Concurrence of Symmetry Breaking and Nonlocality Phase Transitions in Diffusion Models

Yifan F. Zhang
Department of Electrical and Computer Engineering
Princeton University
Princeton, NJ 08544, USA
yz4281@princeton.edu

Fangjun Hu
QuEra Computing Inc.
1284 Soldiers Field Road, Boston, MA 02135, USA
fhu@quera.com

Guangkuo Liu
JILA and Department of Physics
University of Colorado Boulder
Boulder, CO 80309, USA
guangkuo.liu@colorado.edu

Mert Okyay
CTQM and Department of Physics
University of Colorado Boulder
Boulder, CO 80309, USA
mert.okyay@colorado.edu

Xun Gao
JILA and Department of Physics
University of Colorado Boulder
Boulder, CO 80309, USA
xun.gao@colorado.edu

Abstract

Diffusion models undergo a phase transition in a critical time window during generation dynamics, with two complementary diagnoses of criticality. The symmetry breaking picture views the critical window as when trajectories bifurcate into different semantic minima of the energy landscape, whereas the nonlocality picture views the critical window as when local denoising fails. We study whether two notions of such phase transitions are concurrent in modern diffusion transformers. By evaluating the dynamics and outcomes of the generation trajectory, we observe a near-simultaneous occurrence of the non-locality and symmetry breaking critical times. Our work is the first to unify the two notions of phase transitions in practice: it provides a concrete diagnostic for when and why diffusion models rely on conditioning and global denoising, enabling principled evaluation of model efficiency and guiding the design of architectures and sampling schemes that avoid unnecessary computation.

Code: https://github.com/zsxqblz/symmetry_nonlocality_transition

1 Introduction

Diffusion models now produce high-quality images, but we still have a limited understanding of *when* different capabilities turn on during generation. A diffusion trajectory is not a uniform process: the model does not use all signals equally at all times. One influential line of work borrows techniques from statistical mechanics and interprets the generation process as a *symmetry breaking* [1] phase

[1] Although "symmetry breaking" is a somewhat imprecise term—given that semantic minima are not necessarily symmetric—we adopt it because it is standard in the literature [Raya and Ambrogioni, 2023] and effectively conveys the underlying physical concept.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

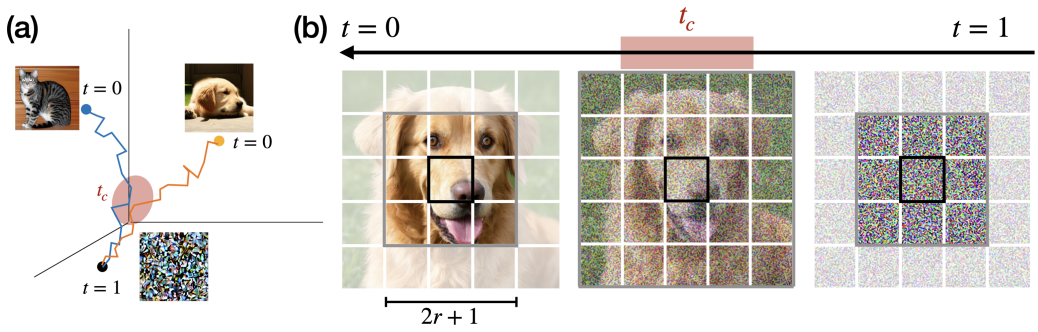

Figure 1: **Two definitions of phase transitions**. (a) Symmetry breaking phase transition. At $t = 1$, the energy landscape has one global minimum. As t approaches the phase transition, multiple local minima start to appear. (b) Nonlocality phase transition. Inside a phase, noise on a patch (black box) can be removed by using information in a small neighborhood (gray box). Near the phase transition, this neighborhood grows to be comparable to the entire image.

transition, where the distribution evolves from a single unstructured basin to multiple class- or prompt-consistent basins [Raya and Ambrogioni, 2023]. This picture predicts the existence of a "critical window" of time during which the dynamics has drastic impact on the output, such as changing the class or prompt semantics, while similar interventions outside that window have much smaller effects.

The statistical mechanics perspective is intuitive (and natural), as the diffusion model essentially mimics a non-equilibrium "cooling" from high-temperature Gaussian noise to low-temperature data distribution [Sohl-Dickstein et al., 2015]. However, this symmetry breaking paradigm is ignorant of the spatial structure of the data. For example, the image distribution undergoes the same symmetry breaking transition regardless of whether the model treats the image as a vectorization of pixels or as a structured 2D array. Since locality bias is a key inductive bias in dataset [Kamb and Ganguli, 2024] and model design, it is important to understand how it interacts with the phase-transition picture.

Tools and concepts to tackle the connection between phase transitions and locality have recently been developed in the many-body physics and quantum error correction community: a complementary notion of criticality, a *nonlocality* phase transition, has been proposed [Sang and Hsieh, 2025]. This perspective puts locality front and center, and asks when local neighborhoods of a site contain enough information for denoising. At a higher noise rate, information is further corrupted, so the denoiser needs to look at a larger neighborhood to recover one patch. The noise level at which the denoising radius diverges is understood as a phase transition in which global information is needed to perform denoising, which may contain information such as the symmetry sector in which the initial condition was. In diffusion models, the identical information-theoretic probe has been found to give rise to a critical window in which local denoising fails [Hu et al., 2025]. This criticality may be inherently distinct from phase transitions in equilibrium statistical mechanics, e.g., because it is based on a purely operational criterion and does not follow from a free-energy argument.

This paper aims to reconcile these two notions in frontier models. Our starting hypothesis is simple: they should be related because semantics typically carry nonlocal information. If a model is actively injecting semantic identity (for example, deciding between a Golden Retriever dog and a golden British Shorthair cat, which can share similar local fur textures but differ in global shape and anatomy), it should also require long-range context. Conversely, if denoising can be done from local texture statistics alone, strong semantic guidance should be less necessary. In short, these arguments suggest that *symmetry breaking predicts when semantic forcing is needed, and nonlocality predicts when nonlocal computation is needed*; if semantics are globally organized, the two windows should align.

Figure 1 clarifies the two notions of phase-transitions. Panel (a) depicts the critical window for symmetry breaking: as diffusion proceeds, trajectories in high-dimensional state space branch into distinct semantic basins (for example, cat-like versus dog-like minima). Panel (b) depicts a locality transition for patch denoising: once global semantic identity is already stable, recovering a local region (such as a dog's nose) may require only a small neighborhood (such as the head), whereas near a critical window where semantic identity is still unresolved or ambiguous, denoising must use broader, often global, context to inject the correct semantics.

1.1 Summary of Main Results

We probe the two definitions of critical window numerically from modern diffusion transformers (DiTs) and examine how they align. We use two complementary families of probes. The *instantaneous probe* directly measures score vectors along different trajectories. It compares conditional versus unconditional scores, and measures the symmetry breaking critical window where the difference is large. It also compares the global versus locally-approximated scores, and measures the nonlocality critical window where the difference is large. If the two critical windows are aligned, the two transitions are concurrent in the operational sense defined below.

We define *concurrence* operationally: after fixing a probe class and a critical-time estimator, we ask whether the critical time extracted from nonlocality data agrees with that extracted from conditioning data. A sharp critical point requires a scaling limit, which a finite model and dataset do not provide, so the claim we test is approximate within-probe temporal alignment rather than a universal critical point shared by all probes (Appendix E.1; limitations in Appendix D).

We complement the instantaneous probe with a series of *integrated probes*, which focus on how the dynamics influence the final generated samples. These probes include the standard forward-backward experiments [Biroli et al., 2024, Li et al., 2025] used to define symmetry breaking critical times, as well as a series of sliding window methods where we switch between conditional and unconditional, global and local denoisers in a time window of the denoising process. We benchmark the quality of generation and ask when global or conditional denoising is necessary to complete the tasks with high quality.

We find approximate within-probe alignment of symmetry breaking and nonlocality in both probe families. For instantaneous probes in the Facebook DiT-XL model [Peebles and Xie, 2023], the conditional versus unconditional gap becomes large in an early window centered near $t \approx 0.2$, and the local versus global gap is large in the same window. In Stable Diffusion 3 (SD3) Medium model [Esser et al., 2024], both gaps concentrate very close to the noisy endpoint $t = 1$. The co-occurrence of the two gaps gives operational evidence that semantic forcing and nonlocal computation turn on together.

In integrated probes, we find approximate alignment as well. Facebook DiT-XL shows a critical window for both symmetry breaking and nonlocality near $t \approx 0.5$. This window is later than the window found in instantaneous probes. This is not a contradiction: instantaneous probes measure local score geometry, whereas integrated probes measure accumulated dynamical influence on the sample, and how the former determines the latter remains open. One possibility is that the model uses conditioning and nonlocal computation earlier than needed. In contrast, the higher-performance model, SD3 medium, shows a concurrent locality and symmetry breaking phase transition close to $t = 1$ in both instantaneous probes and integrated probes. Appendix E.1 and Table 1 quantify these statements; Appendix E adds locality controls and robustness checks across classes, prompts, guidance scales, samplers and two further DiT models.

2 Theoretical Background

We briefly review the theoretical background here; Appendix B.1 reviews related work in more detail.

2.1 Diffusion models and the score function

Let $x_0 \sim p_{\mathrm{data}}$ denote a clean data sample (e.g., an image in pixel or latent space), and let x_t denote its noised version at diffusion time $t \in (0, 1]$ such that $t = 1$ is the high-noise limit. The continuous-time Gaussian forward noising process can be written as an SDE

$$\mathrm{d}x_t = f(x_t, t)\,\mathrm{d}t + g(t)\,\mathrm{d}W_t, \tag{1}$$

whose time-reversal admits the reverse-time SDE [Ho et al., 2020]

$$\mathrm{d}x_t = \big[f(x_t, t) - g^2(t)\, s^*(x_t, t)\big]\mathrm{d}t + g(t)\,\mathrm{d}\bar{W}_t, \tag{2}$$

where $\bar{W}_t$ is a standard Brownian motion running backwards in time. The function $s^*(x_t, t)$ is called the score at time t, and it is the gradient of the log density $s^*(x_t, t) = \nabla_{x_t} \log q_t(x_t)$ where q_t is the marginal distribution of x_t. Diffusion models learn a parametric approximation $s_\theta(x_t, t)$ to $s^*(x_t, t)$

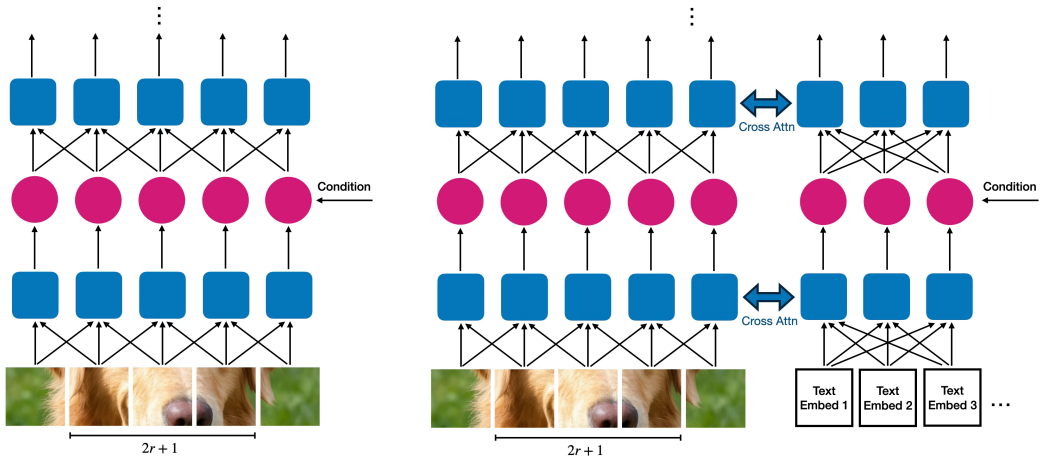

Figure 2: **DiT with local attention**. (a) In class-conditioned DiT, we restrict image-token attention to a local window with radius r and window size $2r + 1$. We visualize a one-dimensional strip of image tokens for simplicity. The conditioning signal still affects each token. (b) In multi-modal DiT (MMDiT), we restrict image token–image token attention to a local window with radius r. Text token–text token and image token–text token attentions remain dense, so conditioning enters through both token-wise modules and cross attention.

