Interpreting the Synchronization Gap: The Hidden Mechanism Inside Diffusion Transformers

Emil Albrychiewicz [*1 2] Andrés Franco Valiente [*1 2 3] Li-Ching Chen [4] Viola Zixin Zhao [5]

Abstract

In this paper, we investigate how the synchronization gap is mechanistically realized within pretrained Diffusion Transformers (DiTs) by constructing an explicit architectural realization of replica coupling by embedding two generative trajectories into a joint token sequence, modulated by a symmetric cross attention gate with variable coupling strength g. Through a linearized analysis of the attention difference, we show that the replica interaction decomposes mechanistically and we empirically validate our theoretical framework on a pretrained DiT-XL/2 model by tracking commitment and per layer internal mode energies. Our results reveal that: (1) the synchronization gap is an intrinsic architectural property of DiTs that persists even when external coupling is turned off; (2) as predicted by our spatial routing bounds, the gap completely collapses under strong coupling $g \to 1$; (3) the gap is strictly depth localized, emerging sharply only within the final layers of the Transformer; and (4) global, low frequency structures consistently commit before local, high frequency details. Ultimately, our findings provide a mechanistic interpretation of how Diffusion Transformers resolve generative ambiguity, isolating speciation transitions to the terminal layers of the network.

[1]Leinweber Institute for Theoretical Physics and Department of Physics,
University of California, Berkeley, CA, 94720-7300, USA [2]Theoretical Physics Group, Lawrence Berkeley National Laboratory
Berkeley, CA 94720-8162, USA [3]Department of Radiation Oncology, University of California, San Francisco [4]Computational Precision Health, University of California, San Francisco [5]Department of Physics,
University of California, Berkeley, CA 94720, USA. Correspondence to: Emil Albrychiewicz <ealbrych@berkeley.edu>, Andrés Franco Valiente <andresfranco@berkeley.edu>, Li-Ching Chen < liching_chen@berkeley.edu>, Viola Zixin Zhao <zhaozixin@berkeley.edu>.

Proceedings of the AI4Physics Workshop at the 43rd International Conference on Machine Learning (AI4Physics@ICML 2026), Seoul, South Korea. 2026. Copyright 2026 by the author(s).

1. Introduction

Diffusion models generate data by learning to reverse a stochastic noising process, progressively transforming Gaussian noise into structured samples (Sohl-Dickstein, Weiss, Maheswaranathan, and Ganguli, 2015; Ho, Jain, and Abbeel, 2020; Song, Sohl-Dickstein, Kingma, Kumar, Ermon, and Poole, 2020; Lai, Song, Kim, Mitsufuji, and Ermon, 2025). Among current architectures, Diffusion Transformers (DiTs) (Peebles and Xie, 2023) have emerged as the foundational standard for generative modeling. By replacing the rigidly structured convolutional U-Net (Ronneberger, Fischer, and Brox, 2015) with a highly scalable sequence of Transformer blocks (Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, and Polosukhin, 2017) that operate on modality agnostic patchified latent tokens (Rombach, Blattmann, Lorenz, Esser, and Ommer, 2022), DiTs have rapidly expanded beyond their origins in high fidelity image synthesis. However, despite this unprecedented empirical success, the internal mechanisms by which these models resolve generative ambiguity when transitioning from unstructured noise to specific, coherent representations remains poorly understood from an interpretability perspective.

A recent line of work based on nonequilibrium statistical physics (Raya and Ambrogioni, 2023; Biroli, Bonnaire, Bortoli, and Mézard, 2024; Kamb and Ganguli, 2024; Sclocchi, Favero, and Wyart, 2025) has begun to address these interpretability questions by identifying sharp dynamical transitions in the reverse generative process. In (Biroli et al., 2024), the authors established that diffusion models trained on structured data distributions undergo two macroscopic phase transitions, a speciation time, at which the trajectory commits to a particular data mode, and a col-

lapse time, at which it locks onto a specific training example. These transitions have since been characterized through replica analysis of Gaussian mixtures (Biroli, Bonnaire, Bortoli, and Mézard, 2024), extended to general class structures via free entropy criteria (Achilli, Benedetti, Biroli, and Mézard, 2026), and connected to entropic signatures observable in trained models (Handke, Stančević, Koulischer, Demeester, and Ambrogioni, 2026). In the setting of coupled multimodal generation, (Albrychiewicz et al., 2026) showed that modeling the interaction between two coupled diffusion trajectories as a pair of Ornstein–Uhlenbeck (OU) processes (Uhlenbeck and Ornstein, 1930) reveals a synchronization gap. In the case of symmetric coupling, this is a temporal window during which the common eigenmode has speciated while the difference eigenmode has not. This gap arises from a spectral hierarchy in the interaction timescales and has been shown to depend on the coupling strength g. The coupled OU processes approach has also been recently extended by (Lu & Tang, 2026) for breaking detailed balance which can be used to accelerate generative process.

These theoretical results, however, are formulated entirely in the language of continuous stochastic processes with analytically tractable score functions. In contrast, a pretrained Diffusion Transformer is not analytically tractable. It is a deep, discrete residual network in which the score function is implicitly defined by the composition of attention layers, pointwise nonlinearities, and adaptive normalization modules. The central question motivating this work is how the synchronization gap phenomenology is realized in the architecture of a Diffusion Transformer, and what mechanism is responsible for its existence?

We answer this question both theoretically and empirically. On the theoretical side, we construct an explicit mapping from the coupled OU system into the self-attention mechanism of a pretrained DiT. By embedding two generation trajectories into a single token sequence and introducing a blockwise normalized, symmetric attention gate which depends on the coupling strength g, we obtain a controlled architectural realization of replica coupling. Linearizing the resulting difference in attention output around the symmetric state, where two replicas are equivalent, we decompose it into two mechanistically distinct terms. The first term, which we refer to as a spatial routing, is the one where unperturbed attention kernel transports a perturbed value signal across token positions, and in the second, the perturbation enters through the softmax Jacobian of the attention weights themselves. We show that these two channels are suppressed by different functions of the coupling strength, $\frac{1-g}{1+g}$ and $\frac{1}{1+g}$, respectively, and argue that the first term is dominant for low frequency modes of replicas difference.

To move beyond the linear regime and to determine a speciation time, we model the local distribution of the replicas difference modes as a symmetric two component Gaussian mixture, following the approach of (Biroli et al., 2024; Albrychiewicz et al., 2026) adapted to the discrete block structure. Projecting the resulting fixed point onto empirical eigenmodes of the initial difference covariance yields a scalar self consistency condition for each mode, with a modewise speciation parameter that decomposes into an attention gated signal-to-noise ratio (SNR). The synchronization gap, the difference in speciation times between leading and trailing modes is then shown to scale as $\mathcal{O}(\frac{1-g}{1+g})$ under an assumption that spatial routing term dominates, predicting its shrinkage at strong coupling.

On the empirical side, we test these predictions on a pretrained DiT-XL/2 model (Peebles and Xie, 2023) using two complementary experimental protocols. The first protocol (Appendix E.2) measures when the model behaviorally commits by coupling two replicas for an initial steps and then letting them evolve independently, measuring the agreement of the final decoded images through feature space cosine similarity (using pretrained ResNet-50 encoder (He, Zhang, Ren, and Sun, 2016)) and scale dependent pixel discrepancies. The second protocol (Appendix E.3), with a sweep across all 28 Transformer layers, identifies where this commitment is represented internally by tracking the empirical mode energies of the hidden state replica difference, evaluated at the speciation time identified by the first protocol.

Our main empirical findings are:

1. The synchronization gap exists in DiT even without coupling. At $g = 0$, the Protocol II experiment reveals a clear separation between leading and trailing modes energies concentrated in the final layers of Transformer c.f., Figure 1.

2. The gap collapses in the strong coupling region as predicted in theory section c.f., Figure 5a–Figure 5c. As g increases from 0 to 1, the internal leading and trailing mode separation is progressively suppressed which is consistent with theoretical prediction.

3. The gap is depth localized. In the weakly coupled regime, the synchronization gap is near zero in early layers and exhibits a transient texture inversion in middle layers, where local features

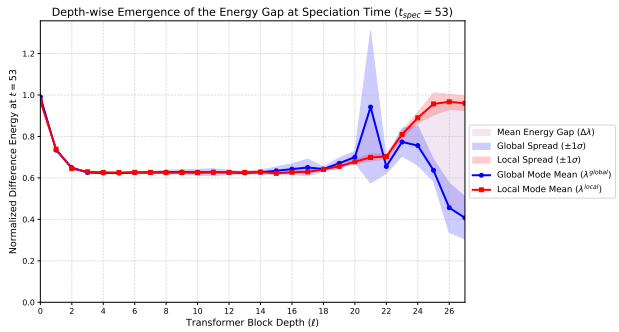


Figure 1. For coupling strength $g = 0$ and speciation step $s = 53$, we sweep across all Transformer layers to evaluate the normalized fixed basis energies of leading and trailing internal difference mode. Even with coupling turned off, the gap is present at deep layers of DiT.

temporarily stabilize before global ones. The gap emerges sharply only within the last $\approx$ 5 layers. This identifies the terminal layers as the site where the network performs frequency based routing.

