45 ± 1.28	0.601 ± 0.002	0.676 ± 0.003

Thus, the first ablation changes only hinge placement, whereas the second additionally changes routing aggregation. All variants retain detached functional responses and gradients through the routing score. The controller definitions, outer weighting, and summation over selected pairs are retained.

Matched training protocol. The ablations use ImageNet at 256×256 resolution, DiT-XL/2, Layer 24, and the same seven Phase 1-selected pairs. We use three continuation seeds, the same pretrained parent, trainable parameter scope, optimizer settings, and timestep-sampling policy. The recorded NC window is $[80, 180)$. Each arm receives 8,000 microsteps with microbatch size two and 16-step gradient accumulation, giving 500 optimizer updates. The comparison uses the predeclared final checkpoint rather than selecting a separate best-FID checkpoint for each arm.

All arms follow the same pilot-to-continuation restart at microstep 256. These are matched resumed experiments, not bitwise replays of the historical uninterrupted training stream. Generation evaluation uses the common 50k-sample DDIM-250 protocol and reports FID, Inception Score, precision, and recall.

Interpretation of batch size. Here B denotes the NC-evaluated subset size, not the accumulated diffusion batch size. When $B = 1$, all three formulations coincide. Gradient accumulation does not combine images before the hinge. We therefore examine actual NC subset sizes and use separate frozen-model, multi-image audits to assess whether aggregation masks individual violations. Because the implemented hinge is capped, we do not assume a universal ordering between the batchwise and samplewise loss values.

Calibrated hinge versus plain shrinkage. We test whether conditional, response-referenced regularization offers benefits over unconditional commutator suppression. Both arms use the same seven selected Layer 24 pairs, fixed controllers, and matched training schedules. Plain shrinkage uses

$$\mathcal{L}_{\text{shrink}} = \kappa \sum_{(i,j)\in\mathcal{P}_{24}} \left(D_{ij}^{(24)}\right)^2,$$

where $D_{ij}^{(24)}$ is the normalized routing commutator score. We calibrate κ once on 16 training microbatches to match initial RMS parameter-gradient strength, then fix it across seeds. Both arms freeze the layer calibration, use unit pair scales, disable adaptive controllers and response-based guards, and retain identical outer and timestep weighting.

The hinge uses detached functional responses: its active, unsaturated penalty encourages smaller routing commutators but stops penalizing a pair once the calibrated constraint is satisfied, potentially at nonzero D. Plain shrinkage instead penalizes nonzero D unconditionally. Functional probes are retained for diagnostics in both arms but do not guide the shrinkage gradient.

These experiments address two questions. First, does NC regularization improve generation beyond ordinary fine-tuning? At Layer 24, the selected NC model achieves better mean FID, IS, and precision than the no-NC control, with lower recall (Table 2).

Second, does conditional suppression offer benefits over unconditional shrinkage? In the separate comparison under matched fixed controllers, the hinge achieves higher mean IS and precision, whereas shrinkage achieves slightly lower mean FID and higher recall (Table 8). The first comparison supports benefits beyond ordinary denoising fine-tuning, while the second reveals a precision–coverage trade-off between the two regularizers.