(or an equivalent parameterization) via denoising score matching [Song et al., 2021]. The learned score s_θ defines reverse-time dynamics that turns noise into a sample from the model distribution by initializing $x_1 \sim \mathcal{N}(0, I)$ and numerically integrating the reverse SDE from $t = 1$ to $t = 0$.

2.2 Symmetry breaking in sampling and classifier-free guidance

The origin of diffusion models in statistical physics invites a study of their dynamics, where early work noticed varying relative importance of different times [Song et al., 2022, Choi et al., 2022, Raya and Ambrogioni, 2023] for sample quality and class. These observations were studied in physics-inspired toy models [Biroli and Mézard, 2023, Ambrogioni, 2025], motivating the terminology 'symmetry-breaking' for when the output sample is placed in a class during the diffusion path.

Classifier-free guidance (CFG) [Ho and Salimans, 2022] mixes a conditional score $s_\theta(x_t, t \mid y)$ and an unconditional score $s_\theta(x_t, t)$ learned by the same network; in energy language it tilts the landscape by an external field that favors basins compatible with y (Appendix B.3). Symmetry breaking is not specific to unconditional sampling: even with conditioning, sampling starts from a unimodal Gaussian and can end in a multimodal distribution, for example when the label does not fully resolve the class structure of the data.

2.3 Local score function via approximate Markovianity

A separate line of work characterizes phases of quantum distributions by local recoverability: between two distributions Sang and Hsieh [2025] proposed the decay length-scale of conditional mutual information (CMI) (the 'Markov length') to diagnose when one is not locally recoverable from the other. Bringing these insights to learning, Hu et al. [2026] has shown that if the CMI remains short-ranged during diffusion, then one can construct a sampler that outputs from this distribution using only local operations. On the other hand, Kumar et al. [2026] shows that distributions with long-ranged CMI are generically hard to learn.

Concretely, split the image into a patch A, a surrounding buffer B and the rest C (Figure 1(b)). If x_A and x_C are conditionally independent given x_B, the score on A depends only on (x_A, x_B); approximate versions of this property are quantified by the decay of the CMI $I(A : C \mid B)$ with distance, whose length scale is the Markov length ξ. Appendix B.2 gives the derivation, the decay ansatz with prefactor I_0, and the result of Hu et al. [2025] that a bounded ξ implies a local denoiser.

These notions only motivate our truncated-attention denoiser (Section 3.1), an operational proxy that reflects the local denoising mechanism but does not measure CMI directly.

3 Experiments

We study two diffusion transformer systems: a Facebook DiT-XL class-conditional ImageNet model and SD3 medium, a text-conditioned MMDiT model. As before, throughout the experiments, $t = 1$ denotes the noisy endpoint and $t = 0$ denotes the clean endpoint; experimental details are in Appendix C.

3.1 Constructing local denoisers

Training a local denoiser could be computationally expensive. Instead, we take any pre-trained DiT and construct a local version of it, which also helps probe the model's internal circuitry. For each image token, we truncate the attention window to a small radius r around itself as shown in Figure 2(a). Note that for Facebook DiT-XL, class and time conditioning is injected into each patch independently through AdaLN [Perez et al., 2018, Peebles and Xie, 2023]. Since the conditioning is still computed globally, in principle the model is capable of providing global conditioning information to each patch even when image-token attention is local. For MMDiT model such as SD3, both image and text tokens enter the joint attention. Therefore, we only restrict image token–image token attention to a local window, while keeping text token–text token and image token–text token attentions dense, as described in Figure 2(b). This way, the conditioning signal can still be globally available to each patch through both cross attention and text token self attention.

The truncation procedure does not produce a strictly local denoiser: context from patches farther than r can still enter by propagating through multiple local attention steps. The truncated DiT is therefore an approximation with strong locality bias. It is operationally meaningful because it is a minimal perturbation of the original model architecture, and local attention could in principle reduce compute.

We use attention truncation because it computes outputs for all image patches in one model evaluation, whereas a strictly local denoiser, which feeds the model only the image tokens near one center patch (plus the text tokens for MMDiT), requires a separate evaluation for each patch. We tested this center-patch denoiser and found score-gap behavior broadly similar to that of the truncated model in both norm and cosine similarity (Appendix E, Table 2).

3.2 Instantaneous probe: score gap

We first probe the phase transition through the instantaneous probe, which measures the behavior of the score function along the sampling and training trajectories. We take the conditional score function $s_\theta(x_t, t \mid y)$ and the unconditional score function $s_\theta(x_t, t)$, define the *conditioning gap* as the l_2 norm of their difference:

$$\Delta\vec{s}_{\mathrm{cond}}(x_t, t \mid y) = s_\theta(x_t, t) - s_\theta(x_t, t \mid y). \tag{3}$$

The conditioning gap provides a direct, model-internal readout of how strongly the condition changes the denoising vector at time t.

For nonlocality, we construct a local denoiser from the pretrained model following Figure 2, which provides a local approximation $s_{\theta,r}(x_t, t)$ to the global score function $s_\theta(x_t, t)$. The *locality gap* is the l_2 norm of their difference,

$$\Delta\vec{s}_{\mathrm{loc}}(x_t, t, r) = s_\theta(x_t, t) - s_{\theta,r}(x_t, t). \tag{4}$$

The above definition is valid for both conditional and unconditional score function. If the trained model can compute the relevant score locally at time t, this gap should be small for modest r. If global information is necessary, the gap should remain large until r is sufficiently large. This gives an operational proxy for the CMI-based nonlocality probe; it is not a direct estimate of CMI or of the Markov length. Throughout this paper, we sample images of 256×256 pixels for Facebook DiT and 512×512 pixels for SD3 medium. Both models we consider have patch size of 16 pixels, so in our study, setting $r = 16$ for Facebook DiT / $r = 32$ for SD3 medium reduces local denoiser to global denoiser.

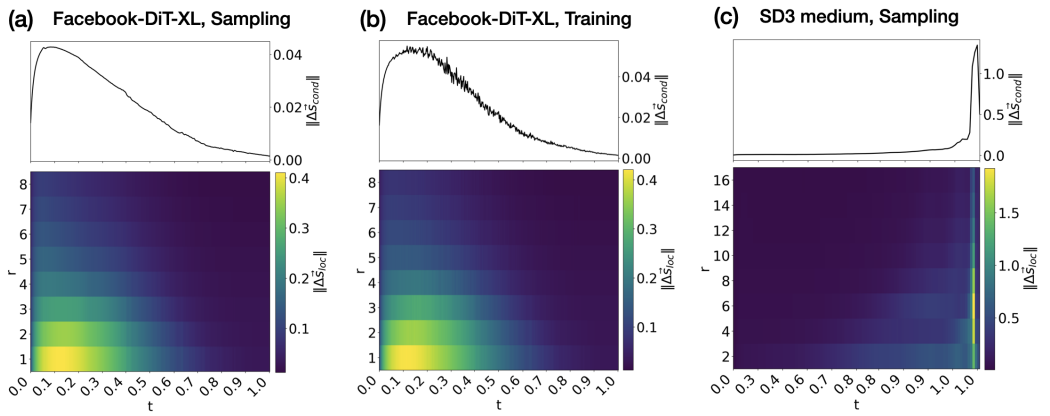

Figure 3: **Conditioning gap vs. locality gap**. Top row: the one-dimensional plot shows the conditioning gap $\|\Delta\vec{s}_{\text{cond}}\|$ as a function of t. Bottom row: the heatmap shows the locality gap $\|\Delta\vec{s}_{\text{loc}}\|$, defined using conditional score function, as a function of attention radius r and time t. (a) Facebook DiT-XL along the sampling trajectory. (b) Facebook DiT-XL along the training trajectory. (c) SD3 medium along the sampling trajectory (the horizontal axis is not linear in t; tick labels give t). In all three panels, the two gaps are approximately aligned in time despite being functionally independent probes.

We first ask whether the two score gaps identify the same time window. Figure 3 plots the conditioning gap on the top row and the locality gap on the bottom row. We normalize the gap by the number of image tokens. Both gaps are raw latent-space l_2 norms that are not weighted by the sampler step size: they indicate when the score changes, not by themselves how strongly that change affects the final sample, which the integrated probes below test. The top row measures the strength of the semantic field induced by conditioning. The bottom row measures the cost of replacing the global denoiser with a local one: large values indicate that the model's score cannot be reproduced from local image context alone.

For Facebook DiT-XL, Figure 3(a,b), the conditioning gap is initially small at t close to zero. This is intuitive, as the semantics are still present in slightly noisy images, so conditioning is unnecessary. As t increases, it becomes large and saturates near $t \approx 0.2$. It then decays as t approaches the noisy endpoint $t = 1$. This behavior is consistent with the phenomenology of symmetry breaking phase transition, and is also consistent with the empirical observation of when conditioning is beneficial [Kynkäänniemi et al., 2024, Jin et al., 2025].

The behavior of the locality gap is largely aligned with the conditioning gap. At small and large t, the locality gap quickly decays at increasing r, indicating that local denoisers approximate global denoiser well. Around $t \approx 0.2$, locality gap remains large even at large r. The behaviors are similar between training and sampling trajectories, indicating a good generalization of the model. For SD3 medium, Figure 3(c), the two gaps behave similarly and are aligned. Prompt dependence is quantified in Appendix E.

The conditioning and locality gaps are independent quantities: one compares conditional and unconditional scores, while the other compares global and local computation under the same conditioning setup. Their alignment therefore gives operational evidence that the symmetry breaking transition and the nonlocality transition are linked.

3.3 Integrated probe: forward-backward error correction

The model-internal details captured by gaps (3) and (4) may be smeared out during sampling (which integrates over time), so we need integrated probes which directly benchmark the downstream samples. The first is the 'forward-backwards experiment' illustrated in Figure 4(a): start from a clean image, add noise forward to time t, and then denoise backward unconditionally. If the noised state remains inside the basin of the original semantic class, the unconditional denoiser recovers an image from the same class. If t crosses the semantic error threshold, the unconditional denoiser can return

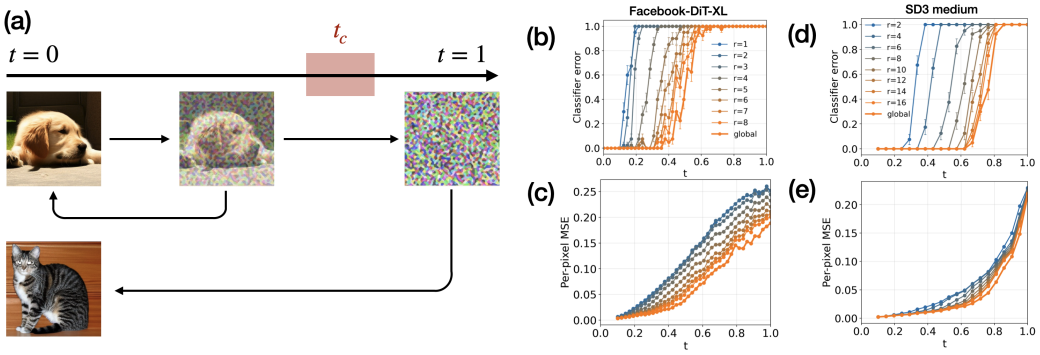

Figure 4: **Forward-backward experiment**. (a) Schematic of the error-correction experiment. When the noise level is below a threshold, the unconditional denoiser recovers the same class; when the noise level is above the threshold, the unconditional denoiser can recover a different class. (b,d) Classification error of images denoised from different noise levels t for Facebook DiT-XL and SD3 medium. Curves labeled by r use local denoisers with attention radius r (see Fig. 2); the global curve uses the original denoiser. (c,e) Per-pixel MSE of the denoised image relative to the clean image.

an image from a different class. This is also a standard probe of symmetry breaking phase transition in early studies by Biroli et al. [2024], Sclocchi et al. [2025c], Li et al. [2025].

Figures 4(b,d) plot the classifier-error rate of the denoised image using Facebook DiT-XL and SD3 medium. Both models show sharp classifier-error jumps as t increases. For Facebook DiT-XL, Figure 4(b), the global threshold is around $t \approx 0.5$. Local denoisers also exhibit sharp thresholds, and those thresholds move toward the global threshold as r increases. This indicates that local denoising has its own error-correction transition, and that larger local neighborhoods recover more of the global denoiser's semantic capability.