4. Global structure commits before local detail. The scale dependent probe of Protocol I confirms that low frequency image structure stabilizes substantially earlier than high frequency details across all tested coupling strengths g c.f., Figure 4a–Figure 4f.

2. Theoretical Framework

We develop an effective theoretical framework connecting the continuous statistical physics of coupled diffusion processes to the discrete architecture of Diffusion Transformers. The goal is not to claim an exact equivalence between the OU description and a pretrained transformer, but rather to extract a testable mechanistic hypothesis. The synchronization gap phenomenology predicted in coupled diffusion theory is implemented in DiTs through spatially selective routing in self-attention, and can be revealed by controlled replica coupling. In contrast to the case discussed in (Albrychiewicz et al., 2026), the synchronization gap that we consider is a temporal window between the commitment times of distinct projected modes of the replica difference channel.

2.1. Coupled Reverse Diffusion: Motivation and Limitations of the Linear Theory

Unless stated otherwise, we will consider the diffusion process as taking values in the latent space of dimension d_z. The variance preserving (VP) forward diffusion process is an OU process which has an Itô SDE realization as

$$dz_t = -\frac{1}{2}\beta_t z_t dt + \sqrt{\beta_t} dW_t, \tag{1}$$

where $z_0 \sim p_{\text{data}}$, β_t is the noise schedule and W_t a standard Wiener process in $\mathbb{R}^{d_z}$. From a physics perspective, this SDE describes the overdamped Langevin dynamics of a particle coupled to a heat reservoir. Under this process, the conditional forward marginal given z_0 is

$$q(z_t \mid z_0) \sim CN(\alpha_t z_0, \sigma_t^2 I), \tag{2}$$

where

$$\alpha_t = e^{-\frac{1}{2}\int_0^t \beta_s ds}, \quad \text{and} \quad \sigma_t^2 = 1 - \alpha_t^2. \tag{3}$$

Following (Albrychiewicz et al., 2026), we study two replicas $z_t^A, z_t^B \in \mathbb{R}^{d_z}$ which are two realizations/trajectories of an identical stochastic process. Here, the superscript A, B are distinct labels, not a free index. These replicas are constructed in such a way that their reverse time dynamics are coupled with a symmetric relaxation matrix with a strength $g \geq 0$. In the continuous OU approximation, the coupled reverse process is

$$dz_t^A = \left[f(z_t^A, t) + g(z_t^A - z_t^B)\right] dt + \sqrt{\beta_t} d\bar{W}_t^A, \tag{4}$$
$$dz_t^B = \left[f(z_t^B, t) + g(z_t^B - z_t^A)\right] dt + \sqrt{\beta_t} d\bar{W}_t^B, \tag{5}$$

where the reverse drift is

$$f(z_t, t) = -\frac{1}{2}\beta_t z_t - \beta_t \nabla_{z_t} \log p_t(z_t) \tag{6}$$

and $s(z_t, t) = \nabla_{z_t} \log p_t(z_t)$ is the score. The reverse process is initialized at the terminal time $t = T$ from the stationary Gaussian distribution $z_T^A, z_T^B \sim \mathcal{N}(0, I)$ with the convention that the time orientation flows backwards $dt < 0$. We assume that the reverse Wiener processes $\bar{W}_t^A$ and $\bar{W}_t^B$ are independent white noise.

Similarly to the approach used in (Albrychiewicz et al., 2026), we introduce common and difference modes

$$u_t = \frac{z_t^A + z_t^B}{\sqrt{2}}, \qquad v_t = \frac{z_t^A - z_t^B}{\sqrt{2}}. \tag{7}$$

However, in this case, we cannot use the exact score and these modes do not fully decouple the system given by (4)-(5). Nevertheless, at the beginning of the reverse process, when noise level is high, we can approximate the marginal. If one approximates the marginal score by a single Gaussian $\mathcal{N}(0, \Sigma_t)$ then the score is linear

$$\nabla_z \log p_t(z) \approx -\Sigma_t^{-1} z, \tag{8}$$

in that regime, the dynamics decouple

$$\mathrm{d}u_t = \left[-\tfrac{1}{2}\beta_t\, u_t + \beta_t\, \Sigma_t^{-1}\, u_t\right] \mathrm{d}t + \sqrt{\tfrac{\beta_t}{2}}\, \mathrm{d}(\bar{W}_t^A + \bar{W}_t^B)\,, \tag{9}$$

$$\mathrm{d}v_t = \left[-\tfrac{1}{2}\beta_t\, v_t + \beta_t\, \Sigma_t^{-1}\, v_t + 2g\, v_t\right] \mathrm{d}t + \sqrt{\tfrac{\beta_t}{2}}\, \mathrm{d}(\bar{W}_t^A - \bar{W}_t^B)\,. \tag{10}$$

The common mode u_t evolves identically to a single uncoupled trajectory. The difference mode v_t acquires an additional restoring term $2g\, v_t$ that accelerates its drift toward zero mean hence acting as a restoring force that pulls the replicas together.

The linear theory predicts a hierarchy of damping rates across directions in latent space. However, by itself it does not identify those directions with spatial frequencies, nor can it generate branch formation. As long as the linearized operator remains stable, the unique fixed point for the difference mode is $v = 0$ [1]. Hence the linear Gaussian theory is best understood as a motivation for a difference channel hierarchy, not yet as a complete account of speciation.

This motivates two extensions: an architectural realization of symmetric coupling inside DiT self-attention, and the use of the simplest non-linear marginal probability that breaks the linear score assumption to evaluate stable branch formation in the difference channel.

2.2. Discrete DiT Model Architecture

To parametrize the reverse dynamics, we use the DiT architecture. For a replica pair (A, B), we concatenate the two token sequences along the token dimension $X_t = [X_t^A; X_t^B] \in \mathbb{R}^{2N \times d_{\text{model}}}$.

For each attention head, let Q, K, and V denote the corresponding query, key, and value matrices, and let d_h be the per-head hidden dimension. The pre-softmax attention logits are

$$S = \frac{QK^{\intercal}}{\sqrt{d_h}}. \tag{11}$$

Under the replica concatenation, these decompose into block form as

$$S = \begin{bmatrix} S_{AA} & S_{AB} \\ S_{BA} & S_{BB} \end{bmatrix}, \qquad V = \begin{bmatrix} V_A \\ V_B \end{bmatrix}. \tag{12}$$

Here S_{AA} and S_{BB} are the intra-replica attention logits, while S_{AB} and S_{BA} are the inter-replica logits. This decomposition will allow us to introduce attention gating by modulating the inter-replica interactions separately from the intra-replica ones. To ensure a clean baseline, we compute attention weights with separate row wise softmax normalizations within each block.

We realize the continuous coupling strength g through a normalized mixture of intra replica and inter replica attention outputs

$$Attn_g(X) = \frac{1}{1+g}\left(\underbrace{\begin{bmatrix} A_{AA}\, V_A \\ A_{BB}\, V_B \end{bmatrix}}_{\text{intra}} + g\, \underbrace{\begin{bmatrix} A_{AB}\, V_B \\ A_{BA}\, V_A \end{bmatrix}}_{\text{inter}}\right). \tag{13}$$

This construction satisfies three essential properties: (i) at $g = 0$, the replicas are exactly decoupled; (ii) the coupling is perfectly symmetric under $A \leftrightarrow B$ exchange; and (iii) the weights $1/(1+g)$ and $g/(1+g)$ sum to 1, the mixture is normalized, so the residual stream scale does not grow trivially with g. This ensures that the pretrained network operates safely within its learned dynamic range. For stability, we keep g in range $[0, 1]$. Equation (13) is our proposed architectural realization of symmetric coupling in a pretrained DiT.

2.3. Linearized Analysis of the Attention Difference

Given the above DiT architecture, we consider the per layer response to a symmetry breaking perturbation h_ℓ of this symmetric state, denotes as $H_{\ell,0}^A = H_{\ell,0}^B$.

We then consider the deformed output of the attention mechanism

$$H_\ell^A = H_{\ell,0} + \frac{h_\ell}{\sqrt{2}}, \qquad H_\ell^B = H_{\ell,0} - \frac{h_\ell}{\sqrt{2}}. \tag{14}$$

To first order in $\|h_\ell\|$, the contribution of the coupled attention layer to the perturbation is

$$\mathrm{Attn}_{g,\ell}^A - \mathrm{Attn}_{g,\ell}^B = \frac{1-g}{1+g}\, 2\, A_{\ell,0}\, \delta h_\ell + \frac{1}{1+g}\, 2\big(\delta A_\ell^{(+)} + g\, \delta A_\ell^{(-)}\big)\, V_{\ell,0} \;+\; \mathcal{O}(\|h_\ell\|^2)\,, \tag{15}$$

which is derived in details in Appendix A. This attention difference decomposes the linear response into two mechanistically distinct first order pathways i.e., one through the value map and one through the attention kernel.

