BREAKDOWN OF LOCAL DENOISING AS SEMANTIC SPECIATION

Anonymous authors
Paper under double-blind review

ABSTRACT

The dynamics of generative models exhibit two apparently distinct temporal windows: a speciation window, in which a sample commits to a semantic class, and a nonlocality window, in which local context windows become insufficient for generation. Motivated by evidence of their near-concurrence in a variety of frontier models, we investigate their relationship through the spatial distribution of semantic information. Under a "common cause" hypothesis, we prove that the nonlocality window must lie in the speciation window. This hypothesis postulates that semantic labels explain a fraction of the correlations between distant tokens, a condition that is natural for many real datasets. We further give conditions under which both windows shrink to a single limiting time as system size grows, defining a "phase transition", and verify this behavior analytically in Gaussian mixtures. Together, these results identify conditions under which semantic information explains the concurrence of speciation and nonlocality, connecting two complementary perspectives on the emergence of semantic structure in generative modeling.

1 INTRODUCTION

Real-world data often contains distinct semantic classes. An image dataset may contain both cats and dogs; a dataset of English text may include sonnets from Shakespeare as well as essays on history. A good generative model must be able to sample from the underlying distribution and thus reproduce this structure Pham et al. (2024); Shah et al. (2025). Such features must therefore emerge during inference. When during generation is this identity determined, and when can a small perturbation still change it? We call the period during which the sample commits to a semantic class the *speciation window* Biroli et al. (2024); Li & Chen (2024). Early empirical studies show that speciation window persists for a very short time Meng et al. (2022); Choi et al. (2022), a phenomenon that has been connected to the thoery of phase transition in physics Raya & Ambrogioni (2023); Biroli et al. (2024); Sclocchi et al. (2025); Takahashi et al. (2026).

A separate question is how much context a model needs when generating one part of a sample and how this context varies during generation. For example, generating a patch of fur may require only nearby texture, while generating an animal's eye might require information from farther away to make sure it only has two, and is at the right location. At some stages, a local neighborhood may be sufficient; at others, restricting the model to that neighborhood may prevent accurate generation. We call the period during which generation requires information of a neighboorhood size that reaches the size of the image *nonlocality window*. Recent work by Hu et al. (2025) has identified this window in simple datasets. Other recent studies analyse the emergence of locality structure through data Lukoianov et al. (2025) and how this structure facilitates generalization Kamb & Ganguli (2024); Niedoba et al. (2024); Hunt et al. (2026), hinting that locality might be a crucial knob to understanding generative modeling.

Although speciation and nonlocality are a priori distinct, recent experiments suggest that their windows closely align. A very recent study Zhang et al. (2026) observed this alignment in two open-source diffusion models, DiT-XL Peebles & Xie (2023) and Stable Diffusion 3 Esser et al. (2024), under an analysis of neural circuitry and controlled experiments. A complementary analysis of patch-based scores connects architectural locality to collective spatial instabilities and the formation of coherent patterns, with growing spatial correlations observed in trained convolutional diffusion

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

models Ambrogioni (2026). Related observations in autoregressive models show that semantic commitment also occurs in short windows beyond diffusion Li et al. (2025). These observations motivate our central question: why should the period when a sample acquires its semantic identity also be the period when generation needs distant context?

In this work, we formalize this connection theoretically under a simple hypothesis—that semantic information is nonlocally encoded. For example, in a dataset containing zebras and leopards, observing stripes in one region can help predict stripes in a distant region because both reflect the animal's species. Unconditioned models develop these coherent features across distant domains, necessitating nonlocal computation to coordinate the generation process. As such, the window of semantic speciation must also include a window of nonlocality, temporally aligning the two phenomena.