For SD3 medium, Figure 4(d), the corresponding threshold is very close to $t = 1$, and local thresholds again approach the global threshold as r grows.

Figures 4(c,e) show a complementary low-level metric. The per-pixel MSE increases smoothly and sublinearly with t, with an almost flat derivative at small t. This is consistent with error correction: small forward noise can be removed without large reconstruction error. However, MSE does not expose the semantic transition. The semantic class can be destroyed abruptly while the pixelwise error curve remains smooth, because semantic information is nonlocal and is not visible in a per-pixel distortion measure.

This experiment also separates two kinds of alignment. Within the forward-backward probe, the global and local denoisers both show sharp thresholds, and the local thresholds converge toward the global threshold with larger radius. Across probes, however, Facebook DiT-XL is misaligned: the instantaneous score-gap window is near $t \approx 0.2$, while the integrated semantic error threshold is near $t \approx 0.5$. Because instantaneous probes characterize local score geometry whereas integrated probes measure accumulated dynamical influence, we compare critical times within a probe (Appendix E.1). One interpretation of the cross-probe offset is that Facebook DiT-XL performs unnecessary conditioning and nonlocal computation at low t. SD3 medium is more aligned across probes, with close to $t = 1$ transitions predicted by both instantaneous and integrated probes.

3.4 Integrated probe: windowed conditioning

We next ask when conditioning is necessary for actual sampling. The symmetry breaking picture predicts that conditioning is most important near the critical window. We first demonstrate this idea with Facebook DiT-XL in Figure 5. Results for SD3 medium are in Appendix A. Figure 5(a) shows that applying conditioning only during $[0.2, 0.7]$, which covers the critical window predicted from forward-backward experiment, can produce a recognizable golden retriever. Figure 5(b) shows the complementary intervention: applying conditioning outside $[0.2, 0.7]$ but removing it inside the window produces a much worse semantic result. Thus the outside-window conditioning is not enough, even though local textures and image statistics may still look plausible.

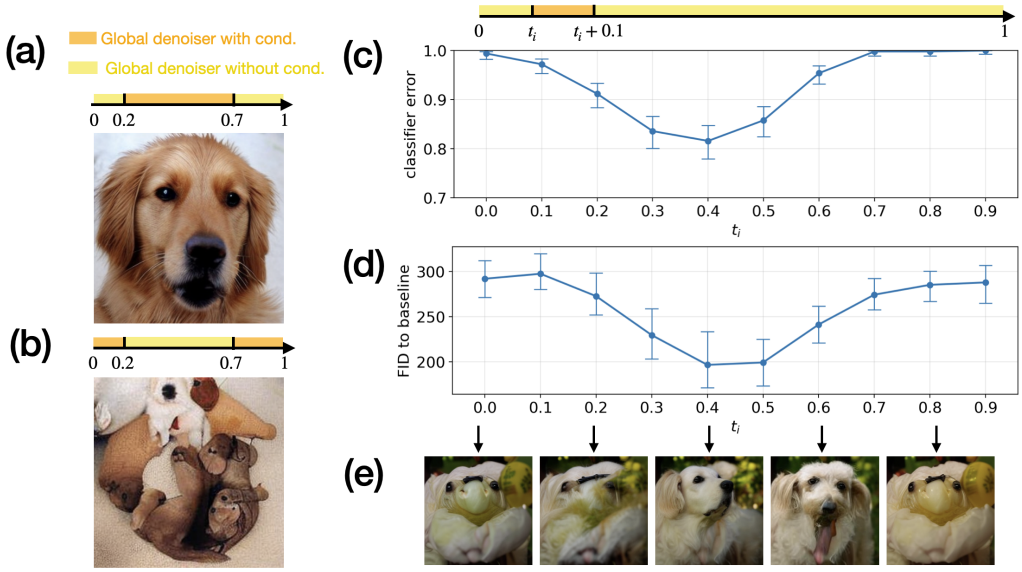

Figure 5: **Time-windowed conditioning in Facebook DiT-XL**. Left column: (a) Golden retriever sample generated with conditioning applied only in the time window $[0.2, 0.7]$ and no conditioning outside the window. (b) Sample generated with conditioning removed in $[0.2, 0.7]$ and applied outside the window. Right column: samples generated with conditioning applied only in short windows $[t_i, t_i + 0.1]$ while scanning t_i. (c) Classification error as a function of t_i. (d) FID to always-conditioned samples as a function of t_i. (e) Visualization of generated samples at different t_i.

To quantitatively pin down the critical window, we choose a short interval $[t_i, t_i+0.1]$ for conditioning and scan over t_i. We expect that when $[t_i, t_i + 0.1]$ overlaps with the critical window the most, the sample has better quality. Figure 5(c) shows that classifier error is lowest around $t_i \approx 0.4$, and Figure 5(d) shows that Fréchet inception distance (FID) to the always-conditioned baseline is also lowest around the same time. The samples in Figure 5(e) visualize this dependence: semantic correctness emerges when conditioning is placed in the critical window. This indicates an alignment with the forward-backward experiment. The SD3 medium windowed-conditioning experiment is shown in the appendix (Figure 7) and exhibits similar alignment.

Classifier error and FID are large overall because the intervention heavily corrupts the sampling dynamics; the quantity of interest is the t_i at which this corruption is smallest.

3.5 Integrated probe: windowed local/global denoising

The windowed-conditioning experiments test when semantic conditioning is needed. We now test when global denoising is needed. In these experiments, a global denoiser is used inside a time window and a local denoiser is used outside it, or vice versa.

We start with visual examples using Facebook DiT-XL (SD3 medium in Appendix A, Figure 8). Figure 6(a)(i) compares the two extremes: an always-global denoiser produces a clean golden retriever, while an always-local denoiser gives "domains" of small dogs (with size comparable to r) smoothly merged together. This can be understood as the creative samples from Kamb and Ganguli [2024]. Figure 6(a)(ii) shows that using the global denoiser during $[0.2, 0.7]$ and a local conditioned denoiser outside that window preserves a globally coherent dog better than the reverse schedule. Figure 6(a)(iii) makes the separation from conditioning sharper: even when the outside-window local denoiser is unconditional, putting global denoising in the critical window remains helpful, whereas replacing the critical window by local denoising is much worse.

To pin down the critical window for nonlocality phase transition, we scan a global conditioning window $[t_i - 0.2, t_i + 0.2]$ and use local unconditional denoiser outside. Figures 6(b,c) show that classifier error and FID are again minimized when the global-denoising window is centered near

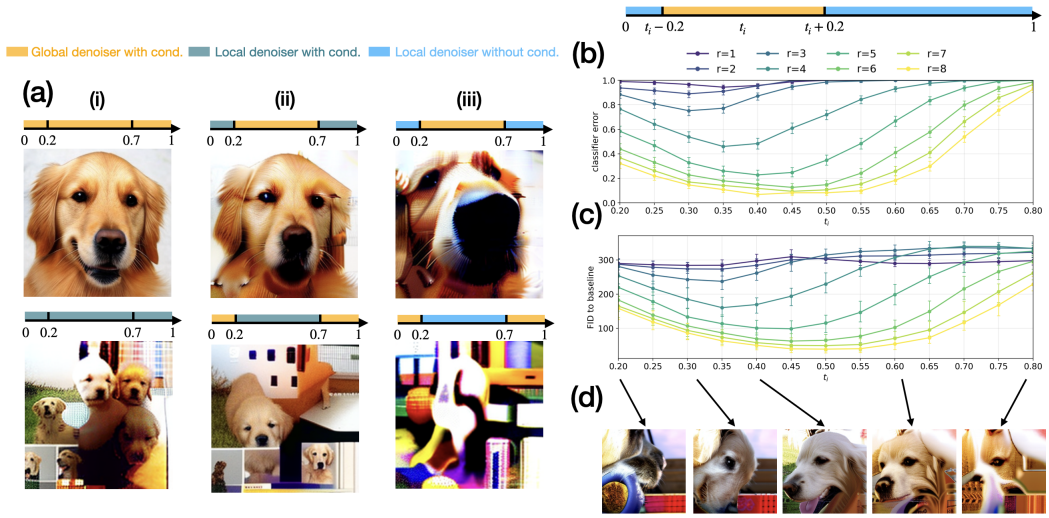

Figure 6: **Time-windowed local denoising in Facebook DiT-XL**. (a) Golden retriever samples generated with different combinations of local and global denoisers. The local denoiser uses radius $r = 3$. (i) Top: global denoiser at all times. Bottom: local denoiser at all times. (ii) Top: global denoiser in $[0.2, 0.7]$ and local conditioned denoiser outside the window. Bottom: local conditioned denoiser in $[0.2, 0.7]$ and global denoiser outside the window. (iii) Top: global denoiser in $[0.2, 0.7]$ and local unconditional denoiser outside the window. Bottom: local unconditional denoiser in $[0.2, 0.7]$ and global denoiser outside the window. Right column: samples generated with a global denoiser in $[t_i - 0.2, t_i + 0.2]$ and a local unconditional denoiser with different r outside the window while scanning t_i. (b) Classification error as a function of t_i. (c) FID to always-global samples as a function of t_i. (d) Visualization of generated samples at different t_i using $r = 3$.

$t_i \approx 0.4$. The samples in Figure 6(d) visualize the same window dependence. The minima also shift slightly to larger t_i as r increases. We hypothesize that this is related to the increasing threshold at increasing r in the forward-backward experiment (Figure 4): at a fixed r the observed critical window would then be earlier than the actual critical window, with the difference vanishing as r increases. Together with previous integrated probe results, we observe an approximate alignment between symmetry breaking and nonlocality transition, through error correction and sampling.

4 Conclusion

Our work provides a unified operational view of two notions of phase transitions in diffusion models. By connecting symmetry breaking induced by conditioning with the onset of nonlocal computation in the score, we show that both phenomena can be detected through the same set of measurable quantities. This yields a concrete diagnostic for identifying when global communication is actually required during sampling, rather than assumed a priori from model design. In particular, the agreement (or mismatch) between instantaneous score-based probes and integrated sampling behavior can indicate how efficiently a model uses its conditioning signal and attention mechanism over time.

These diagnostics open a path toward diffusion models that organize computation around the critical window. Future architectures can gate cross-token attention or restrict receptive fields outside this window, while training objectives, noise schedules and locality constraints can concentrate learning near the critical interval. The typical score-gap profile across trajectories (Figure 9) also motivates samplers that switch between local and global denoisers using a window calibrated once before inference (Appendix F).

Acknowledgments and Disclosure of Funding

G. L. and X. G. acknowledge support from NSF PFC grant No. PHYS 2317149.

The authors declare no competing interests.

References

Luca Ambrogioni. The statistical thermodynamics of generative diffusion models: Phase transitions, symmetry breaking, and critical instability. *Entropy*, 27(3), 2025. ISSN 1099-4300. doi: 10.3390/e27030291. URL https://www.mdpi.com/1099-4300/27/3/291.

Luca Ambrogioni. How out-of-equilibrium phase transitions can seed pattern formation in trained diffusion models, 2026. URL https://arxiv.org/abs/2603.20092.

Giulio Biroli and Marc Mézard. Generative diffusion in very large dimensions. *Journal of Statistical Mechanics: Theory and Experiment*, 2023(9):093402, September 2023. ISSN 1742-5468. doi: 10.1088/1742-5468/acf8ba. URL http://dx.doi.org/10.1088/1742-5468/acf8ba.

Giulio Biroli, Tony Bonnaire, Valentin de Bortoli, and Marc Mézard. Dynamical regimes of diffusion models. *Nature Communications*, 15(1):9957, November 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-54281-3. URL https://doi.org/10.1038/s41467-024-54281-3.

Black Forest Labs. FLUX.2 Klein 9B. https://huggingface.co/black-forest-labs/FLUX.2-klein-9B, 2026. Hugging Face model card.

Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11462–11471, 2022. doi: 10.1109/CVPR52688.2022.01118.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. URL https://arxiv.org/abs/2403.03206.