The first term, $A_{\ell,0}\, \delta h_\ell$, is a value perturbation routed by the unperturbed attention kernel. The symmetric attention pattern $A_{\ell,0}$ is held fixed, while the perturbation perturbs the value vectors, and the resulting signal is transported across token positions. Consequently, we will occasionally call this the spatial routing term. In other words, this is the contribution from

[1] For later discussion we suppress index t for clarity.

the value path with the attention map frozen. Its prefactor $(1-g)/(1+g)$ decreases with g and vanishes at $g=1$, so strong inter-replica coupling suppresses this channel completely.

The second term, $\big(\delta A_\ell^{(+)} + g\,\delta A_\ell^{(-)}\big)V_{\ell,0}$, is a perturbation of the attention kernel. Here the perturbation enters through the query key softmax pathway, changing the attention weights themselves, which are then applied to the background values $V_{\ell,0}$. This is the Jacobian contribution from the attention pattern modulation channel. Unlike the routing term, it is suppressed only by $1/(1+g)$ and therefore remains nonzero even at $g=1$. Thus increasing g shifts the leading linear response from a fixed kernel routing of a perturbed value signal towards a perturbation of the attention kernel itself.

2.4. The Linearized Propagator

To further study the contribution of the coupled attention layer to the attention difference (15), we introduce the per-layer linear operators:

$$R_\ell = 2\,A_{\ell,0}\,J_{\ell,V}\,, \quad P_\ell(g)\,h_\ell = 2\big(\delta A_\ell^{(+)} + g\,\delta A_\ell^{(-)}\big)\,V_{\ell,0}\,, \tag{16}$$

where $J_{\ell,V}$ is the effective Jacobian absorbing the per-layer normalization and value projection. We also define the gating functions:

$$\rho(g) = \frac{1-g}{1+g}\,, \qquad \xi(g) = \frac{1}{1+g}\,. \tag{17}$$

Linearizing the MLP around the symmetric state, the full one block difference mode update incorporates both the pointwise MLP Jacobian $J_{\ell,0}^{MLP}$ and the attention contributions. Dropping $\mathcal{O}(\epsilon^2)$ cross-terms (see Appendix C for the full expansion), we obtain the effective per-layer Jacobian of the difference mode update around the symmetric state:

$$K_g \approx \mathbb{I}_\ell + J_{\ell,0}^{MLP} + \rho(g)\,R_\ell + \xi(g)\,P_\ell(g)\,. \tag{18}$$

The linearized propagator K_g describes how the perturbation is transported through the network, but a linear map around a single fixed point cannot produce speciation, it can only damp or amplify the difference. To derive a speciation criterion, we require a nonlinear score with multiple attractors. Therefore, we model the local marginal distribution of the perturbation at reverse step s and layer ℓ using the simplest nontrivial case which is a symmetric two component Gaussian mixture (Biroli et al., 2024)[2]

$$p_{s,\ell}(h) = \tfrac{1}{2}\,\mathcal{N}\big(h;\,+m_{s,\ell},\,C_{s,\ell}\big) + \tfrac{1}{2}\,\mathcal{N}\big(h;\,-m_{s,\ell},\,C_{s,\ell}\big)\,, \tag{19}$$

where $m_{s,\ell} \in \mathbb{R}^D$ is the branch separation vector and $C_{s,\ell} \succ 0$ is the local strictly positive definite covariance matrix. The corresponding score provides the necessary cubic like bifurcation structure that admits nontrivial fixed points

$$\nabla_h \log p_{s,\ell}(h) = -C_{s,\ell}^{-1}\,h + C_{s,\ell}^{-1}\,m_{s,\ell}\,\tanh\big(m_{s,\ell}^\top C_{s,\ell}^{-1}\,h\big)\,. \tag{20}$$

We model the deterministic one block update by combining the linearized propagator K_g with a local score gain $\gamma_{s,\ell} > 0$, yielding:

$$h^+ = K_g\,h + \gamma_{s,\ell}\Big[-C_{s,\ell}^{-1}\,h + C_{s,\ell}^{-1}\,m_{s,\ell}\,\tanh\big(m_{s,\ell}^\top C_{s,\ell}^{-1}\,h\big)\Big]\,. \tag{21}$$

To find the local stability criterion for this map, such that the bifurcation point where the symmetric solution $v=0$ becomes locally unstable and branch formation begins, we solve for the fixed point (see Appendix D for details)

$$\Big[(\mathbb{I} - K_g) + \gamma_{s,\ell}\,C_{s,\ell}^{-1}\Big]\,v = \gamma_{s,\ell}\,C_{s,\ell}^{-1}\,m_{s,\ell}\,\tanh\big(m_{s,\ell}^\top C_{s,\ell}^{-1}\,v\big)\,. \tag{22}$$

This is the discrete, attention-gated analogue of the continuous fixed point in symmetric coupled OU analysis of (Albrychiewicz et al., 2026).

2.5. Modewise Signal-to-Noise Ratio Formula

We now present a mean field theory to determine the layer-wise bifurcation point as a function of the coupling strength g. We project the dynamics onto fixed empirical modes. Let $C_{0,\ell}^{\text{emp}} = \frac{1}{M}V_{0,\ell}^\top V_{0,\ell} \in \mathbb{R}^{D\times D}$ be the empirical covariance of the M hidden state difference vectors at the initial reverse step $s=0$ $(t=T)$ and layer ℓ, where where $V_{0,\ell} \in \mathbb{R}^{M\times D}$ stacks the M sampled difference vectors row wise and D is the dimension of each difference vector at layer ℓ . We obtain its leading eigenvectors $\{r_k^{(\ell)}\}_{k=1}^K$ via the Nyström method on the dual Gram matrix $G_{0,\ell} = \frac{1}{M}V_{0,\ell}V_{0,\ell}^\top$. For any reverse step s, we define the scalar modal projections

$$c_k = r_k^\top C_{s,\ell}\,r_k\,, \quad m_k = r_k^\top m_{s,\ell}\,, \quad \eta_k(g) = r_k^\top K_g\,r_k\,, \tag{23}$$

where we suppress the (s,ℓ) indices on the left hand side for brevity.

[2] An extensions to multiple classes was introduced in (Achilli et al., 2026).

Under a mean field approximation, we set $v = a_k\, r_k$ to derive an independent scalar self consistency equation for each mode. This decoupling assumes the empirical modes $\{r_k\}$ orthogonally diagonalize $C_{s,\ell}$ enforced by our construction in Section 3 and the branch separation vector $m_{s,\ell}$ aligns with this principal subspace. The argument of the tanh nonlinearity then collapses to a scalar order parameter $u_k = m_{s,\ell}^\top C_{s,\ell}^{-1}\, v = a_k(m_k/c_k)$. Because the tanh series lacks even powers, crossmode nonlinearities enter strictly at $\mathcal{O}(\|v\|^3)$. Furthermore, linear mode mixing from off diagonal elements of K_g is practically negligible since $\{r_k\}$ span the dominant variance directions. Thus, near the symmetric bifurcation point ($|u_k| \ll 1$), we can safely treat the leading modes independently.

We expand the projected one block gain $\eta_k(g)$:

$$\eta_k(g) = 1 + \lambda_k^{MLP} + \rho(g)\,\chi_k + \xi(g)\,\pi_k\,, \tag{24}$$

where $\lambda_k^{MLP} = r_k^\top J_{MLP}\, r_k$, $\chi_k = r_k^\top R\, r_k$, and $\pi_k = r_k^\top P_g\, r_k$ represent the pointwise MLP, frequency selective spatial routing, and pattern modulation contributions, respectively. Assuming $C_{s,\ell}$ is diagonal in the $\{r_k\}$ basis, the fixed point equation (22) simplifies to:

$$u_k = \kappa_{v,k}(s,\ell;\, g)\ \tanh(u_k)\,, \tag{25}$$

with the modewise speciation parameter defined as:

$$\kappa_{v,k}(s,\ell;\, g) = \frac{\gamma_{s,\ell}\, m_k^2}{c_k\Big((1-\eta_k(g))\, c_k + \gamma_{s,\ell}\Big)}\,. \tag{26}$$

Equation (25) exhibits a standard mean field bifurcation: for $\kappa_{v,k} \le 1$, the only solution is $u_k = 0$ (replicas remain identical), whereas for $\kappa_{v,k} > 1$, two symmetric nonzero solutions emerge (replicas commit to distinct branches). The modal speciation time $s_{spec}^{(k)}$ is strictly defined by the critical condition:

$$\kappa_{v,k}\big(s_{spec}^{(k)},\, \ell;\, g\big) = 1\,. \tag{27}$$

Following (Albrychiewicz et al., 2026), we factor the speciation parameter into a score gain and a signal-to-noise ratio, $\kappa_{v,k} = \gamma_{s,\ell}\ SNR_{v,k}$. Substituting $\mu_k = 1/c_k$ as the modal precision, the expanded SNR is:

$$SNR_{v,k}(s,\ell;\, g) = \frac{m_k^2\, \mu_k^2}{\gamma_{s,\ell}\mu_k - \lambda_k^{MLP} - \rho(g)\,\chi_k - \xi(g)\,\pi_k}\,. \tag{28}$$

This formulation remains well defined provided the denominator is strictly positive ($\eta_k(g) < 1 + \gamma_{s,\ell}/c_k$). If this condition is violated, the linearized difference mode amplifies faster than the score can restore it, breaking the perturbative mean field picture.