To make this intuition precise, we identify two information quantities that characterize semantic speciation and nonlocal computation, and establish a quantitative relation between them. *Mutual information* (MI) between local parts of samples and the class label quantifies the amount of the semantic information exposed in local regions, revealing the speciation window Handke et al. (2025; 2026); *Conditional mutual information* (CMI) between parts of the sample characterizes the error incurred when compute is restricted to a local region Hu et al. (2025), revealing the nonlocality window. We derive upper and lower bounds on CMI using MI's, under a *common-cause hypothesis*: the semantic labels explain a fraction of the dependence across regions. This formalizes the nonlocal encoding of semantic information as contributions to CMI. These bounds result in the containment of the nonlocality window in the speciation window. Since our theory is fundamentally information-theoretic, the theorem applies to the generation dynamics of both autoregressive and diffusion models, independent of the noising process. The concrete examples and experiments in this paper focus on diffusion models; direct tests in autoregressive models are left to future work, although recent work Zhu et al. (2026) observes token-entropy spikes at transitions into erroneous reasoning, suggesting a possible link.

We anchor these statements in exact analytical calculations in Gaussian mixture models, where score functions and the aforementioned information-theoretic functions are obtained. We also derive the time scales of the semantic and nonlocality windows in this setting. A 'thermodynamic limit' (where the data dimension is taken to infinity) closes the semantic and nonlocal windows and yields a sharp *transition*. Beyond Gaussian mixtures, we also construct a more general condition for window closure, requiring semantic classes to separate faster than within-class fluctuations as system size grows. Our results unify semantic and architectural perspectives imposed by dynamics along generation trajectories.

2 Background and Related Work

2.1 Local scores by decaying conditional mutual information

A diffusion model is completely determined by obtaining the score function of the data distribution over diffusion time. If the underlying data distribution is $p_{\rm data}$ the score function takes the form $s(x,t) = \nabla_x \log p_{\rm data}(x,t)$ where $p_{\rm data}(x,t)$ is the data distribution convolved with the Gaussian noise along the diffusion path. We take $X_0 \sim p_{\rm data}$ as a sample from the data distribution, and use the interpolation convention with $0 \le t \le 1$, where

$$X_t = (1-t)X_0 + tZ\,, \quad Z \sim \mathcal{N}(0,\mathbb{I})\,. \tag{1}$$

A brief review of diffusion models is provided in App. A.

Our interest is in local diffusion models, in which the generation of a pixel is determined only by its neighbourhood and not need the whole image Kamb & Ganguli (2024); Niedoba et al. (2024); Hu et al. (2025); Hunt et al. (2026). If such a task is possible, operationally, we expect the score function to take a local form. We define locality via a tripartition of the image, shown at the left/right of the bottom row of Fig. 1, where the pixel (or small patch) to denoise is denoted A, a square annulus around A of radius r is denoted B, and the rest of the image is denoted C. If A is close to the edges of the image, the parts of the regions outside the image domain is ignored. Images are assumed to live in $\mathbb{R}^d$, where $d = N \times N$ and N is the linear dimension; we work with square images for simplicity and details of the results do not depend on the aspect ratio. Each image draw $x \in \mathbb{R}^d$ is then partitioned as $x = (x_A, x_B, x_C)$ and as a shorthand, we denote $x_{AB} = (x_A, x_B)$. Thus,