Yiqiu Han, Xiaoyang Huang, Zohar Komargodski, Andrew Lucas, and Fedor K Popov. Entropic order. *Nature Communications*, 2025.

Florian Handke, Félix Koulischer, Gabriel Raya, and Luca Ambrogioni. Measuring semantic information production in generative diffusion models, 2025. URL https://arxiv.org/abs/2506.10433.

Florian Handke, Dejan Stančević, Felix Koulischer, Thomas Demeester, and Luca Ambrogioni. The entropic signature of class speciation in diffusion models, 2026. URL https://arxiv.org/abs/2602.09651.

Ali Hassani. Neighborhood attention: Dynamic restriction of self attention. Technical Report AREA-202307, University of Oregon, Department of Computer Science, 2023. URL https://www.cs.uoregon.edu/Reports/AREA-202307-Hassani.pdf.

Ali Hassani and Humphrey Shi. Dilated neighborhood attention transformer, 2022. URL https://arxiv.org/abs/2209.15001.

Ali Hassani, Steven Walton, Jiachen Li, Shen Li, and Humphrey Shi. Neighborhood attention transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6185–6194, 2023. doi: 10.1109/CVPR52729.2023.00599.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Fangjun Hu, Guangkuo Liu, Yifan F. Zhang, and Xun Gao. Local diffusion models and phases of data distributions, 2025. URL https://arxiv.org/abs/2508.06614.

Fangjun Hu, Christian Kokail, Milan Kornjača, Pedro L. S. Lopes, Weiyuan Gong, Sheng-Tao Wang, Xun Gao, and Stefan Ostermann. Learning and generating mixed states prepared by shallow channel circuits. *arXiv:2604.01197 [quant-ph]*, 2026. URL https://arxiv.org/abs/2604.01197.

Hugging Face and Microsoft. Resnet-50 v1.5. https://huggingface.co/microsoft/resnet-50, 2025.

Cheng Jin, Qitan Shi, and Yuantao Gu. Stage-wise dynamics of classifier-free guidance in diffusion models. *arXiv preprint arXiv:2509.22007*, 2025.

Mason Kamb and Surya Ganguli. An analytic theory of creativity in convolutional diffusion models. *arXiv preprint arXiv:2412.20292*, 2024.

Tarun Advaith Kumar, Yijian Zou, Amir-Reza Negari, Roger G. Melko, and Timothy H. Hsieh. Unlearnable phases of matter. *arXiv:2602.11262 [cond-mat.dis-nn]*, 2026. URL https://arxiv.org/abs/2602.11262.

Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. DOI: 10.52202/079017-3892.

Marvin Li and Sitan Chen. Critical windows: non-asymptotic theory for feature emergence in diffusion models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Marvin Li, Aayush Karan, and Sitan Chen. Blink of an eye: a simple theory for feature localization in generative models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=QvqnPVGWAN.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in neural information processing systems*, 35:5775–5787, 2022.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *Machine Intelligence Research*, 22 (4):730–751, 2025.

Haiqi Lu and Ying Tang. Steering dynamical regimes of diffusion models by breaking detailed balance, 2026. URL https://arxiv.org/abs/2602.15914.

Artem Lukoianov, Chenyang Yuan, Justin Solomon, and Vincent Sitzmann. Locality in image diffusion models emerges from data statistics. *arXiv preprint arXiv:2509.09672*, 2025.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=aBsCjcPu_tE.

N. D. Mermin and H. Wagner. Absence of ferromagnetism or antiferromagnetism in one- or two-dimensional isotropic heisenberg models. *Phys. Rev. Lett.*, 17:1133–1136, Nov 1966. doi: 10.1103/PhysRevLett.17.1133. URL https://link.aps.org/doi/10.1103/PhysRevLett.17.1133.

Matthew Niedoba, Berend Zwartsenberg, Kevin Murphy, and Frank Wood. Towards a mechanistic explanation of diffusion model generalization. *arXiv preprint arXiv:2411.19339*, 2024.

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

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Bao Pham, Gabriel Raya, Matteo Negri, Mohammed J Zaki, Luca Ambrogioni, and Dmitry Krotov. Memorization to generalization: The emergence of diffusion models from associative memory. In *NeurIPS 2024 Workshop on Scientific Methods for Understanding Deep Learning*, 2024. URL https://openreview.net/forum?id=zVMMaVy2BY.

Sai Niranjan Ramachandran, Manish Krishan Lal, and Suvrit Sra. Cross-fluctuation phase transitions reveal sampling dynamics in diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=b4X6cz1F9l.

Gabriel Raya and Luca Ambrogioni. Spontaneous symmetry breaking in generative diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=lxGFGMMSVl.

Shengqi Sang and Timothy H. Hsieh. Stability of mixed-state quantum phases via finite markov length. *Phys. Rev. Lett.*, 134:070403, Feb 2025. doi: 10.1103/PhysRevLett.134.070403. URL https://link.aps.org/doi/10.1103/PhysRevLett.134.070403.

Antonio Sclocchi, Alessandro Favero, Noam Itzhak Levi, and Matthieu Wyart. Probing the latent hierarchical structure of data via diffusion models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025a. URL https://arxiv.org/abs/2410.13770.

Antonio Sclocchi, Alessandro Favero, and Matthieu Wyart. A phase transition in diffusion models reveals the hierarchical nature of data. *Proceedings of the National Academy of Sciences*, 122 (1):e2408799121, 2025b. doi: 10.1073/pnas.2408799121. URL https://www.pnas.org/doi/abs/10.1073/pnas.2408799121.

Antonio Sclocchi, Alessandro Favero, and Matthieu Wyart. A phase transition in diffusion models reveals the hierarchical nature of data. *Proceedings of the National Academy of Sciences*, 122(1): e2408799121, 2025c. doi: 10.1073/pnas.2408799121.

Kulin Shah, Alkis Kalavasis, Adam R. Klivans, and Giannis Daras. Does generation require memorization? creative diffusion models using ambient diffusion. *arXiv preprint arXiv:2502.21278*, 2025.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022. URL https://arxiv.org/abs/2010.02502.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Stability AI. Stable diffusion 3.5 large. https://huggingface.co/stabilityai/stable-diffusion-3.5-large, 2026. Hugging Face model card.

Dejan Stančević and Luca Ambrogioni. The information dynamics of generative diffusion. *Entropy*, 28(2), 2026. ISSN 1099-4300. doi: 10.3390/e28020195. URL https://www.mdpi.com/1099-4300/28/2/195.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

Tomoei Takahashi, Takashi Takahashi, and Yoshiyuki Kabashima. Dynamical regimes of discrete diffusion models, 2026. URL https://arxiv.org/abs/2604.10961.

Enrico Ventura, Beatrice Achilli, Luca Ambrogioni, and Carlo Lucibello. Emergence of distortions in high-dimensional guided diffusion models, 2026. URL https://arxiv.org/abs/2602.00716.

Jiafu Wu, Yabiao Wang, Jian Li, Jinlong Peng, Yun Cao, Chengjie Wang, and Jiangning Zhang. Swin dit: Diffusion transformer using pseudo shifted windows. *arXiv preprint arXiv:2505.13219*, 2025.

Manato Yaguchi, Kotaro Sakamoto, Ryosuke Sakamoto, Masato Tanabe, Masatomo Akagawa, Yusuke Hayashi, Masahiro Suzuki, and Yutaka Matsuo. The geometry of phase transitions in diffusion models: Tubular neighbourhoods and singularities. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=ahVFKFLYk2. Featured Certification.

Jinmin Yi, Kangle Li, Chuan Liu, Zixuan Li, and Liujun Zou. Universal decay of mutual information and conditional mutual information in gapped pure- and mixed-state quantum matter. *Phys. Rev. Lett.*, 136:116604, Mar 2026. doi: 10.1103/mqp8-y1m7. URL https://link.aps.org/doi/10.1103/mqp8-y1m7.

Zhihang Yuan, Hanling Zhang, Pu Lu, Xuefei Ning, Linfeng Zhang, Tianchen Zhao, Shengen Yan, Guohao Dai, and Yu Wang. Ditfastattn: Attention compression for diffusion transformer models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. DOI: 10.52202/079017-0037.

Yifan Zhang and Sarang Gopalakrishnan. Conditional mutual information and information-theoretic phases of decohered gibbs states, 2025a. URL https://arxiv.org/abs/2502.13210.

Yifan F. Zhang and Sarang Gopalakrishnan. Stability of mixed-state phases under weak decoherence, 2025b. URL https://arxiv.org/abs/2511.01976.

A Additional Data

Figure 7 and Figure 8 provide the SD3 medium versions of the two windowed sampling experiments discussed in the main text.

One may ask if the score gap plots are typical behaviors for different trajectories. Figure 9 plots the standard deviation of different score gaps due to the randomness in the trajectory. We observe a small fluctuation across all score gaps, indicating that the observed behavior is typical.

We also provide additional data comparing the score gaps along conditional and unconditional sampling trajectories in Figure 10(a-d). For conditional sampling trajectories we report the conditional locality gap since it is relevant to this type of trajectory. Similarly, for unconditional sampling trajectories we report the unconditional locality gap. We observe the same behavior for both types of trajectories. We also compare the conditional and unconditional locality gap for Facebook DiT along the training trajectory. We observe a very small difference.

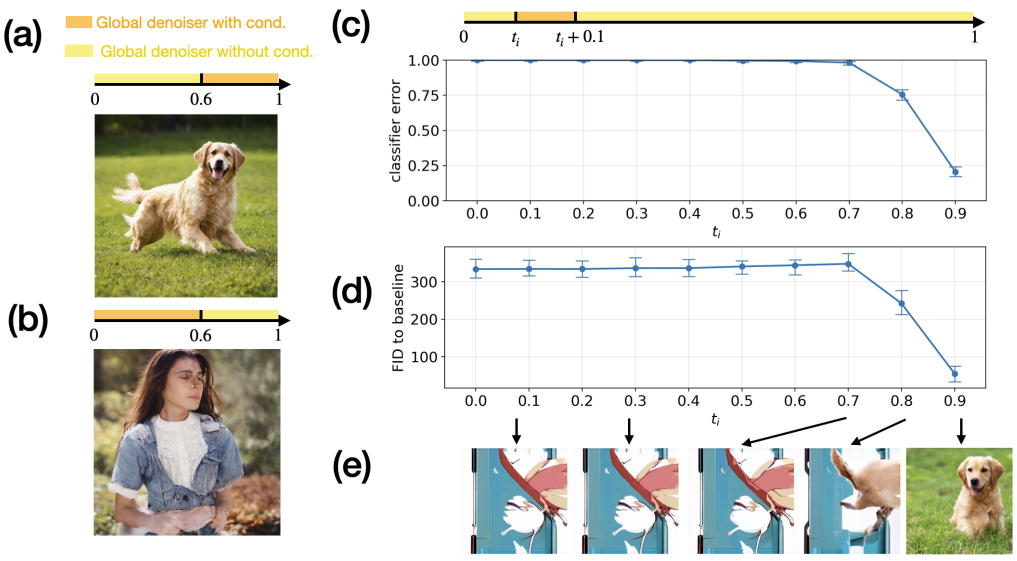

Figure 7: **Windowed conditioning in SD3 medium**. Left column: (a) Golden retriever sample generated with conditioning applied only in the time window $[0.6, 1.0]$ and no conditioning outside the window. (b) Sample generated with conditioning removed in $[0.6, 1.0]$ and applied outside the window. Right column: samples generated with conditioning applied only in short windows $[t_i, t_i + 0.1]$ while scanning t_i. (c) Classification error as a function of t_i. (d) FID to always-conditioned samples as a function of t_i. (e) Visualization of generated samples at different t_i.