To obtain an analytic theory, we introduce a phenomenological, empirically testable closure where the branch separation propagates multiplicatively: $m_{k,s,\ell} = G_{v,k}(s,\ell;\, g)\, m_{k,\text{init}}$, with the cumulative gain:

$$G_{v,k}(s,\ell;\, g) = \prod_{(s',\ell')\prec(s,\ell)} \eta_{k,s',\ell'}(g)\,. \tag{29}$$

Here, $(s',\ell') \prec (s,\ell)$ denotes blocks processed earlier in the reverse trajectory. The propagated SNR becomes:

$$SNR_{v,k}(s,\ell;\, g) = \frac{G_{v,k}(s,\ell;\, g)^2\, m_{k,\text{init}}^2}{c_{k,s,\ell}\Big((1-\eta_{k,s,\ell}(g))\, c_{k,s,\ell} + \gamma_{s,\ell}\Big)}\,. \tag{30}$$

This is the discrete counterpart of the continuous SNR formula. The OU processes eigenvalues are replaced by the learned attention routing gains χ_k, and the coupling enters through $\rho(g)$, $\xi(g)$ and the cumulative discrete gain $G_{v,k}^2(g)$.

2.6. The Synchronization Gap and Its Collapse

We will now focus on deriving the central experimental observable that we are interested in i.e., the synchronization gap between global and local difference modes.

Let k_{hi} denote a leading mode associated with coarse, globally organized structure and k_{lo} a trailing mode associated with finer detail. In the coherent low frequency regime characterized in Appendix B, we assume the following structural properties. First, the spatial routing contribution dominates the pattern modulation correction,

$$|\pi_k| \ll |\chi_k|\,, \tag{31}$$

we provide a structural reason why this is valid in Appendix B. Next, the MLP term is only weakly mode dependent as we discussed in Section 2.4

$$\lambda_k^{MLP} \approx \lambda^{MLP} \quad . \tag{32}$$

Finally, as a modeling hypothesis consistent with the routing dominant regime, we assume that attention routes leading coarse modes more strongly than trailing fine modes,

$$\chi_{k_{\text{hi}}} > \chi_{k_{\text{lo}}}, \tag{33}$$

from RG theory (Wilson, 1975) perspective this can be interpreted as a coarse graining operator preserving k_{hi} (long wavelength) while heavily suppresses k_{lo} (respectively short wavelength). Under these structural properties the propagator (24) simplifies to

$$\eta_k(g) \approx 1 + \lambda^{MLP} + \rho(g)\,\chi_k\,, \tag{34}$$

and the routing dominant SNR becomes

$$SNR_{v,k}(s,\ell;\, g) \approx \frac{m_k^2}{c_k\Big((-\lambda^{MLP} - \rho(g)\,\chi_k)\,c_k + \gamma_{s,\ell}\Big)}\,. \tag{35}$$

If the remaining modal factors (m_k, c_k) are comparable, then $SNR_{v,k_{\rm hi}} > SNR_{v,k_{\rm lo}}$, which implies

$$s_{spec}^{(k_{\rm hi})}(\ell;\, g) < s_{spec}^{(k_{\rm lo})}(\ell;\, g), \tag{36}$$

the global mode speciates before the local mode. The modewise synchronization gap at layer ℓ and coupling g is

$$\Delta s_v(\ell;\, g) := s_{spec}^{(k_{\rm lo})}(\ell;\, g) - s_{spec}^{(k_{\rm hi})}(\ell;\, g)\,. \tag{37}$$

A positive value indicates that the global mode has committed while the local mode remains unresolved. Using the assumptions from above, the SNR difference between modes picks up a factor of $\rho(g)$ from the routing term. To first order in the spectral split

$$SNR_{v,k_{\rm hi}}(s,\ell;\, g) - SNR_{v,k_{\rm lo}}(s,\ell;\, g) = \mathcal{O}\left(\frac{1-g}{1+g}\right)\,, \tag{38}$$

hence the spatial routing contribution to the SNR difference vanishes as $g \to 1$.

2.7. Summary of Testable Predictions

The derived theoretical framework yields four concrete empirical predictions.

1. Commitment Hierarchy: Latent low frequency structures commit to a basin of attraction significantly earlier than high frequency textures ($\tau_{\rm low} < \tau_{\rm high}$). The speciation time $\tau_{\rm spec}$ is expected to track this transition but is not required by the theory to lie strictly between them.

2. Depth Localization: Since the speciation time depends on the layer the synchronization gap will not be uniform across them.

3. Existence of the Natural Gap: At $g = 0$, fully decoupled replicas sharing an initialization z_T exhibit an intrinsic synchronization gap whose magnitude and mode dependence are predicted by the routing dominant SNR formula (35).

4. Coupling Induced Collapse: As symmetric coupling g increases toward 1, the spatial routing difference is suppressed, forcing both the internal layerwise gap and the behavioral commitment gap to collapse.

3. Empirical Evaluation

To verify the theoretical predictions of the linearized difference channel, we test a pretrained DiT-XL/2 model using two complementary experimental protocols. Protocol I probes the output space speciation time and scale dependent commitment via a decoupling intervention, while Protocol II sweeps the internal network depth to measure the layerwise synchronization gap.

(Detailed variable definitions, variance preserving initialization procedures, and exhaustive measurement metrics for both protocols are provided in Appendix E).

3.1. Protocol I: Output Speciation and Scale-Dependent Commitment

In Protocol I, two replicas are initialized with a shared macroscopic structure (a shared initial latent state and an antisymmetric noise perturbation). They are coupled with strength g up to an intervention step $t_{\rm int}$, after which they evolve independently ($g = 0$) using a stochastic DDPM sampler.

We define the speciation time $\tau_{\rm spec}(g)$ as the inflection point where the final decoded images commit to the same semantic basin, measured via feature space cosine similarity. As summarized in Table 1, $\tau_{\rm spec}(g)$ decreases monotonically as coupling strength increases. This confirms that stronger symmetric coupling suppresses spatial routing differences earlier, reducing the temporal window over which replicas can diverge.

Table 1. Summary of speciation times determined using the cosine similarity method and corresponding 95% confidence intervals versus coupling strength g.

Coupling strength g	$\tau_{\rm spec}(g)$ [95% CI]
0.1	63.6 [50.0, 96.6]
0.3	48.0 [33.7, 85.3]
0.5	46.3 [31.6, 82.6]
0.7	45.6 [35.0, 59.7]
0.9	45.0 [35.8, 56.0]
1.0	44.2 [35.6, 58.8]

To determine how spatial routing resolves generative ambiguity across scales, we decompose the image discrepancy into global (coarse) and local (fine) commitment times, denoted $\tau_{\rm g}(g)$ and $\tau_{\rm l}(g)$. The output space synchronization gap is directly quantified as:

$$\Delta\tau(g) = \tau_{\rm l}(g) - \tau_{\rm g}(g) \tag{39}$$

A positive gap formally indicates that global structures resolve before high frequency local details. Empiri-

cally, we validate the existence of this output space gap and show that it is robust: after the weak coupling regime, the gap stabilizes near $\Delta\tau \approx 39$–41 steps across all medium and strong coupling strengths (see Appendix F for full sweeps).

3.2. Protocol II: Internal Mode Stabilization and the Layerwise Gap

Protocol II provides the primary mechanistic test of our theory. Here, the replicas remain coupled at a constant g for the entire reverse trajectory. We evaluate the normalized fixed basis energies of leading and trailing internal difference modes at the speciation time $\tau_{\text{spec}}(g)$, sweeping across all 28 Transformer layers ℓ.

The internal energy gap ratio is defined as:

$$\mathcal{G}_{\text{int}}(\ell; g) = \frac{\bar{\lambda}_{\text{trail}}(\ell; g)}{\bar{\lambda}_{\text{lead}}(\ell; g)} \tag{40}$$

Values of $\mathcal{G}_{\text{int}} > 1$ indicate that trailing (local) empirical modes retain more energy than leading (global) modes at the moment of species commitment.

The depth sweeps in Figure 5a and 5c support the main theoretical claim. At weak coupling ($g = 0.1$), the internal synchronization gap is strongly depth localized; leading and trailing energies remain close through early layers but separate sharply in the final blocks. Conversely, at strong coupling ($g = 0.9$), the hierarchy collapses entirely, and the energies are nearly superposed across the full depth. This empirically validates that the effective spatial routing term is weighted by $\rho(g) = \frac{1-g}{1+g}$, suppressing the internal gap as $g \to 1$.