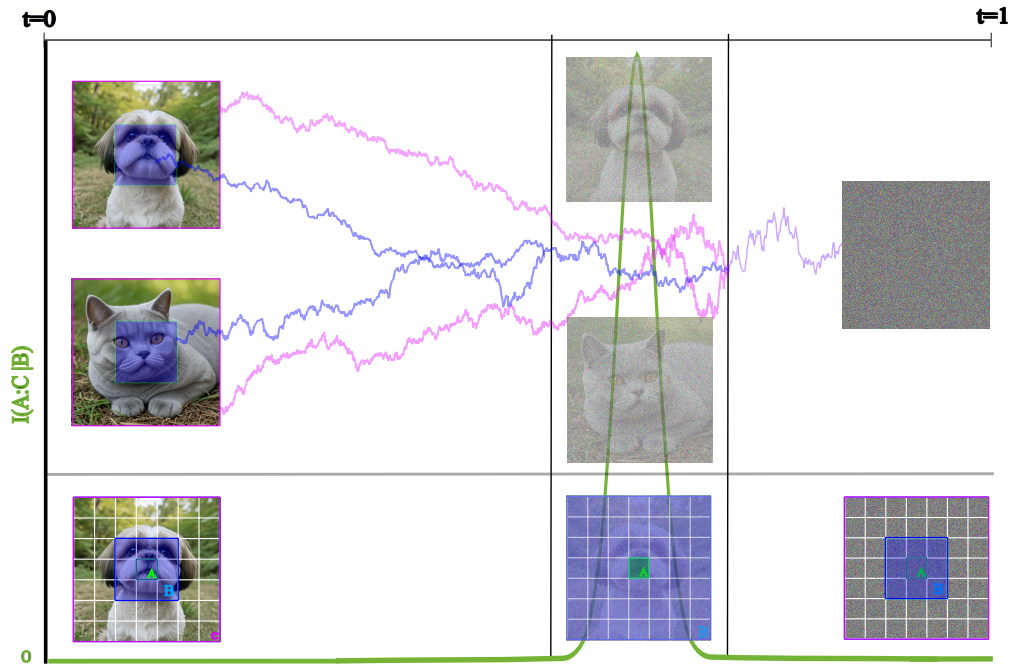


Figure 1: **Semantic speciation locates the window in which local denoising breaks down.** Denoising trajectory is from $t = 1$ (right) to $t = 0$ (left). The upper panel depicts a semantic speciation from white noise to cat or dog for local region (blue) and global region (magenta); the corresponding thresholds define a speciation window (black vertical lines). In the lower row, A is the patch to denoise (green), B its minimal surrounding context (blue) for below-threshold denoising, and C the rest of the image (magenta). The green curve depicts $I_t(A; C \mid B)$ whose peak characterizes a breakdown of exact denoising using local context alone. Under the common-cause hypothesis, Theorem 1 places the threshold-defined nonlocality window inside the speciation window.

implied by the operational constraint, a local score acts on A and depends only on A and B (we use a region R and the pixels x_R in R interchangeably). If the ideal data distribution is known, this would be the score function of the AB marginal, given by $\nabla_A \log p(x_{AB})$ (see equation 34). In the absence of the full data distribution, how can such a score be constructed, and when is it useful to do so?

Inspired from work by Sang & Hsieh (2025) in mixed-state phases in quantum systems, Hu et al. (2025) bounds the recovery error from using a local denoiser instead of the global one, using the decay length scale of the CMI of the ABC tripartition. The CMI is defined as the change in mutual information between A and C remaining after revealing the annulus B, i.e.

$$I(A; C \mid B) = I(A; BC) - I(A; B)\,. \tag{2}$$

The CMI, by definition, is the expectation of the Kullback-Leibler (KL) divergence between the joint conditional $p(x_A, x_C \mid x_B)$ and its factorized conditional $p(x_A \mid x_B)p(x_C \mid x_B)$; if the CMI is zero, the joint distribution on ABC factorizes and the regions form a Markov chain $A - B - C$ Hu et al. (2025); Zhang et al. (2026). As a result, differentiating the score function on A kills the C dependence completely, where $\partial_{x_A} \ln p(x) = \partial_{x_A} \ln p_{AB}(x_A, x_B)$ for $p = p_{AB}p_{C|B}$—giving exact locality of the score function. If the CMI is not zero, it still controls the error in denoising a noisy distribution by bounding the total variation. We relegate the careful definition of the quantities in this statement, as well as proving the statement itself (Theorem 3) to Appendix A.