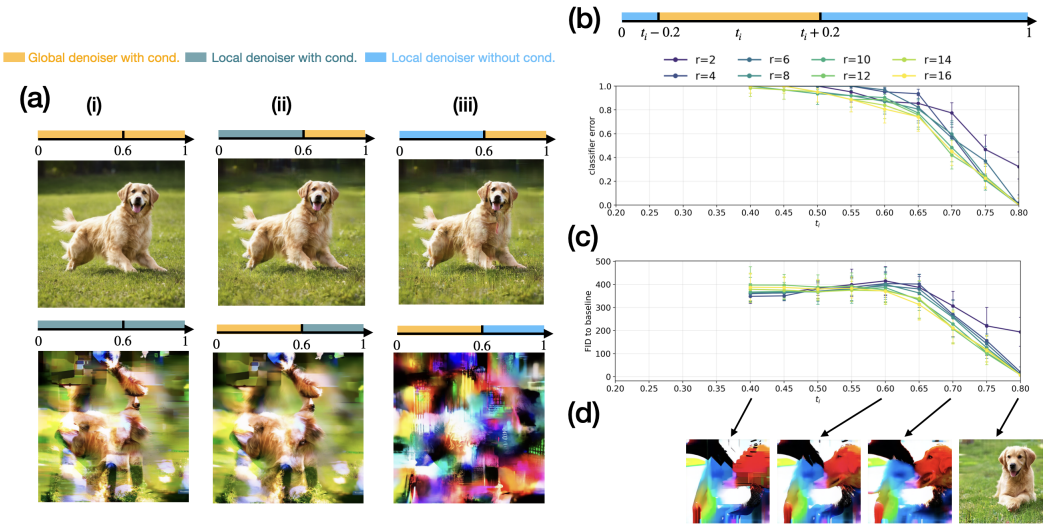

Figure 8: **Windowed local denoising in SD3 medium**. (a) Golden retriever samples generated with different combinations of local and global denoisers. The local denoiser uses radius $r = 6$. (i) Top: global denoiser at all times. Bottom: local denoiser at all times. (ii) Top: global denoiser in $[0.6, 1.0]$ and local conditioned denoiser outside the window. Bottom: local conditioned denoiser in $[0.6, 1.0]$ and global denoiser outside the window. (iii) Top: global denoiser in $[0.6, 1.0]$ and local unconditional denoiser outside the window. Bottom: local unconditional denoiser in $[0.6, 1.0]$ and global denoiser outside the window. Right column: samples generated with a global denoiser in $[t_i - 0.2, t_i + 0.2]$ and a local unconditional denoiser with different r outside the window while scanning t_i. (b) Classification error as a function of t_i. (c) FID to always-global samples as a function of t_i. (d) Visualization of generated samples at different t_i using $r = 6$.

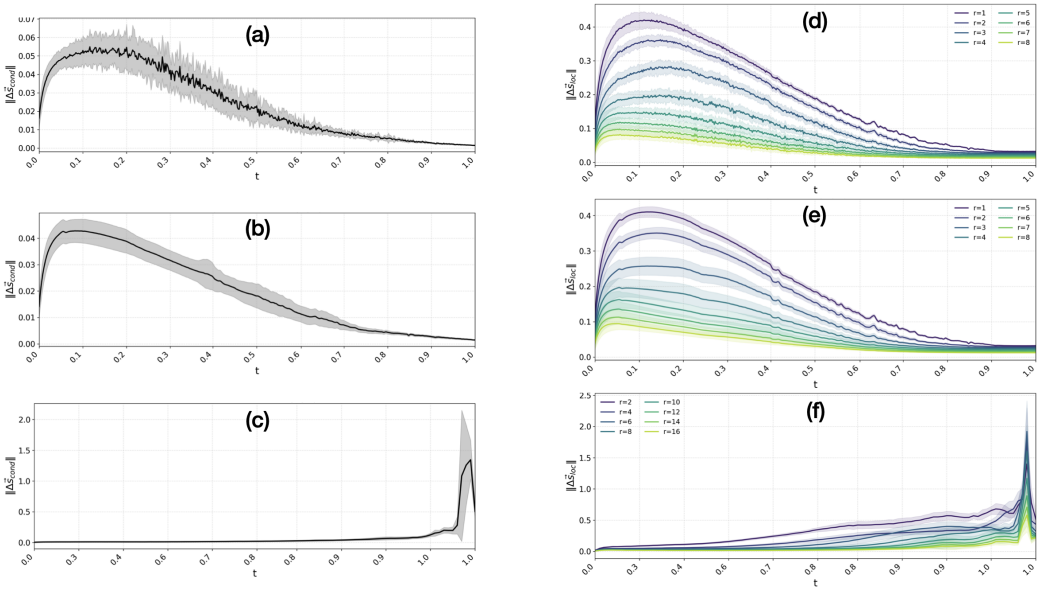

Figure 9: **Fluctuation of score gap**. Plot of score gaps with colored area marking the standard deviation across different trajectories. (a) Conditioning gap and (b) locality gap for Facebook DiT along the training trajectory. (c) Conditioning gap and (d) locality gap for Facebook DiT along the conditional sampling trajectory. (e) Conditioning gap and (f) locality gap for SD3 medium along the conditional sampling trajectory.

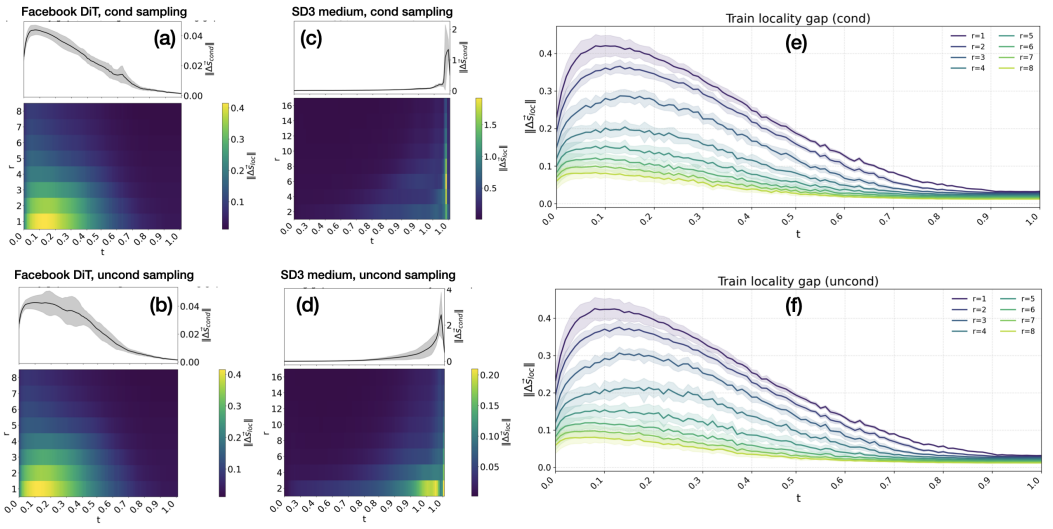

Figure 10: **Score gaps along conditional and unconditional trajectories**. (a) Conditioning gap (top) and locality gap defined using conditional score function, for Facebook DiT along the conditional sampling trajectory. (b) Conditioning gap (top) and locality gap defined using unconditional score function, for Facebook DiT along the unconditional sampling trajectory. (c) Conditioning gap (top) and locality gap defined using conditional score function, for SD3 medium along the conditional sampling trajectory. (d) Conditioning gap (top) and locality gap defined using unconditional score function, for SD3 medium along the unconditional sampling trajectory. (e,f) Locality gap for Facebook DiT along the training trajectory, for (e) conditional and (f) unconditional score functions. (e,f) are visually indistinguishable because their differences are around 0.01. All colored areas mark the standard deviation across different trajectories.

B Detailed Review of Earlier Work and Theoretical Background

We give a more detailed review of earlier work and theoretical background in this section.

B.1 Review of earlier work

Using a diffusion process for generative modeling was proposed by Sohl-Dickstein et al. [2015] and their modern formulations as DDPM and score-SDE were set in Ho et al. [2020] and Song et al. [2021], respectively. Early empirical work, such as Choi et al. [2022] or Meng et al. [2022], already noted that information injection during denoising is not uniform in time, and these observations rapidly inspired a detailed study into dynamics of denoising.

Its origin in non-equilibrium physics invited a study of generative diffusion via a statistical physics lens. Raya and Ambrogioni [2023] found hints of phase transitions in diffusion models, and further work by Biroli and Mézard [2023] and Ambrogioni [2025] analyze symmetry breaking (origin of phase transitions) toy models for data distributions. Existence of a phase transition in generation motivates understanding the full denoising process, which was analyzed asymptotically by Biroli et al. [2024],and non-asymptotically by Li and Chen [2024]for when features emerge during generation. Other works by Sclocchi et al. [2025b], Ventura et al. [2026], Yaguchi et al. [2025] studied details of the generation process, such as how various regimes and their time-lengths may be determined by the structure of the data itself.

Recent work expands this phase-transition view toward memory, information diagnostics, and control. Pham et al. [2024] study the transition from memorization to generalization through associative-memory spurious states, and Li et al. [2025] develop a broader feature-localization theory for generative models. Handke et al. [2025] measure class-semantic information production along diffusion trajectories, and Stančević and Ambrogioni [2026]connect entropy production, score divergence, and branching dynamics, and Ramachandran et al. [2026] use cross-fluctuations to detect sampling transitions and improve downstream generation and classification. Lu and Tang [2026] study how nonreversible perturbations can steer dynamical regimes by breaking detailed balance. Very recent extensions study entropic signatures of class speciation [Handke et al., 2026],out-of-equilibrium pattern formation in trained diffusion models [Ambrogioni, 2026],and analogous regimes in discrete diffusion models [Takahashi et al., 2026]. Our symmetry breaking probe belongs to this lineage, but instead of inferring semantic commitment from generated samples, we directly measure the classifier-free guidance (CFG) score correction $s_\theta(x_t, t \mid y) - s_\theta(x_t, t)$ as a model-internal field.

A separate line of work characterizes phases by conditional independence and local recoverability. Given two distributions, Sang and Hsieh [2025] proposed the decay length-scale of conditional mutual information (CMI) (called the 'Markov length') as a diagnostic of mixed-state phase transitions, which diverges if the two are in different phases and hence not locally recoverable. Zhang and Gopalakrishnan [2025b] proved that under weak local decoherence, the mixed-state phases are stable, and Yi et al. [2026] proved that finite Markov length is a generic feature of mixed-state phases. Further work by Zhang and Gopalakrishnan [2025a] proved that the Markov length is finite during diffusion for high-temperature Gibbs states while noting that at low temperature, it may diverge.

Hu et al. [2025] brought this Markov-length perspective directly to diffusion models by defining phases of data distributions analogously to mixed state phases: between two phases, local denoisers fail near a narrow transition window where Markov length diverges, while it remains finite away from this window, which was shown explicitly by calculating the CMI during image generation. Prior work by Hu et al. [2026] has further shown that if a distribution admits a diffusion noise path along which the CMI remains short-ranged, then one can construct a generative procedure that samples from this distribution using only local operations. Conversely, distributions with long-ranged CMI are known to be hard to learn [Kumar et al., 2026]. Taken together, these results suggest that the Markov length provides a quantitative indicator of the difficulty of learning and generating a distribution. Our locality-gap heatmaps are an operational analogue of their CMI heatmaps: instead of estimating an information-theoretic decoder, we replace the global denoiser by a local-attention denoiser and directly measure the resulting score and sample degradation. The local-attention intervention itself is related to Swin Transformer by Liu et al. [2021], Neighborhood Attention Transformer by Hassani et al. [2023], neighborhood attention and dilated neighborhood attention by Hassani [2023], Hassani and Shi [2022], DiTFastAttn by Yuan et al. [2024], and Swin DiT by Wu et al. [2025], but our use is diagnostic rather than architectural.

Our contribution is to connect the two phase-transition pictures experimentally. Prior diffusion work asks when semantic decisions, speciation, or collapse occur; prior Markov-length work asks when local neighborhoods suffice for recovery or denoising. We ask whether these two criticalities align in real diffusion transformers by comparing conditioning gaps with locality gaps at the score level and with error-correction and windowed-sampling behavior at the sample level. This also connects to limited-interval guidance by Kynkäänniemi et al. [2024], stage-wise CFG dynamics by Jin et al. [2025], ambient-diffusion studies of memorization by Shah et al. [2025], and latent-hierarchy probing by Sclocchi et al. [2025a].