A second important observation is that the internal gap is concentrated in late blocks. This depth localization is also consistent with the theoretical picture in which layer dependent modulation and residual composition cause the effective difference channel Jacobian to vary strongly across depth. In particular, the results suggest that the deepest transformer blocks are where the leading/trailing hierarchy becomes relevant in the uncoupled or weakly coupled regime, while strong coupling suppresses that hierarchy before it can fully emerge.

3.3. Summary

Taken together, the results support the following picture. Strong symmetric replica coupling sharply suppresses the internal, depth localized hierarchy between leading and trailing difference modes, in agreement with the spatial routing dominant theory. Nevertheless, after coupling is removed, fine output detail still requires a longer period of synchronized evolution than coarse structure in order to survive later independent stochastic refinement. The main novel message is therefore a two level synchronization phenomenon, coupling collapses the internal hidden state hierarchy, while a residual coarse to fine output commitment lag can persist in decoded image space.

4. Conclusion

We have presented a theoretical and empirical study of the synchronization gap in Diffusion Transformers, providing a controlled framework for analyzing how pretrained DiTs resolve generative ambiguity across spatial scales. By constructing an explicit realization of symmetric replica coupling within self-attention, we decomposed the network's first order response into spatial routing and pattern modulation channels. Using a symmetric Gaussian mixture to obtain a speciation criterion, we derived a scalar self-consistency condition wherein the modewise SNR predicts the collapse of the internal synchronization gap at strong coupling, scaling as $\mathcal{O}(\frac{1-g}{1+g})$. Empirical tests on DiT-XL/2 confirmed this collapse, revealing that while moderate coupling ($g \geq 0.3$) superposes leading and trailing mode energies within the final Transformer blocks, a residual global to local commitment lag persists in the decoded image space. Taken together, these results indicate that the VAE decoder and cumulative reverse steps introduce scale dependent processing beyond the single block linearized theory.

Several limitations of the present work should be noted: the local score gain $\gamma_{s,\ell}$ was analyzed for its sign and monotonicity rather than via direct measurement, the mean field projection omits $\mathcal{O}(\|v\|^3)$ crossmode couplings, and the derivation assumes a diagonal covariance $C_{s,\ell}$ in the empirical basis. Nevertheless, this mechanistic characterization explains why recent temporal feature forecasting methods preserve global semantics while degrading local detail, providing a structural justification for stage and depth aware feature caching to preserve fine image fidelity. Ultimately, this framework provides a quantitative language for analyzing how spatial information is routed through Transformer depth, connecting the macroscopic phenomenology of phase transitions to the microscopic mechanics of self-attention.

References

Achilli, B., Benedetti, M., Biroli, G., and Mézard, M. Theory of speciation transitions in diffusion models with general class structure. arXiv preprint arXiv:2602.04404, 2026.

Albrychiewicz, E., Valiente, A. F., and Chen, L.-C. Dynamical regimes of multimodal diffusion models. arXiv preprint arXiv:2602.04780, 2026.

Biroli, G., Bonnaire, T., Bortoli, V. D., and Mézard, M. Dynamical regimes of diffusion models. Nature Communications, 15(1):9957, 2024.

Handke, F., Stančević, D., Koulischer, F., Demeester, T., and Ambrogioni, L. The entropic signature of class speciation in diffusion models. arXiv preprint arXiv:2602.09651, 2026.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778, 2016.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 33:6840–6851, 2020.

Kamb, M. and Ganguli, S. An analytic theory of creativity in convolutional diffusion models. arXiv preprint arXiv:2412.20292, 2024.

Lai, C.-H., Song, Y., Kim, D., Mitsufuji, Y., and Ermon, S. The principles of diffusion models. arXiv preprint arXiv:2510.21890, 2025.

Lu, H. and Tang, Y. Steering dynamical regimes of diffusion models by breaking detailed balance. arXiv preprint arXiv:2602.15914, 2026.

Peebles, W. and Xie, S. Scalable diffusion models with transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4195–4205, 2023.

Raya, G. and Ambrogioni, L. Spontaneous symmetry breaking in generative diffusion models. Advances in Neural Information Processing Systems, 36:66377–66389, 2023.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10684–10695, 2022.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241, 2015.

Sclocchi, A., Favero, A., and Wyart, M. A phase transition in diffusion models reveals the hierarchical nature of data. Proceedings of the National Academy of Sciences, 122(1):e2408799121, 2025.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In International Conference on Machine Learning, pp. 2256–2265, 2015.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456, 2020.

Uhlenbeck, G. E. and Ornstein, L. S. On the theory of the brownian motion. Physical Review, 36(5):823, 1930.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. Advances in Neural Information Processing Systems, 30, 2017.

Wilson, K. G. The renormalization group: Critical phenomena and the kondo problem. Reviews of Modern Physics, 47(4):773, 1975.

A. Detailed Derivation of (15)

In this appendix, we suppress the layer index ℓ. The attention output for replica A is:

$$\mathrm{Attn}_g^A = \frac{1}{1+g}\left[A_{AA}\,V_A + g\,A_{AB}\,V_B\right]. \tag{41}$$

Substituting the first order expansions:

$$A_{AA}\,V_A = (A_0 + \delta A^{(+)})(V_0 + \delta V) \tag{42}$$
$$= A_0V_0 + A_0\,\delta V + \delta A^{(+)}V_0 + \mathcal{O}(\|v\|^2)\,, \tag{43}$$
$$A_{AB}\,V_B = (A_0 + \delta A^{(-)})(V_0 - \delta V) \tag{44}$$
$$= A_0V_0 - A_0\,\delta V + \delta A^{(-)}V_0 + \mathcal{O}(\|v\|^2)\,. \tag{45}$$

Summing these terms yields:

$$\mathrm{Attn}_g^A = \frac{1}{1+g}\Big[(1+g)\,A_0\,V_0 + (1-g)\,A_0\,\delta V \tag{46}$$
$$+\big(\delta A^{(+)} + g\,\delta A^{(-)}\big)\,V_0\Big] + \mathcal{O}(\|v\|^2)\,. \tag{47}$$

By the exact exchange symmetry of the system, Attn_g^B is obtained by mapping $v \to -v$, which flips the signs of the linear perturbations ($\delta V \to -\delta V$ and $\delta A^{(\pm)} \to -\delta A^{(\pm)}$). Taking the difference $\mathrm{Attn}_g^A - \mathrm{Attn}_g^B$ cancels the base term A_0V_0 and doubles the linear terms, yielding the desired result (15).

B. Routing Dominance for Low Frequency Difference Modes

Below we provide a justification for one of a key structural assumptions that we use to analyze the synchronization gap collapse mechanism. In Section 2.6 we stated that for low frequency difference channel v modes the spatial routing term in the attention difference (15) dominates the pattern modulation term. The argument has two layers. First, we prove exact identities that follow solely from row wise softmax, and the fact that its entries are nonnegative and normalized to unity. Second, we derive a quantitative residual bound in terms of the effective attention width. As in the theoretical framework section we view token space matrices such as V_0 and δV as elements of $R^{N\times d_{\mathrm{model}}}$.

Let $\mathbf{1} \in R^N$ denote the token vector of all ones and define the orthogonal projectors

$$P_0 := \frac{1}{N}\,\mathbf{1}\mathbf{1}^\top, \qquad P_\perp := I - P_0. \tag{48}$$

Thus P_0 projects onto the subspace of constant tokens and $P_\perp$ onto the mean free tokens subspace. For any token space matrix $X \in R^{N\times d_{\mathrm{model}}}$, we have the decomposition

$$X = P_0X + P_\perp X. \tag{49}$$

$A_0 \in \mathbb{R}^{N\times N}$ is a row stochastic attention matrix produced by row wise softmax, and $\delta A^{(\pm)}$ is the first order perturbation induced by a logit perturbation δS. Since blockwise softmax normalizes each of them separately, let us denote both as δA. Then following is true, for each row i, the softmax Jacobian is

$$\frac{\partial[A]_{0,ij}}{\partial[S]_{ik}} = [A]_{0,ij}\big(\delta_{jk} - [A]_{0,ik}\big), \tag{50}$$

where comparing to the notation of Section 2.2 we suppressed channel indices for clarity. To the first order, the perturbation is

$$\delta[A]_{ij} = \sum_{k=1}^{N} \frac{\partial[A]_{0,ij}}{\partial[S]_{ik}}\,\delta[S]_{ik} \tag{51}$$
$$= [A]_{0,ij}\left(\delta[S]_{ij} - \sum_{k=1}^{N}[A]_{0,ik}\delta[S]_{ik}\right). \tag{52}$$

Summing over j gives

$$\sum_{j=1}^{N}\delta[A]_{ij} = \sum_{j=1}^{N}[A]_{0,ij}\delta[S]_{ij} \tag{53}$$
$$-\left(\sum_{k=1}^{N}[A]_{0,ik}\delta[S]_{ik}\right)\sum_{j=1}^{N}[A]_{0,ij} = 0. \tag{54}$$

Hence $\delta A\,\mathbf{1} = 0$ and therefore $\delta A = \delta A P_\perp$ along with normalization fact which implies $A_0P_0 = P_0$.