While the CMI is an information-theoretic quantity that satisfies data-processing inequality in the first two inputs, A and C, and thus would decay if only one was noised, it does not satisfy such an inequality on the conditioned variable B. Adding noise to B may degrade correlations between

A or C and B, strengthening the ones between A and C, leading to an increase in CMI Zhang & Gopalakrishnan (2025). Along the noise trajectory, CMI may grow too long-ranged while the context window B remains small, such that local denoising incurs larger and larger errors. We call the window in which a local denoiser fails and the global denoiser succeeds the "nonlocality window".

A problem with CMI as an empirical probe of phenomenology is that information theoretic quantities are notoriously hard to sample in high-dimensional data Poole et al. (2019). As a consequence, score functions (for diffusion models) are used as operational probes in diagnosing properties of the underlying data structure Premkumar (2026); Zhang et al. (2026). While intuitive, is it justified to study the error in the denoiser itself instead of the errors in its outputs? In Appendix A.4, we prove the following bound

$$\Delta_{\mathrm{loc}}(t)^4 \leq \frac{4d_A(d_A+3)}{t^4} I_t(A:C \mid B)\,, \quad \Delta_{\mathrm{loc}}^2(t) = \mathbb{E}_{p_t(x)} \|\nabla_A \log p_t(x) - \nabla_A \log p_t(x_{AB})\|^2\,, \tag{3}$$

where $d_A = \dim A$, which implies that the "locality gap" Δ_{loc} can be used to probe the CMI. This result complements the results of Hu et al. (2025): their theorem shows that CMI bounds the error incurred by a local denoiser (see Appendix A for the careful statement) whereas our result bounds the score matching error in the denoiser *itself*. This establishes that the locality gap itself can serve as a probe of the nonlocality transition, which already was empirically demonstrated by Zhang et al. (2026).

2.2 Semantic speciation

Another type of transition is studied through semantic speciation: as forward noise increases, the noisy observation carries less information about the original image's class. This is commonly probed by forward–backward (FB) experiments, which denoise a corrupted image and assess whether the reconstruction retains its semantic identity Biroli et al. (2024); Sclocchi et al. (2025); Zhang et al. (2026).

An ideal FB experiment draws a clean sample X_0 which intrinsically comes with a label S_0, adds noise to a region R of the sample to obtain $X_{R,t}$. Given that observation, it draws a fresh clean image and label from the posterior $p(X_{R,0}, S \mid X_{R,t})$. Call the returned label $\widehat{S}$. Finally, it checks whether $\widehat{S} = S$.

Conditioned on $X_{R,t}$, the original label S and the returned label $\widehat{S}$ are independently drawn from the same distribution

$$q_s \equiv p_t(s \mid X_{R,t}).$$

Therefore, the label agreement probability for $X_{R,t}$ is

$$\Pr(S = \widehat{S} \mid X_{R,t}) = \sum_s \Pr(S = s \mid X_{R,t}) \Pr(\widehat{S} = s \mid X_{R,t}) = \sum_s q_s^2. \tag{4}$$

Averaging over the distribution of $X_{R,t}$ gives the success rate of FB experiment for region R at time t,

$$P_{\mathrm{FB}}(R,t) = \mathbb{E} \sum_s q_s^2. \tag{5}$$

Our analysis considers exact posterior sampling and true semantic labels. If the clean observation determines its label, then $P_{\mathrm{FB}}(R,0) = 1$; at complete noise it approaches $\sum_s p(s)^2$, equal to $1/K$ for K balanced classes. Empirical forward–backward experiments, however, need not reach ideal endpoints because the reverse sampler and classifier are imperfect. Sclocchi et al. (2025) report a drop in the peak of the source–reconstruction classifier-logit cosine distribution, while Zhang et al. (2026) report a rise in reconstruction classification error. We also study observations restricted to a spatial region R. Using the same type of classifier-logit cosine probe, our ImageNet (Deng et al., 2009; Russakovsky et al., 2015) crop experiment shows a rapid switch of the empirical peak that occurs earlier for smaller regions (Appendix B). We call the interval of rapid semantic change the "speciation window", whose starting and ending points are marked by the local and global speciation, respectively, see Section 3.2.2.