Understanding locality bias has also been a recurrent topic in literature. Kamb and Ganguli [2024] shows that convolutional diffusion model exploits locality bias explicitly which leads to generalization / creativity. Niedoba et al. [2024] similarly shows the emergence of locality bias in convolutional model and DiT and designs mechanistically interpretable local denoisers to approximate global denoisers. Lukoianov et al. [2025] shows that even for architectures like transformer that are ignorant of locality, they can learn locality structure through data statistics. Intriguingly, while their numerical technique is drastically different, they observe a time window in which model exhibits long-range correlations. This already alludes to the existence of the nonlocality phase transition. Complementary to these works, our main contribution is to understand *when* locality bias can be exploited, and how it correlates with symmetry breaking and semantic forcing.

B.2 Local computation of the score under a Markov property

We provide a more detailed theoretical background about local denoising under approximate Markov property. We encode locality via a rectangular tripartition into the image vector $X = (x_A, x_B, x_C)$ as shown in Figure 1: A is the center patch in the black box, B is the surrounding buffer in gray box, and C is the rest of the image that we fade out. We would like to understand when local generation of A is possible, i.e., when the score on region A can be determined only by the information in B, without needing to access C. This is the case if the distribution $p(x_A, x_B, x_C)$ forms a Markov chain

$$x_A - x_B - x_C, \quad \text{i.e.} \quad p(x_A \mid x_B, x_C) = p(x_A \mid x_B), \tag{5}$$

which we call the *Markov property*. Then the score on region A depends only on (x_A, x_B):

$$\begin{aligned} \nabla_{x_A} \log p(x_A, x_B, x_C) &= \nabla_{x_A} \log\Big(p(x_A \mid x_B)\, p(x_B, x_C)\Big) \\ &= \nabla_{x_A} \log p(x_A \mid x_B). \end{aligned} \tag{6}$$

This motivates a local score approximation to $s_A(x_A, x_B, x_C)$, where

$$s_A^{\text{local}}(x_A, x_B) = \nabla_{x_A} \log p(x_A \mid x_B), \tag{7}$$

which is exact whenever the Markov property holds. We expect the Markov property to hold (at least approximately) in real images. For example, to generate the nose of a dog, the model may only need to look at a local neighborhood containing the snout and eyes, and not the rest of the image.

In practice, the Markov property is only approximate. A standard way to quantify this is the conditional mutual information (CMI)

$$I(A : C \mid B) = \mathbb{E}_{p(x_A, x_B, x_C)}\left[\log \frac{p(x_A, x_C \mid x_B)}{p(x_A \mid x_B) p(x_C \mid x_B)}\right]. \tag{8}$$

We define the Markov length ξ as the length-scale of the empirical decay of CMI with separation: for a neighborhood B of radius r around A and its complement C, we posit

$$I\big(A : C \mid B\big) \approx I_0 \, \exp(-d_{AC}/\xi)\,, \tag{9}$$

where d_{AC} is a distance between regions A and C (typically $d_{AC} = r$ when B_r is an r-neighborhood) and I_0 is a distance-independent scale constant that sets the magnitude of the CMI. Eq. (9) is an ansatz describing how the CMI of a given distribution decays, not a derived equality. Intuitively, ξ is the length scale beyond which information outside the buffer B is negligible for predicting A. This was formalized by Hu et al. [2025] as Theorem 1 therein:

Proposition 1 (Local denoising from finite Markov length)**.** *Consider the $L_x \times L_y$ image space $X \in \mathbb{R}^{L_x \cdot L_y}$. Fix a diffusion time t and let $p_t(X)$ denote the noisy image distribution. Suppose the Markov length $\xi(t) \le \xi_{\max}$ for all $t \le t_c$. Then there exists a local denoiser with radius $O(\xi_{\max} \log(L_x L_y/\epsilon))$ that can produce $p_0(X)$ given $p_t(X)$ with $t \le t_c$, up to total variation distance of ϵ.*

Operationally, if Eq. (9) holds, the error of recovering a center patch decreases exponentially as the conditioning window grows, and the Markov length is the scale beyond which it no longer decreases appreciably. Hu et al. [2025] estimate the CMI of MNIST and FashionMNIST variationally and observe this decay.

Local Gibbs distributions have zero Markov length. By the Hammersley–Clifford theorem, under a positivity condition, Gibbs distributions with geometrically local interactions are Markov random fields. Concretely, let B shield A from C (every path from A to C passes through B), with thickness larger than the interaction range. Writing the distribution as the exponential of a sum of energy terms, no term couples x_A and x_C, so

$$p(x) \propto e^{H(x_A)+H(x_A,x_B)}\, e^{H(x_C)+H(x_B,x_C)}\, e^{H(x_B)},$$

and conditioning on x_B gives $p(x_A, x_C \mid x_B) = p(x_A \mid x_B)\, p(x_C \mid x_B)$. Hence $I(A : C \mid B) = 0$ and the Markov length is zero. Our motivation requires only approximate local recovery, for which a finite Markov length suffices; exact local recovery is not necessarily desirable, since some approximation may help generalization [Kamb and Ganguli, 2024].

In general, even if the data distribution satisfies the Markov property exactly, the noised distribution p_t may not, and the Markov length $\xi(t)$ may vary, perhaps even diverge, as a function of t [Hu et al., 2025]. The divergence of $\xi(t)$ at some critical time t_c is understood as a nonlocality phase transition.

CMI and the data-processing inequality. The CMI does not satisfy the data-processing inequality in the conditioning variable. Hence a clean distribution satisfying Eq. (9) can, after independent local noise is applied to each variable, have a Markov length that grows and even diverges; this is what allows CMI to diagnose phase transitions along the diffusion process. By contrast, two-point correlations, such as the mutual information between distant regions, do satisfy the data-processing inequality under independent local noise, so if they decay in the clean data they keep decaying along the entire forward process. Unlike in equilibrium thermodynamics, they therefore cannot locate the critical time of this non-equilibrium process.

Several theoretical works have characterized how $\xi(t)$ behaves in discrete spin systems, in both classical and quantum settings. Zhang and Gopalakrishnan [2025a] shows that for high-temperature Gibbs states, $\xi(t)$ remains finite for all t up to one. This indicates that if a state is in the high-temperature trivial thermodynamic phase, then nonlocality phase transition cannot happen during diffusion. At low temperature, $\xi(t)$ can diverge at some t_c less than one. In fact, one can prove that it has to if the state exhibits long-range two-point correlations, through a simple light-cone argument. This indicates that nonlocality phase transition can be inevitable in sampling low-temperature distributions. Nevertheless, Zhang and Gopalakrishnan [2025b] proves that t_c has to be bounded away from zero by a constant amount, indicating that the low-temperature phase is *stable*. Operationally, this implies that a local denoiser should always exist at sufficiently low noise rate.

So far, all existing rigorous results only apply to discrete spin systems and extending the results to continuous systems remains an open challenge. We expect many of the proof techniques to break down due to physical reasons, such as the existence of Goldstone mode which challenges stability [Mermin and Wagner, 1966], and the existence of entropic ordering at high temperature, which challenges the lack of phase transitions [Han et al., 2025]. These examples are fine-tuned so we do not expect them to appear in realistic datasets. However, any rigorous results and proof technique would have to be able to rule out these adversarial examples.

B.3 Classifier-free guidance as a symmetry-breaking field

In many applications we want to sample the conditional distribution $p_\theta(x \mid y)$ for a specified condition y (in addition to the marginal $p_\theta(x)$), which requires access to the conditional score $\nabla_{x_t} \log p_\theta(x_t \mid y)$ for all t. The modern approach to obtain a conditional sampler is known as classifier-free guidance (CFG) [Ho and Salimans, 2022], where the *same* parameters learn both a conditional score $s_\theta(x_t, t \mid y)$ and an unconditional score $s_\theta(x_t, t)$. During sampling, CFG mixes the two scores to form a 'guided score'

$$s_{\mathrm{cfg}}(x_t, t \mid y) = s_\theta(x_t, t) + w\big(s_\theta(x_t, t \mid y) - s_\theta(x_t, t)\big), \tag{10}$$

where $w \geq 0$ is the guidance scale. To make contact with physics, we write scores as gradients of energies U such that $s = -\nabla U$, and CFG can be viewed as sampling from an energy tilted by an

external field:

$$U_{\mathrm{cfg}}(x_t, t \mid y) = (1-w)U_{\mathrm{uncond}}(x_t, t) + w\, U_{\mathrm{cond}}(x_t, t \mid y) + \text{const.} \tag{11}$$

Viewing each class as a local minimum of the conditional energy landscape $U_{\mathrm{cond}}(x_t, t \mid y)$, the CFG weighted term plays the role of a bias that breaks the symmetry. The absence of a symmetry-breaking field implies that there needs to be spontaneous symmetry breaking for the model to pick a class, which has been associated with experimental failures of forward-backward experiments [Ambrogioni, 2026, Sclocchi et al., 2025b]. Conditioning does not remove symmetry breaking altogether: sampling still starts from a unimodal Gaussian, and the conditional distribution can be multimodal, for example when the label does not fully resolve the class structure of the data, so the model may still need nonlocal computation.

B.4 Why the two transitions may coincide

This subsection collects heuristic and theoretical arguments for why the two transitions may coincide. None of them is a general theorem; they motivate, rather than establish, the empirical observations of the main text.

Semantic information is nonlocal. By long-range correlations we mean semantic information, such as object identity, shape, pose and spatial layout, that constrains local features to be mutually consistent. If an image depicts a cat rather than a dog, distant regions must jointly exhibit compatible ears, muzzle, paws, fur texture, body proportions and tail. During sampling, the model must commit to a semantic mode and propagate that choice across the image: this is where symmetry breaking happens, and it is also where nonlocal denoising is needed, because features generated in distant regions must agree with the choice. Once the mode is selected, most subsequent denoising fills in textures, edges and other local details, and can be performed locally.

An error-correction view. Many semantics are robust to weak noise (a weakly noisy dog is still a dog), so data with different semantics can be viewed as codewords of an error-correcting code, and the symmetry-breaking critical time as the noise level beyond which the semantics can no longer be read off. For data with locality structure, decoding under weak noise can be local; for example, good low-density parity-check codes under weak noise can be decoded by belief propagation with a bounded number of local message-passing steps. As the noise grows, decoding requires aggregating evidence from larger neighborhoods; at the decodability threshold no bounded radius suffices in the large-system limit, and above it the codeword cannot be recovered even globally. Concurrence is therefore expected when semantics behave as noise-robust clusters. There is no reason to expect all semantics to do so; our forward-backward experiments (Section 3.3) indicate that the classes we test are robust in this sense.

A solvable toy model: the noisy repetition code. Consider n spins on a two-dimensional lattice at zero temperature, so that the clean distribution puts probability $1/2$ on the all-down and $1/2$ on the all-up configuration; the semantic class is the phase $s \in \{\pm 1\}$. Noise flips each spin independently with probability p, and $p = 1/2$ is maximal noise. For $p < 1/2$, a global majority vote recovers the phase with error probability at most $\exp\{-2n(1/2-p)^2\}$ by Hoeffding's inequality, so global recovery succeeds with high probability whenever $(1/2-p)^2 n \to \infty$. A local recovery map sets each spin to the majority of its radius-L square neighborhood $N_L(i)$. With $m_i = |N_L(i)|$ and $m_{\min} = \min_i m_i$, Hoeffding's inequality and a union bound give

$$\Pr(\hat{s}_i \neq s) \le \exp\{-2m_i(1/2-p)^2\}, \qquad \Pr(\exists i : \hat{s}_i \neq s) \le n \exp\{-2m_{\min}(1/2-p)^2\},$$

so exact recovery holds with probability at least $1 - n^{-k}$ whenever $2m_{\min}(1/2-p)^2 \ge (k+1)\log n$, that is, up to boundary constants, $L^2 = \Omega\big(\log n/(1-2p)^2\big)$ suffices, i.e., a radius $L = O\big(\sqrt{\log n}/(1-2p)\big)$. For each fixed $p < 1/2$ a logarithmic window area suffices, while the required radius diverges like $(1-2p)^{-1}$ as $p \uparrow 1/2$; at $p = 1/2$ the observations are independent of the phase and no local or global rule beats chance. In this toy model the point where the required local radius diverges is exactly the information-theoretic threshold of the noisy repetition code. In general the critical point need not lie at maximal noise: in the toric code under noise, analyzed by Sang and Hsieh [2025], the relevant phenomenon can be described by a classical loop gas whose semantic information is the homology class of loops winding around the lattice, and the noise level at which

broken loops can no longer be repaired locally coincides with the one at which they corrupt the homology class, at a finite critical point. This argument controls only the regime below the critical point, where semantic information is still present, not the behavior near maximal noise.