Using these results for δA we find

$$\delta A V_0 = \delta A P_\perp V_0, \tag{55}$$

This result is exact and it shows that pattern modulation can only respond to spatial variation in the common mode value field. It cannot see the constant token component at all.

Now for the spatial routing term we obtain

$$A_0\delta V = P_0\delta V + A_0P_\perp\delta V, \tag{56}$$

so this term preserves the constant token subspace, which was removed by pattern modulation term.

The structural identities above already imply a strong separation between routing and pattern modulation for coherent token space signals. However, we can provide a stricter bound on pattern modulation residuals. For row i of A_0 define effective attention index as

$$N_{\mathrm{eff}}^{(i)} := \frac{1}{\sum_{j=1}^{N} A_{0,ij}^2}, \tag{57}$$

bounded by N, equal to $N_{\text{eff}}^{(i)} = 1$ for a single token attention and $N_{\text{eff}}^{(i)} = N$ for a uniform row. Let ϵ_j denote the rows of the mean free residual $P_\perp V_0$. Then we can write explicitly

$$(\delta A V_0)_i = \sum_{j=1}^{N} \delta A_{ij}\, \epsilon_j. \tag{58}$$

Applying Cauchy–Schwarz,

$$\left\| \sum_{j=1}^{N} \delta A_{ij} \epsilon_j \right\|^2 \leq \left(\sum_{j=1}^{N} \delta A_{ij}^2 \right) \left(\sum_{j=1}^{N} \|\epsilon_j\|^2 \right). \tag{59}$$

From (51),

$$|\delta A_{ij}| = A_{0,ij} \left| \delta S_{ij} - \sum_{k=1}^{N} A_{0,ik} \delta S_{ik} \right| \leq 2 A_{0,ij} \|\delta S_i\|_\infty, \tag{60}$$

where

$$\|\delta S_i\|_\infty := \max_{1 \leq j \leq N} |\delta S_{ij}|. \tag{61}$$

Therefore,

$$\sum_{j=1}^{N} \delta A_{ij}^2 \leq 4 \|\delta S_i\|_\infty^2 \sum_{j=1}^{N} A_{0,ij}^2 = \frac{4 \|\delta S_i\|_\infty^2}{N_{\text{eff}}^{(i)}}. \tag{62}$$

and that gives per row i bound

$$\|(\delta A V_0)_i\|^2 \leq \frac{4 \|\delta S_i\|_\infty^2}{N_{\text{eff}}^{(i)}} \sum_{j=1}^{N} \|\epsilon_j\|^2. \tag{63}$$

If we sum over all rows we get a norm bound

$$\|\delta A V_0\| \leq \frac{2\sqrt{N}}{\sqrt{N_{\text{eff}}^{\min}}} \|\delta S\|_{\max} \|P_\perp V_0\|, \tag{64}$$

where the narrowest attention width is $N_{\text{eff}}^{\min} = \min_i N_{\text{eff}}^{(i)}$. Assuming that the common mode field is dominated by the constant token component

$$\|\delta A V_0\| \leq \frac{2\sqrt{N}}{\sqrt{N_{\text{eff}}^{\min}}} \|\delta S\|_{\max} \varepsilon_{V_0} \|P_0 V_0\|, \tag{65}$$

for some constant $0 \leq \varepsilon_{V_0} \leq 1$.

For the spatial routing term, the triangle inequality gives

$$\|A_0 \delta V\| = \|P_0 \delta V + A_0 P_\perp \delta V\| \geq \|P_0 \delta V\| - \|A_0 P_\perp \delta V\|. \tag{66}$$

Assuming that for low frequency a constant component is dominating we can write

$$\|P_\perp \delta V\| \leq \varepsilon_{\delta V}\, \|P_0 \delta V\|, \tag{67}$$

for some constant $0 \leq \varepsilon_{\delta V} < 1$, we refer to this condition as spatial coherency of tokens. Since the norm of $P_\perp$ is one then

$$\|A_0 P_\perp\| \leq \lambda_\perp, \tag{68}$$

where a constant $\lambda_\perp \leq ||A_0||$ and altogether for (66) we have

$$\|A_0 \delta V\| \geq \left(1 - \lambda_\perp \varepsilon_{\delta V}\right) \|P_0 \delta V\|. \tag{69}$$

Finally, we can combine (65) with (69) to obtain

$$\frac{\|\delta A V_0\|}{\|A_0 \delta V\|} \leq \frac{2\sqrt{N} \|\delta S\|_{\max}\, \varepsilon_{V_0}}{\sqrt{N_{\text{eff}}^{\min}} (1 - \lambda_\perp \varepsilon_{\delta V})} \frac{\|P_0 V_0\|}{\|P_0 \delta V\|}. \tag{70}$$

This equation identifies the regime in which spatial routing dominates pattern modulation in the attention difference. In particular, if (i) the attention rows are sufficiently diffuse so that $N_{\text{eff}}^{\min}$ is large, (ii) the common-mode value field V_0 is dominated by its constant / slowly varying token component, so that $\varepsilon_{V_0} \ll 1$, (iii) the difference mode v perturbation δV is likewise coherent in token space, so that $\varepsilon_{\delta V} \ll 1$, and (iv) the attention operator is not too expansive on the mean free token subspace, so that $1 - \lambda_\perp \varepsilon_{\delta V}$ remains bounded away from zero, then the ratio of pattern modulation to spatial routing is small

$$\frac{\|\delta A V_0\|}{\|A_0 \delta V\|} \ll 1. \tag{71}$$

Moreover, when $N_{\text{eff}}^{\min} \approx N$, the $N_{\text{eff}}^{-1/2}$ factor provides additional suppression of the pattern modulation residual. In the DiT-XL/2 setting used in our experiments, $N = 1024$, so diffuse attention heads can provide a substantial numerical suppression through the factor $N_{\text{eff}}^{-1/2}$. Thus, in the coherent low frequency regime relevant to early branch formation, the dominant coupling sensitive contribution comes from the spatial routing term rather than the pattern modulation term.

C. Algebraic Expansion of the Propagator Cross Terms

The sequential residual structure of the DiT block means that the MLP acts on the post attention hidden state $\tilde{H}_\ell$. Denoting the attention contribution to the perturbation mode v compactly as $\Delta\text{Attn}_{g,\ell} =$

$[\rho(g)\, R_\ell + \xi(g)\, P_\ell(g)]\, h_\ell + \mathcal{O}(\|h_\ell\|^2)$, the full one block difference mode update prior to truncation is:

$$h_\ell^+ = (\mathbb{I}_\ell + J_{\ell,0}^{MLP})\big(\mathbb{I}_\ell + \rho(g)\, R_\ell + \xi(g)\, P_\ell(g)\big)\, h_\ell + \mathcal{O}(\|h_\ell\|^2)\,. \tag{72}$$

Algebraically expanding the product yields $h_\ell^+ = K_g\, h_\ell + \mathcal{O}(\|h_\ell\|^2)$, where:

$$\begin{aligned} K_g = \mathbb{I}_\ell + J_{\ell,0}^{MLP} + \rho(g)\, R_\ell + \xi(g)\, P_\ell(g) \\ + J_{\ell,0}^{MLP}\big[\rho(g)\, R_\ell + \xi(g)\, P_\ell(g)\big]\,. \end{aligned} \tag{73}$$

The cross term arises because the MLP processes the attention updated residual stream. However, since the attention and MLP residual branches are gated by learned vectors initialized at zero ($\mathcal{O}(\epsilon)$), the individual contributions scale as $\mathcal{O}(\epsilon)$ while the cross term scales as $\mathcal{O}(\epsilon^2)$. Thus, the cross term is safely dropped in the main text approximation.

D. Score Expansion and Propagator Repartitioning

A potential concern is that the linearized part of the Gaussian mixture score overlaps with the propagator K_g. Expanding the mixture score around $v = 0$ gives:

$$\nabla_v \log p_{s,\ell}(v) = -\Lambda_{\text{eff}}\, v + \mathcal{O}(\|v\|^3)\,, \tag{74}$$

where the effective precision of the mixture is:

$$\Lambda_{\text{eff}} = C_{s,\ell}^{-1} - C_{s,\ell}^{-1}\, m_{s,\ell}\, m_{s,\ell}^\top\, C_{s,\ell}^{-1}\,. \tag{75}$$

One could absorb all linear in v score contributions into a redefined propagator:

$$\tilde{K}_g = K_g - \gamma_{s,\ell}\, \Lambda_{\text{eff}}\,, \tag{76}$$

leaving only the genuinely nonlinear remainder:

$$\gamma_{s,\ell}\, C_{s,\ell}^{-1}\, m_{s,\ell}\Big[\tanh\big(m_{s,\ell}^\top C_{s,\ell}^{-1}\, v\big) - m_{s,\ell}^\top C_{s,\ell}^{-1}\, v\Big] = \mathcal{O}(\|v\|^3)\,. \tag{77}$$

This decomposition is analogous to separating the free propagator from the self-energy in a diagrammatic expansion in quantum field theory, where $-\gamma_{s,\ell}\, \Lambda_{\text{eff}}$ acts as the one loop self energy correction. However, this repartitioning leaves the fixed point vector and scalar self consistency equations exactly invariant, as both are derived from the full tanh non-linearity rather than the linearization. All testable predictions depend on the total effective Jacobian at the symmetric point $(K_g - \gamma_{s,\ell}\, \Lambda_{\text{eff}})$, so the physical conclusions remain identical regardless of the partition.