3 CONNECTING SPECIATION AND NONLOCALITY

We now connect the nonlocality window, in which local context is insufficient for accurate denoising, to the speciation window, in which semantic labels become uncertain under forward noising. After giving an intuitive argument using score functions, we use an information-theoretic formulation to establish conditions under which the nonlocality window is contained in the speciation window.

3.1 A SCORE-BASED ARGUMENT OF WINDOW CONTAINMENT

Here we provide a heuristic argument using score function that builds operational intuition for why nonlocality window is contained in speciation window. A natural way to connect the locality gap to semantic information is to subtract the localized version of Bayes' theorem from its global version:

$$\begin{aligned}\nabla_A \log p_t(x_t) - \nabla_A \log p_t(x_{AB,t}) = &- \big[\nabla_A \log p_t(s \mid x_t) - \nabla_A \log p_t(s \mid x_{AB,t})\big] \\ &+ \big[\nabla_A \log p_t(x_t \mid s) - \nabla_A \log p_t(x_{AB,t} \mid s)\big]. \end{aligned} \tag{6}$$

The second bracket is the vector inside the locality gap, conditioned on the semantic label s. In consistency with common-cause hypothesis 1, we expect an informative label to reduce this gap by specifying shared aspects of the image's global organization, leaving less dependence on distant context:

$$\mathbb{E}\Big[\|\nabla_A \log p_t(x_t \mid s) - \nabla_A \log p_t(x_{AB,t} \mid s)\|_2^2\Big] < \mathbb{E}\Big[\|\nabla_A \log p_t(x_t) - \nabla_A \log p_t(x_{AB,t})\|_2^2\Big],$$

with expectations taken over noisy images and their semantic labels.

Under this score-gap suppression assumption, a small unconditional gap implies that both brackets on the right-hand side of equation 6 are small in squared expectation. Conversely, when the unconditional gap is nonzero, the conditional gap cannot account for it entirely, so the first bracket must also be nonzero in squared expectation.

To interpret the first bracket, let s be the label of the clean image. The quantities $p_t(s \mid x_t)$ and $p_t(s \mid x_{AB,t})$ are the posterior probabilities assigned to that label using global and local observations, respectively. Their log-gradients measure how these probabilities respond to perturbations in A. At low noise, both observations can identify the label reliably, and we expect their semantic posteriors to be relatively insensitive to small perturbations. At high noise, both observations become uninformative about the label, and their posteriors approach the prior. The difference in posterior responses is therefore expected to be most pronounced between the loss of reliable local label information and the loss of global label information. This motivates the conjecture that the nonlocality window lies within the speciation window. Section 3.2 makes this connection precise using mutual information and an explicit common-cause assumption.

3.2 THE INFORMATION-THEORETIC DEFINITION OF WINDOWS

3.2.1 NONLOCALITY WINDOW

As shown by Hu et al. (2025) (and reproduced in our notation in App. A), the conditional mutual information $I_t(A; C \mid B)$ bounds the error of local recovery. Therefore we define the breakdown window of local denoising as the window where CMI is above a threshold.

Definition 1 (Nonlocality window)**.** *For an annulus tripartition ABC, and an information tolerance $\delta > 0$, define the start and end of the nonlocality window as*

$$t_{\text{nonloc}}^{\text{start},\delta} := \inf\{t \in [0,1] : I_t(A; C \mid B) > \delta\}, \tag{7}$$

$$t_{\text{nonloc}}^{\text{end},\delta} := \sup\{t \in [0,1] : I_t(A; C \mid B) > \delta\}. \tag{8}$$

3.2.2 SPECIATION WINDOW

We show that the mutual information $I_t(S; R)$ has the same window where it sharply decreases from the maximal $H(S)$ to the minimal value 0, using the following lemma proved in Appendix C.