When concurrence may fail. Our working conjecture is that the correspondence holds when the data has well-defined classes and long-range correlations are mediated by class membership: the CMI decays with distance away from class mergers or splits, while a merger or split produces the nonlocal critical window. If the two transitions had different mechanisms, the two diagnostics could identify different noise windows: a clustering probe could show a merger or split at one time while the locality probe peaks at another. A topological transition could produce a locality change without a clear cluster merger or split; in a Mexican-hat landscape, for example, a point-like region becomes a ring of minima. Overlapping hierarchical transitions could appear as several unresolved or broadened peaks rather than a single aligned window. All models we test are DiTs, chosen so that locality can be tuned through the attention radius; related phenomenology has been reported for UNet-based models on smaller datasets with a different methodology [Hu et al., 2025]. We regard universality across architectures as a working hypothesis, not a result.

C Details on Experimental Methodology

We provide details about all experiments in this section. All experiments are run on at most a single 80G A100 GPU. Most jobs are run on a 1/7 MIG partition of a 80G A100 GPU. All job times are less than a day.

C.1 Local attention implementations

All local-attention experiments use the same pretrained backbone as the corresponding global baseline and modify only the attention pattern. We replace the attention layer with a radius-r Chebyshev window on the 2D token grid. Queries, keys, and values are projected with the original layer weights; for each image token, the module gathers only keys and values whose grid coordinates satisfy $\max(|\Delta x|, |\Delta y|) \leq r$, applies softmax within that window, and reuses the original output projection. For MMDiT joint attention: image-to-image attention is windowed, while text/context tokens remain globally visible so that conditioning is not sparsified.

In practice, this can be implemented directly, or can be implemented by leveraging masking options in existing attention kernels, which gives a significant speedup.

C.2 Conditioning and Locality Gap

The conditioning and locality gap are defined in the main text. The training trajectory is generated by taking an image from the dataset, and evolving it forward by adding noise. We only compute the training trajectory for Facebook DiT since its training dataset is ImageNet which is publicly accessible [Deng et al., 2009]. Also note that there is no distinction between conditional and unconditional training trajectory. The sampling trajectory is generated by starting from complete noise and evolving it backward using the diffusion model. Sampling trajectory can be either conditional or unconditional. We also note that all score gaps and their l_2 norms are directly computed in the latent space. We observe that score gaps in latent space translate to a larger gap in the pixel space when t is closer to zero. This tilts the score gap plot towards the beginning. We hypothesize that this enhanced sensitivity is due to the formation of data manifold near clean image which could enhance perturbation in the latent space. We report the l_2 norm divided by the number of image tokens.

For Facebook DiT-XL results presented in Figure 3, we take 500 steps along the training and sampling trajectory. We take 20 samples at each time step. We use the HuggingFace diffusers implementation of DPMSolverMultistepScheduler [Lu et al., 2022, 2025]. Image size is fixed to 256×256 by the model. For conditioning, we use class 207 (golden retriever). We do not observe qualitative differences between different class labels (see also the sweeps in Appendix E).

For SD3 medium, we take 100 steps along the sampling trajectory. We take 10 samples at each time step. We use the StableDiffusion3Pipeline from HuggingFace diffusers. We use an image size of 512×512. We use a CFG strength of 4.0 in both cases. For conditioning, we use prompt "a golden

retriever playing in a park, high detail, soft lighting". Again, we do not observe qualitative differences between different prompts (Appendix E).

For both models, the VAE has a eight-to-one compression ratio, and they both organize a two-by-two latent patch into a token. Therefore, both models have a token patch size of 16×16 in the pixel space. This means that for Facebook DiT-XL, setting $r = 16$ reduces local attention to global dense attention. For SD3 medium in our setup, this r becomes 32.

Subsequent experiments inherit the same setup unless otherwise noted.

C.3 Forward-backward experiments

For Facebook DiT, we use 20 clean-generation steps, 50 denoising steps, and evaluate at 40 linearly spaced noise levels with $t_{\mathrm{norm}} \in [0.1, 1.0]$. For each noise level, the experiment draws 40 independent noise realizations.

For both models, we use microsoft/resnet-50 for classification [He et al., 2016, Hugging Face and Microsoft, 2025]. We report the standard deviation of the classifier error which is analytically determined from the Bernoulli distribution, since classification is either correct or incorrect. For a datapoint with classification error p and with n samples, the one-sigma error is given by $\sqrt{p(1-p)/n}$.

C.4 Conditioning window and locality window experiments

We use 40 sampling steps for all window experiments. For conditioning window experiments for both models, we use a window size of 0.1 and scan t_i from 0 to 0.9 by steps of 0.1 (setting $t_i > 0.9$ truncates the window size). We use 500 samples per t_i. For locality window experiments for Facebook DiT, we use a window size of 0.4 and scan t_i from 0.2 to 0.8 by steps of 0.05 (setting $t_i > 0.8$ or $t_i < 0.2$ truncates the window size). We use 500 samples per t_i. For SD3, we start t_i at 0.4 and use 62 samples per t_i to reduce compute.

Our FID follows from the standard definition [Heusel et al., 2017]. We use the pool3 layer of the inception V3 network [Szegedy et al., 2016] as features in computing the Gaussian Fréchet distance. We are in a sample-deficient regime, which biases FID significantly upward; our analysis uses only the t_i at which FID is smallest, not its absolute value. We compute the one-sigma error by bootstrap resampling.

D Limitation of Our Work

Our empirical study is currently limited to two representative diffusion transformer systems: a class-conditional ImageNet DiT and a text-conditioned MMDiT/SD3 model. Score-gap probes were additionally run on SD3.5 Large [Stability AI, 2026] and FLUX.2 Klein 9B [Black Forest Labs, 2026] (Appendix E); integrated sampling and forward-backward tests on these models are not included in the present evaluation. Although these models cover two common conditioning mechanisms and show consistent concurrence between the symmetry-breaking and nonlocality critical windows, a broader evaluation across architectures, model scales, datasets, samplers, and prompt distributions is needed before claiming universality. Because DiT-XL and SD3 differ in architecture, conditioning, data, resolution, noise schedule and parameterization, comparisons between them do not isolate model quality; for example, the larger forward-backward threshold of SD3 does not test whether a stronger model is a better decoder. In addition, our truncated-attention DiT should be understood as an approximate local denoiser rather than a strictly local one, where information can still propagate across distant patches through multiple layers of local attention. It is an architectural intervention on a pretrained model that may introduce distribution shift, and it is neither a strict local score estimator nor a direct measurement of CMI. In our main experiments the conditioning signal remains globally available to every patch (through AdaLN in DiT and dense text attention in MMDiT), so the intervention restricts image-token communication only. A strictly local center-patch control (Appendix E) broadly reproduces the score-gap behavior of the truncated model but aligns less tightly with the conditioning gap: the Pearson correlation is 0.432 for SD3 and 0.583 for DiT-XL sampling, and the fitted-peak intervals do not overlap in these two cases (Table 2). Constructing a numerically faithful strictly local score estimator for modern pretrained diffusion models is itself challenging;

future work could test whether the same concurrence persists under more diverse architectures and under stricter or independently trained local denoisers.

Our conclusions are also empirical rather than theoretical. The observed concurrence is supported by score-gap measurements and windowed sampling interventions, but we have not proven that symmetry-breaking and nonlocality transitions must coincide in general; Appendix B.4 gives heuristic arguments for when they should coincide and when they may not. We do not specify an equivalence margin for concurrence: several within-probe uncertainty intervals do not overlap, and the SD3 integrated-window grids are not matched (Appendix E.1). Critical times also depend on the probe and the estimator: for DiT-XL, score-gap fitted peaks, score-gap centers of mass and window optima lie near 0.06–0.09, 0.30–0.32 and 0.43–0.50, and center-of-mass discrepancies reach 0.27 for SD3 with DPM samplers. Schedule-aware variants of the score gaps give mixed results, and how instantaneous score perturbations determine integrated sampling behavior remains open (Appendix E). Our sample-level metrics, classifier error and FID to an always-conditioned or always-global baseline, measure relative damage; they do not fully establish semantic alignment or perceptual quality, especially for text prompts beyond ImageNet-like categories, and the SD3 locality-window scan uses only 62 samples per t_i. Some quantitative estimates, especially FID and sliding-window results, are based on finite samples and should be interpreted as evidence for relative timing rather than precise estimates of the phase transition window. Developing sharper statistical tests and a theory connecting these operational probes to Markov length or conditional mutual information remains an important direction for future work.

We are also not able to quantify the width of the critical window. We expect that different definitions give rise to different width. Only in well-defined "thermodynamic limit" can one give a unified definition (and critical window can also collapse to a critical point in this limit). Ideally, it would be useful to understand the criticality better by probing certain universal scaling behavior as in statistical mechanics. However, it is currently unclear what is the correct scaling variable and scaling hypothesis that captures the behavior of diffusion models. We leave such a scaling analysis to future work.

The design and sampling directions in the Conclusion are untested: we do not measure wall-clock or FLOP savings from local attention, and the calibrated samplers of Appendix F are not evaluated.

E Quantitative and robustness analyses

This appendix reports the quantitative within-probe comparison, locality controls and robustness analyses. Peaks are fitted by a local quadratic around the observed maximum; for score gaps we also use a center of mass (COM) weighted by the positive score-gap magnitude; locality score probes use $r = 8$. Intervals are 5%–95%, from trajectory-level resampling for score gaps and from bootstrapping the FID and classification-error estimates for windows. All $|\Delta t|$ values are calculated from the reported centers and are not separate experiments.

E.1 Quantitative within-probe comparison

We fix a probe and a critical-time estimator (a fitted peak, or for score gaps a center of mass, COM; Appendix E) and extract one critical time from nonlocality data, t_{loc}, and one from conditioning data, t_{cond}. Table 1 in Appendix E lists the estimates and the discrepancy $|t_{\mathrm{loc}} - t_{\mathrm{cond}}|$, which we calculate from the reported centers. Within a probe and estimator the discrepancy is 0.004–0.043 for score gaps and about 0.01 for the DiT-XL windows. The Pearson correlation between locality and conditioning data is 0.976 (DiT-XL) and 0.851 (SD3) for score gaps, and 0.73–0.88 for integrated windows. We do not specify an equivalence margin, five of the eight interval pairs do not overlap, and the SD3 window grids are not matched, so we describe the result as approximate within-probe temporal alignment. Across probes and estimators the times differ substantially: for DiT-XL, score-gap fitted peaks lie near 0.06–0.09, score-gap COMs near 0.30–0.32, and window optima near 0.43–0.50. This is expected for a finite system without a scaling limit.

Schedule-aware variants of the score gaps give mixed results, and sweeps over guidance scale, prompt categories, samplers, noise schedules and two further DiTs give locality–conditioning correlations of 0.60–0.98, with fitted-peak discrepancies below 0.09 but COM discrepancies up to 0.27 (Appendix E).

Table 1: **Within-probe critical times.** Critical times with 5%–95% intervals ($r = 8$); $|\Delta t|$ is calculated from the reported centers. †SD3 window grids are not matched (locality grid ends near 0.8, conditioning grid at 0.95).

Probe, estimator	Model	t_{loc}	t_{cond}	$\lvert\Delta t\rvert$	Overlap
Score gap, fitted peak	DiT-XL	0.061 [0.054, 0.074]	0.085 [0.071, 0.113]	0.024	yes
Score gap, COM	DiT-XL	0.318 [0.316, 0.320]	0.304 [0.302, 0.306]	0.014	no
Score gap, fitted peak	SD3	0.991 [0.991, 0.992]	0.995 [0.994, 0.998]	0.004	no
Score gap, COM	SD3	0.882 [0.878, 0.885]	0.925 [0.913, 0.934]	0.043	no
Window, FID peak	DiT-XL	0.496 [0.446, 0.548]	0.484 [0.449, 0.529]	0.012	yes
Window, class-error peak	DiT-XL	0.443 [0.415, 0.488]	0.432 [0.407, 0.453]	0.011	yes
Window, FID peak†	SD3	0.832 [0.800, 0.849]	0.950 [0.950, 0.950]	0.118	no
Window, class-error peak†	SD3	0.835 [0.807, 0.900]	0.950 [0.950, 0.950]	0.115	no

Correlations for integrated windows. For the integrated-window probes, the Pearson correlation between locality and conditioning data is 0.842 (FID degradation) and 0.884 (classifier error) for DiT-XL, and 0.727 and 0.779 for SD3; for SD3 the grids are not matched (Table 1).