E. Empirical Setup

E.1. Variance Preserving Initialization

To ensure the difference mode v measurement remains within the prior distribution, we initialize replicas with an antisymmetric perturbation:

$$z_T \sim \mathcal{N}(0, \mathbb{I})\,, \quad \delta \sim \mathcal{N}(0, \mathbb{I})\,, \quad z_T^{A,B} = \frac{z_T \pm \sigma\, \delta}{\sqrt{1+\sigma^2}} \tag{78}$$

This guarantees $\mathrm{Var}(z_T^A) = \mathrm{Var}(z_T^B) = \mathbb{I}$ while injecting a controlled difference signal $v_T = \sqrt{2}\, \sigma\, \delta / \sqrt{1+\sigma^2}$. The inter replica correlation is $\varrho(\sigma) = \frac{1-\sigma^2}{1+\sigma^2}$.

E.2. Protocol I: Speciation Time and Scale Dependent Commitment

Output agreement is measured using a pretrained ResNet-50 encoder ϕ:

$$a_{\text{feat}}(t_{\text{int}}; g) = \cos\Big(\phi(x^A(t_{\text{int}}; g)),\, \phi(x^B(t_{\text{int}}; g))\Big) \tag{79}$$

Scale dependent commitment utilizes adaptive average pooling (P) and bilinear upsampling (U) to isolate coarse (d_{low}) and fine (d_{high}) discrepancies:

$$d_{\text{low}}(t_{\text{int}}; g) = \|P(x_0^A) - P(x_0^B)\|_2^2 \tag{80}$$

$$d_{\text{high}}(t_{\text{int}}; g) = \|\big(x_0^A - UP(x_0^A)\big) - \big(x_0^B - UP(x_0^B)\big)\|_2^2 \tag{81}$$

Thus d_{low} measures discrepancy after strong spatial averaging, whereas d_{high} measures discrepancy in the corresponding high frequency residual. We extract the macroscopic τ_{g} and microscopic τ_{l} commitment times from the 50% thresholds of their respective sigmoid fits. The output space synchronization gap is directly quantified as

$$\Delta\tau(g) = \tau_{\text{l}}(g) - \tau_{\text{g}}(g). \tag{82}$$

A positive gap formally indicates that the continuous physics of the DiT resolves low frequency global structure before high frequency local detail. In what follows, the terms global and local refer specifically to this output space scale decomposition, which is distinct from the internal hidden state basis defined in Protocol II.

E.3. Protocol II: Internal Mode Stabilization and the Layerwise Gap

Our second empirical protocol probes how the continuous synchronization gap is represented internally across the discrete Transformer depth ℓ. In this setup, the replicas remain coupled at a constant g for the entire reverse trajectory.

We extract the flattened hidden state difference vector for M paired seeds to form $V_{s,\ell} \in \mathbb{R}^{M \times D}$. Using the empirical covariance $C_{s,\ell}^{\text{emp}}$ and initial eigenvalues $\lambda_k^{(0,\ell)}$, the normalized fixed basis energy is:

$$\lambda_k(s,\ell) = \frac{\|V_{s,\ell} r_k^{(\ell)}\|^2}{\|V_{0,\ell} r_k^{(\ell)}\|^2} = \frac{r_k^{(\ell)\top} C_{s,\ell}^{\text{emp}} r_k^{(\ell)}}{\lambda_k^{(0,\ell)}} \tag{83}$$

We aggregate these into leading ($\bar{\lambda}_{\text{lead}}$) and trailing ($\bar{\lambda}_{\text{trail}}$) bands based on the initial covariance ordering to form the gap ratio $\mathcal{G}_{\text{int}}$.

F. Exhaustive Empirical Sweeps

This appendix provides the complete suite of empirical measurements across all tested coupling strengths, including the speciation probe (cosine similarity), the scale-dependent decoupling commitment, and the internal depth emergence sweeps.

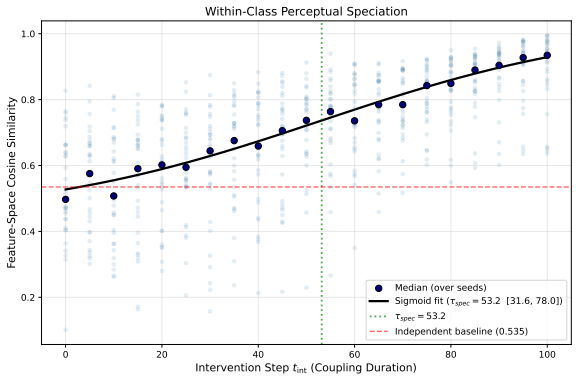
F.1. Protocol I: Figures

The following plots show the median feature space cosine similarity between paired outputs as a function of intervention step t_{int}, alongside the sigmoid fits used to extract $\tau_{\text{spec}}(g)$.

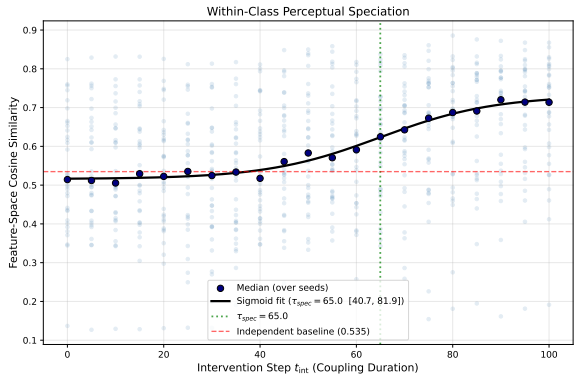
The following plots illustrate the global and local commitment scores versus coupling strength g. These curves confirm that coarse output structure stabilizes earlier than fine detail across the entire coupling range.

F.2. Protocol II: Figures

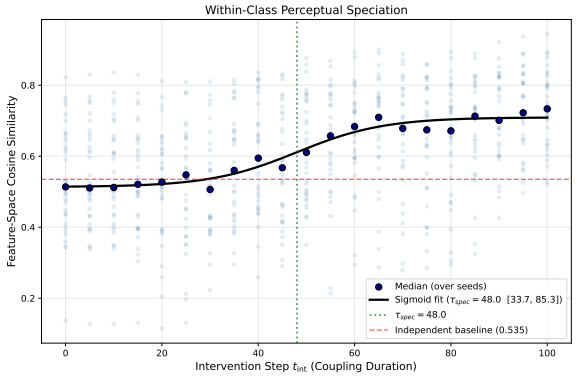
The following plots show the normalized fixed basis energies of leading and trailing internal difference modes across all 28 Transformer layers, evaluated exactly at the speciation time $\tau_{\text{spec}}(g)$.