Lemma 1 (Two-sided sandwich bounds)**.** *Let $p(s)$ be the prior distribution of labels $s \in S$, $|S| = L$, and write the posterior as $q_s \equiv p_t(s \mid x_{R,t})$. Let $P_{\mathrm{FB}}(R,t)$ denote the success probability of a forward-backward experiment, as described in equation 5. On one hand, we have*

$$\frac{1}{2}\mathbb{E}\sum_s (q_s - p(s))^2 \le I_t(S;R) \le \sum_s \mathbb{E}\frac{(q_s - p(s))^2}{p(s)}. \tag{9}$$

On the other hand,

$$\log P_{\mathrm{FB}}(R,t) \ge I_t(S;R) - H(S) \ge -h_2\left(P_{\mathrm{FB}}(R,t)\right) - \left(1 - P_{\mathrm{FB}}(R,t)\right)\log(L-1), \tag{10}$$

where $h_2(p) = -p\log p - (1-p)\log(1-p)$ is the binary entropy.

The first inequality 9 applies to the zero–information endpoint: $I_t(S;R) = 0$ if and only if $q_s - p(s) = 0$ for all s and $x_{R,t}$, reducing the posterior-sampling FB experiment to prior sampling. The second inequality 10 applies to the full-information endpoint, i.e., $I_t(S;R) = H(S)$ if and only if we succeed at $P_{\mathrm{FB}}(R,t) = 1$. Therefore, $P_{\mathrm{FB}}(R,t)$ and $I_t(S;R)$ share the same window where they both drop from their respective maxima to minima.

The mutual information for a global region $I_t(S;ABC)$ must be greater than that for a local region $I_t(S;B)$ by data processing inequality, indicating that $I_t(S;B)$ has an earlier drop than $I_t(S;ABC)$. Therefore, we choose our definition for speciation window taking this locality nuance into consideration. As the score function arguments in Section 3.1 suggest, we define the speciation window starting with nonzero local response to A and ending with global response to A.

Definition 2 (Speciation window)**.** *For an annulus tripartition ABC, and an information tolerance $\delta > 0$, define the start and end of the speciation window as*

$$t_{\mathrm{spec}}^{\mathrm{start},\delta} := \inf\{t \in [0,1] : I_0(S;AB) - I_t(S;B) \ge \delta\}, \tag{11}$$

$$t_{\mathrm{spec}}^{\mathrm{end},\delta} := \sup\{t \in [0,1] : I_t(S;ABC) \ge \delta\}. \tag{12}$$

The start time indicates the earliest time when local label recognition ability drops by a certain amount, and the end time indicates the latest time when global recognition ability retains that amount.

3.2.3 Common-cause hypothesis

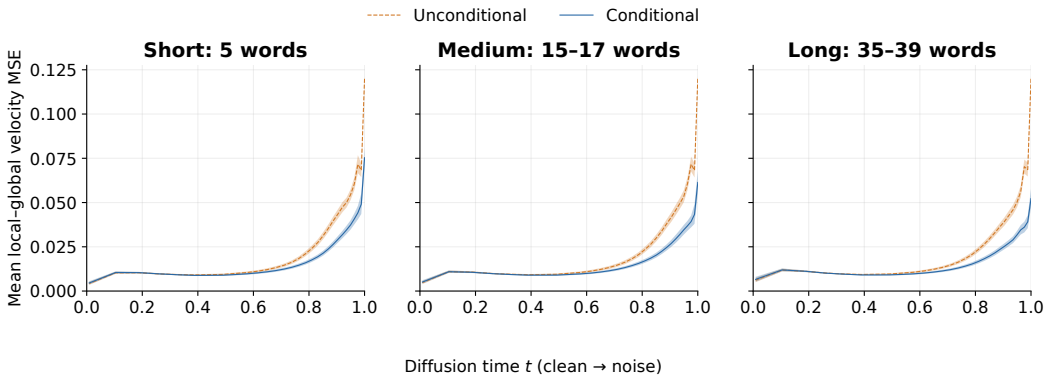


Figure 2: Semantic conditioning reduces the average local–global prediction gap in SD3. A smaller conditional gap means that restricting distant image-token interactions changes the model prediction less when semantic information is supplied. This is consistent with the common-cause interpretation: the description explains part of the shared image structure that would otherwise require distant context.