Strict center-patch locality control. At a point along the sampling (or training) trajectory we pass in only the image tokens near a center patch, together with the text tokens for MMDiT, perform dense attention, and use only the center-token output to compute score gaps. This is inefficient, but gives a strictly local one-patch denoiser, whose score-gap behavior is broadly similar to that of the masked-attention denoiser in both norm and cosine similarity. Table 2 reports the statistics given with this control. Correlations are lower than for the masked-attention probe in several rows (SD3: 0.432; DiT-XL sampling: 0.583), and the DiT-XL sampling-trajectory fitted peaks differ by 0.172 with non-overlapping intervals. Locality gaps computed without conditioning behave similarly. We do not believe the conditioning pathway can be fully localized, since semantic information is nonlocal.

Table 2: **Center-patch control.** Pearson correlation between locality and conditioning data, and fitted-peak and COM critical times with 5%–95% intervals. $|\Delta t|$ is calculated from the reported centers; "no" marks non-overlapping intervals.

Model	Traj.	Pearson	Loc. fit	Cond. fit	$\lvert\Delta t\rvert$ fit	Loc. / Cond. COM	$\lvert\Delta t\rvert$ COM
DiT-XL	train	0.840	0.173 [0.079, 0.323]	0.184 [0.082, 0.311]	0.011	0.222 / 0.214	0.008
DiT-XL	sample	0.583	0.032 [0.002, 0.064]	0.204 [0.099, 0.424]	0.172 (no)	0.168 / 0.242	0.074
SD3	sample	0.432	0.928 [0.890, 0.950]	0.983 [0.970, 0.998]	0.055 (no)	0.903 / 0.935	0.032
SD3.5	sample	0.552	0.900 [0.847, 0.969]	0.982 [0.866, 1.000]	0.082	0.819 / 0.911	0.092
FLUX.2	sample	0.973	0.999 [0.933, 1.000]	0.998 [0.992, 1.000]	0.001	0.921 / 0.958	0.037

COM intervals (loc. / cond.): DiT-XL train [0.153, 0.241] / [0.169, 0.280]; DiT-XL sample [0.129, 0.195] / [0.179, 0.318]; SD3 [0.871, 0.912] / [0.905, 0.960]; SD3.5 [0.795, 0.858] / [0.825, 0.935]; FLUX.2 [0.877, 0.964] / [0.893, 0.973]. All COM interval pairs overlap.

Spatial versus random masks. We repeated the experiment of Figure 6(a)(ii, top), with global denoising inside the window, replacing the spatial local mask outside the window by a random fixed mask that keeps the same number of unmasked tokens. Random masking produces much larger score gaps and heavily corrupts samples even when the global denoiser is used near the critical window (Table 3), so the locality gap reflects spatially organized communication rather than the number of visible tokens alone.

Table 3: **Spatial versus random masks** outside the global window (16 samples per row).

Model	Mask outside global window	FID to global baseline	Classifier errors
DiT-XL	spatial local	57.6	0/16
DiT-XL	random fixed	206.8	3/16
SD3	spatial local	2.93	0/16
SD3	random fixed	297.6	14/16

Robustness sweeps and additional models. Score-gap experiments were repeated across CFG strength, class labels and prompts, SD3 prompt categories, samplers and noise schedules, and two additional text-to-image DiTs, SD3.5 Large [Stability AI, 2026] and FLUX.2 Klein 9B [Black Forest Labs, 2026]. For the additional models only score-gap probes are reported; integrated sampling and forward-backward tests on them are not included in the present evaluation. Table 4 lists the reported values. The Pearson correlations range from 0.603 to 0.977. Fitted-peak discrepancies are at most 0.082 (DiT-XL, DPM default). COM discrepancies are larger for SD3 and reach 0.272 with DPM samplers, so agreement under one estimator need not carry over to another.

Table 4: **Robustness of the score-gap comparison**: locality/conditioning critical centers for the fitted peak and the COM, and the Pearson correlation. $|\Delta t|$ is calculated from the reported centers. Random-prompt rows compare models; the other rows vary one setting.

Sweep	Model	Setting	Pearson	Fit loc./cond.	$\lvert\Delta t\rvert$	COM loc./cond.	$\lvert\Delta t\rvert$
Random prompts	DiT-XL	—	0.945	0.006 / 0.045	0.039	0.302 / 0.301	0.001
	SD3	—	0.848	0.988 / 0.990	0.002	0.890 / 0.927	0.037
	SD3.5	—	0.697	0.942 / 1.000	0.058	0.828 / 0.896	0.068
	FLUX.2	—	0.787	0.982 / 0.993	0.011	0.857 / 0.896	0.039
CFG scale	DiT-XL	$g = 2$	0.975	0.044 / 0.077	0.033	0.314 / 0.293	0.021
		$g = 4$	0.977	0.061 / 0.085	0.024	0.318 / 0.304	0.014
		$g = 6$	0.969	0.043 / 0.074	0.031	0.311 / 0.292	0.019
	SD3	$g = 2$	0.603	0.983 / 0.992	0.009	0.856 / 0.947	0.091
		$g = 4$	0.776	0.992 / 0.995	0.003	0.882 / 0.922	0.040
		$g = 6$	0.862	0.984 / 0.984	0.000	0.874 / 0.932	0.058
Prompt category	SD3	multi-object	0.741	0.943 / 0.984	0.041	0.885 / 0.936	0.051
		non-animal	0.901	0.980 / 0.980	0.000	0.909 / 0.951	0.042
		relational/spatial	0.811	0.981 / 0.988	0.007	0.894 / 0.952	0.058
		spatial composition	0.930	0.987 / 0.987	0.000	0.898 / 0.955	0.057
		texture-dominated	0.856	0.990 / 0.984	0.006	0.861 / 0.953	0.092
Sampler / schedule	DiT-XL	DPM default	0.904	0.036 / 0.118	0.082	0.201 / 0.232	0.031
		DDIM default	0.905	0.052 / 0.129	0.077	0.229 / 0.257	0.028
		DPM trailing	0.903	0.036 / 0.116	0.080	0.202 / 0.233	0.031
		DDIM trailing	0.947	0.049 / 0.098	0.049	0.213 / 0.229	0.016
	SD3	FlowMatch default	0.871	0.987 / 0.989	0.002	0.873 / 0.939	0.066
		FlowMatch shift 1	0.940	0.972 / 0.971	0.001	0.728 / 0.864	0.136
		FlowMatch shift 5	0.849	0.992 / 0.993	0.001	0.908 / 0.949	0.041
		DPM default	0.885	0.898 / 0.898	0.000	0.633 / 0.905	0.272
		DPM trailing	0.885	0.898 / 0.898	0.000	0.634 / 0.905	0.271

Schedule-aware diagnostics. We considered three schedule-aware variants of the score gaps: the cosine deviation $\cos(s_t, s'_t) = \langle s_t, s'_t\rangle/(\|s_t\|\,\|s'_t\|)$, which measures directional change but not magnitude; the gap weighted by the reverse-SDE drift coefficient, $g(t)^2\big(\mathbb{E}_{p_t}\|\Delta s_t\|^2\big)^{1/2}$; and the gap weighted by the interpolant noise variance, $\sigma_t^2\|\Delta s_t\|$. We use the weighted gaps as heuristic measures of how strongly a score perturbation enters the sampling dynamics, not as bounds on the final-sample error. In DiT-XL, the peak disappears in cosine deviation for both the local-versus-global and the conditional-versus-unconditional comparison, so the original feature is mainly a magnitude effect; after schedule normalization it shifts from about $t = 0.2$ to $t = 0.3$, closer to but still distinct from the $t \approx 0.5$ forward-backward threshold. In SD3, the $g(t)^2$-weighted gap still peaks near $t = 1$, while the σ_t^2-weighted peak is less conspicuous. How instantaneous score perturbations determine integrated sampling behavior therefore remains open.

F Proposal of Calibrated local/global sampling

The score-gap curve, and hence the critical window, is typical across trajectories (Figure 9) and, per the sweeps of Appendix E, broadly similar across classes and prompts. Under this premise the window need not be detected online for each sample; it can be calibrated once before inference. We give two pseudo-algorithms. We have not evaluated their sample quality or computational cost.

Let $s_{\mathrm{glob}}(x_t, t)$ be the global denoiser and $s^{(r)}_{\mathrm{loc}}(x_t, t)$ the local denoiser of radius r, and let

$$\Delta_r(t) = \mathbb{E}_{x_t}\left[\|s_{\mathrm{glob}}(x_t, t) - s^{(r)}_{\mathrm{loc}}(x_t, t)\|^2\right]$$

be the average score gap at time t, estimated on a small calibration set.

Algorithm 1: fixed radius, calibrated global window. Parametrize the global window by a center c and width w, $W(c, w) = \{t : |t - c| \le w/2\}$. Given a budget of B discretized timesteps of global computation, choose

$$(c^\star, w^\star) = \arg \max_{c,w:\, |W(c,w)|=B} \sum_{t \in W(c,w)} \Delta_r(t),$$

equivalently minimizing $\sum_{t \notin W(c,w)} \Delta_r(t)$.

1. *Calibration* (once; input r, B, calibration samples): estimate $\Delta_r(t)$ for every diffusion time t, then choose $(c^\star, w^\star)$ maximizing the score gap covered by $W(c, w)$ subject to $|W(c, w)| = B$.
2. *Inference*: at each reverse step t, use $s_{\mathrm{glob}}(x_t, t)$ if $t \in W(c^\star, w^\star)$ and $s^{(r)}_{\mathrm{loc}}(x_t, t)$ otherwise.

A larger budget allows a wider window, or several high-gap windows, which covers at least as much score-gap mass.

Algorithm 2: radius calibrated by a score-gap tolerance. Given a tolerance ϵ and candidate radii R that include $r = \infty$, with $s^{(\infty)}_{\mathrm{loc}} = s_{\mathrm{glob}}$ and hence $\Delta_\infty(t) = 0$, set

$$r^\star(t) = \min\{r \in R :\ \Delta_r(t) \le \epsilon\}.$$

1. *Calibration* (once): for every t and every $r \in R$, estimate $\Delta_r(t)$ and record $r^\star(t)$.
2. *Inference*: at each reverse step t, use $s^{(r^\star(t))}_{\mathrm{loc}}(x_t, t)$, which is the global denoiser whenever $r^\star(t) = \infty$.

Outside the critical window a small radius often suffices because the score gap is small, while inside it $r^\star(t)$ grows sharply and may become global, so the calibrated schedule $r^\star(t)$ is another operational readout of the bifurcation structure.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our main scientific and technical contributions are clearly stated in the abstract and introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in Section D.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The paper's own derivations are elementary illustrations (Appendices B.2–B.4: the local Gibbs factorization, the classifier-guidance decomposition and the noisy repetition-code bounds); each states its assumptions and gives a complete derivation. All other theoretical results are explicitly cited from other works, and the heuristic arguments of Appendix B.4 are labeled as such.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We have provided detailed discussion in Section C. We have also provided all code and generated data.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We have also provided all code and generated data.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We have provided detailed discussion in Section C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Classifier-error and FID results are reported with one-sigma error bars, whose computation is described in Section C right after the corresponding experiment; score-gap variability across trajectories is shown in Figure 9, and Tables 1 and 2 report 5%–95% intervals (Section E).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have discussed the compute resources needed in the beginning of Section C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This work is in compliance with NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This work is about validating the phenomenology of phase transition in diffusion models. At this point, we do not see any direct societal impact of this phenomenology and our validation of this phenomenology.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This work does not pose such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

    Justification: The models and datasets are cited in the paper; their license names and terms of use are not explicitly listed.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: All codes and data are included and documented.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We acknowledge the use of Codex in generating all code for all experiments and the use of ChatGPT in writing this manuscript. All scientific arguments and core methodologies are generated by humans.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.