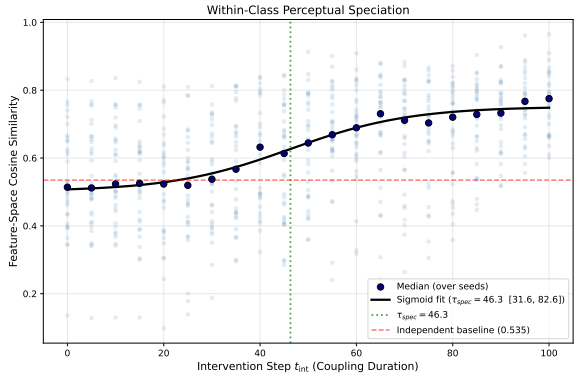
(a) $g = 0.0$



(b) $g = 0.1$

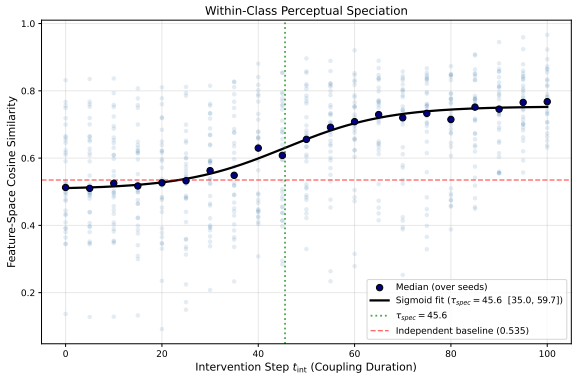


(c) $g = 0.3$

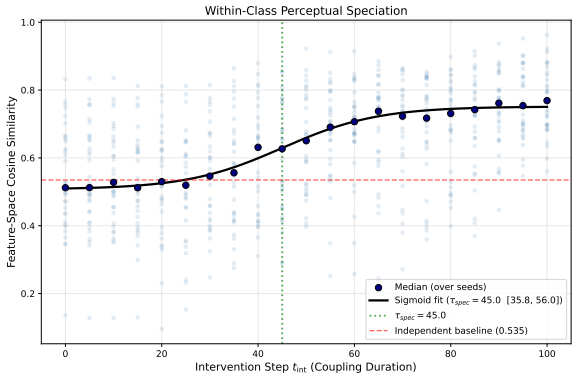


(d) $g = 0.5$

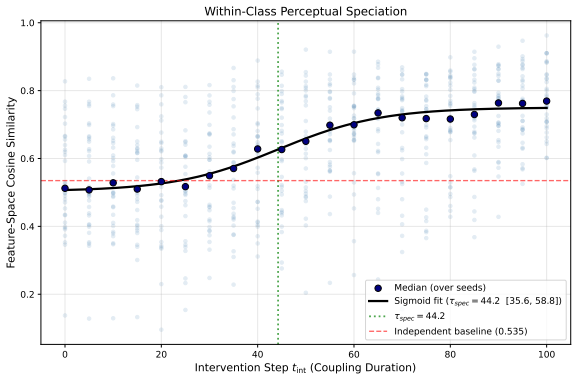
Figure 2. Exhaustive sweep of the Protocol I speciation probe (Part 1). As symmetric coupling increases, the spatial routing differences are suppressed earlier, reducing the temporal window over which replicas can diverge.



(a) $g = 0.7$

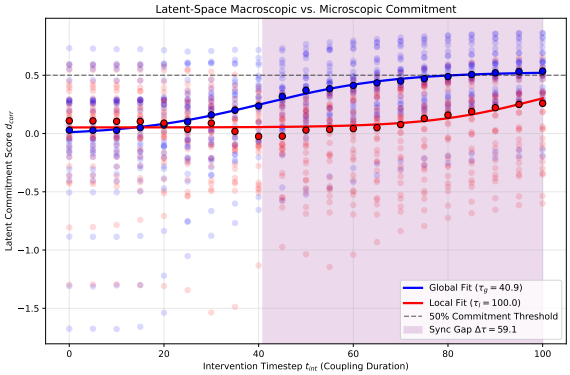


(b) $g = 0.9$

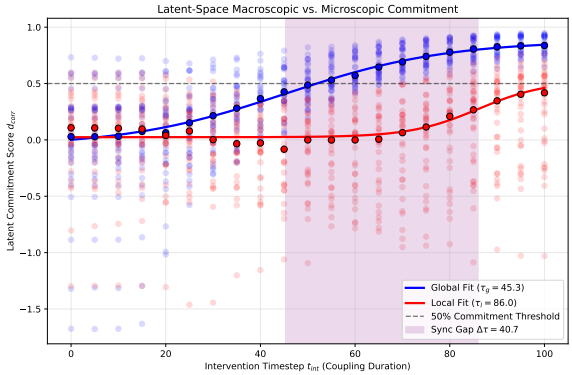


(c) $g = 1.0$

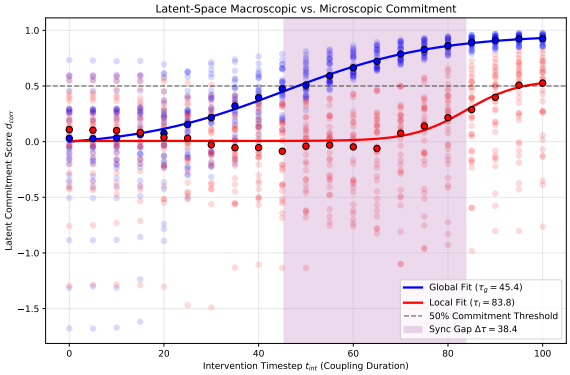
Figure 3. Exhaustive sweep of the Protocol I speciation probe (Part 2).



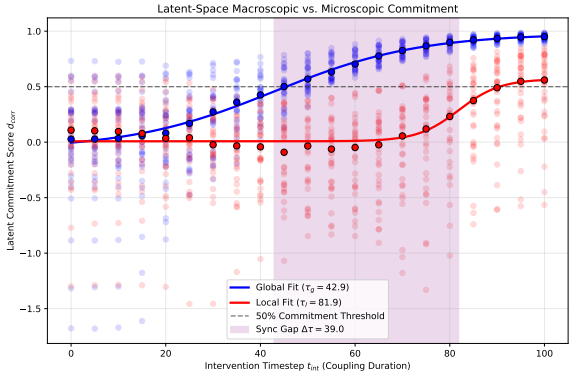
(a) $g = 0.1$



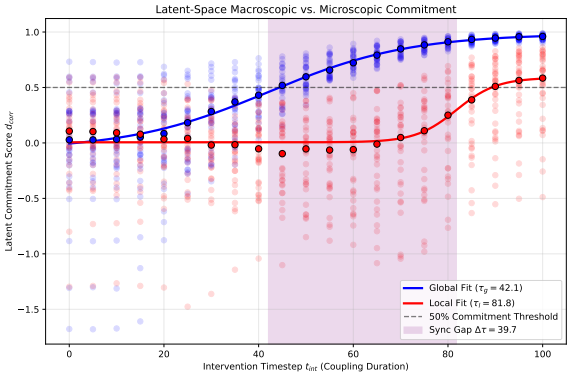
(b) $g = 0.3$



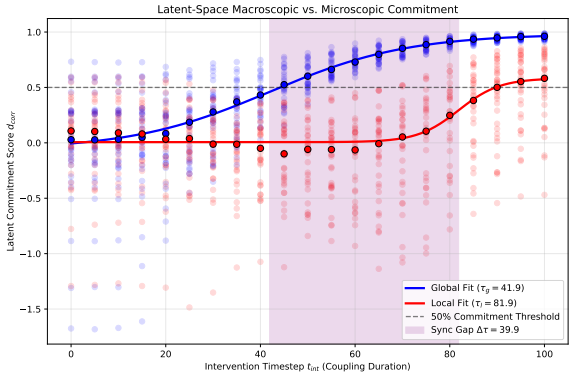
(c) $g = 0.5$



(d) $g = 0.7$

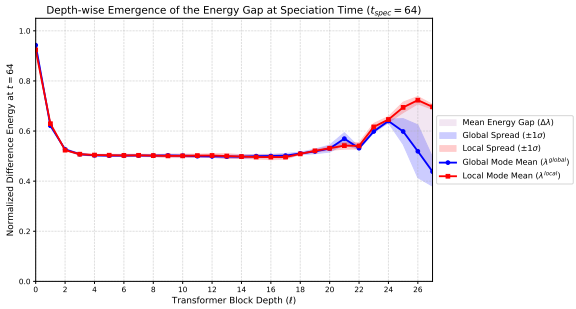


(e) $g = 0.9$

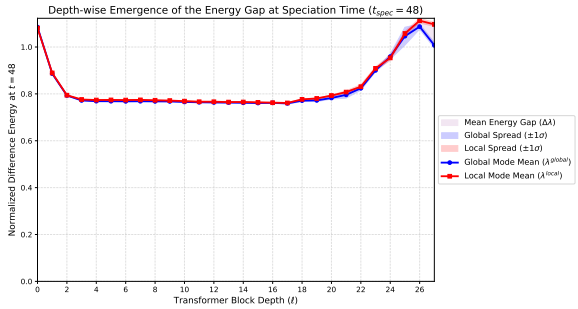


(f) $g = 1.0$

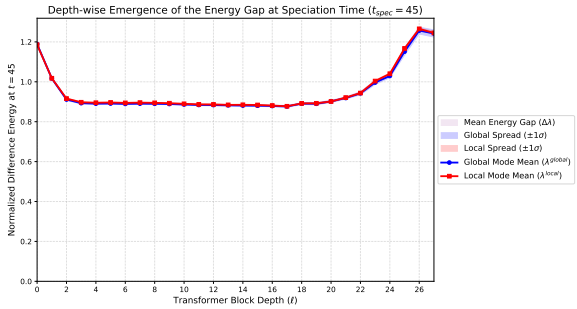
Figure 4. Each plot shows a global and local commitment scores vs coupling strength g. These show that coarse output structure stabilizes substantially earlier than fine detail throughout the explored coupling range.



(a) $g = 0.1,\ \tau_{\text{spec}} = 64$



(b) $g = 0.3,\ \tau_{\text{spec}} = 48$



(c) $g = 0.9,\ \tau_{\text{spec}} = 45$

Figure 5. For particular coupling strength g and speciation time, we sweep across all Transformer layers to evaluate the normalized fixed basis energies of leading and trailing internal difference modes. These sweeps reveal that the gap is located at deep layers of the Transformer and as coupling strength g increases the gap collapses. At $g = 0.1$, the internal gap localizes heavily in the deepest blocks. As g increases, the gap is monotonically suppressed until at $g = 0.9$.