Conditioning on semantic labels can act in two ways. One is synergistic, where it increases conditional mutual information,

$$I(A;C \mid B,S) \ge I(A;C \mid B).$$

This occurs when label S contains information that can only be determined by A and C together, but not from any of them alone. For example, count of total objects or parity of a set of bits. The other is redundant, where it decreases conditional mutual information,

$$I(A;C \mid B,S) \leq I(A;C \mid B).$$

This occurs when label S acts as a common cause for A and C, such as the label of a cat image, which explains the correlation between the cat's head and tail. We assume that the latter is the case for natural datasets, and we postulate the following *common-cause hypothesis*.

Hypothesis 1 (Common-cause hypothesis)**.** *For an annulus tripartition ABC, semantic labels S, and a positive constant $0 < \alpha \leq 1$, we have*

$$I_t(A;C \mid B,S) \leq (1-\alpha) I_t(A;C \mid B) \tag{13}$$

for all time $t \in [0,1]$.

Such labels always exist: consider the extreme case where the label S specifies the entire clean image, then $I(A;C \mid B,S) = 0$, in which case we have $\alpha = 1$. In general, we expect that the more informative the label is, the larger α is.

Gaussian mixtures with identity covariance satisfy this hypothesis for the labels matching the mixture labels at $\alpha = 1$ as shown in Appendix E. For natural datasets, Figure 2 provides empirical evidence consistent with the common-cause interpretation: semantic conditioning reduces the average local–global prediction gap in SD3, suggesting that shared semantic information reduces reliance on distant image context. We average over 64 scenes described using 5, 15–17, and 35–39 words, respectively, giving 192 prompts in total and each prompt uses three random seeds. We observe a time-weighted gap reduction of 20.7%, 23.4%, and 24.4% for short, medium, and long descriptions (see Appendix D for the experimental setup, score parameterization, and limitations).

3.2.4 Nonlocality window is contained in speciation window

The two windows are defined using different information quantities. The common-cause hypothesis connects them.

Theorem 1 (Containment of the nonlocality window)**.** *Fix an annulus tripartition ABC. Suppose that, for all $t \in [0,1]$, $S \to X_{AB,0} \to X_{AB,t}$ is a Markov chain and Hypothesis 1 holds with the same $\alpha > 0$. Choose $0 < \delta < \sup_{t\in[0,1]} I_t(A;C \mid B)$. Then*

$$t_{\text{spec}}^{\text{start},\alpha\delta} \leq t_{\text{nonloc}}^{\text{start},\delta} \leq t_{\text{nonloc}}^{\text{end},\delta} \leq t_{\text{spec}}^{\text{end},\alpha\delta}. \tag{14}$$

Proof. The common-cause hypothesis gives, at each time,

$$\alpha I_t(A;C \mid B) \leq I_t(A;C \mid B) - I_t(A;C \mid B,S). \tag{15}$$

By the chain rule, the difference on the right has two useful forms. The first is

$$\begin{aligned} I_t(A;C \mid B) - I_t(A;C \mid B,S) &= I_t(S;AB) - I_t(S;B) - I_t(S;ABC) + I_t(S;BC) \\ &\leq I_0(S;AB) - I_t(S;B). \end{aligned} \tag{16}$$

The last step uses $I_t(S;AB) \leq I_0(S;AB)$, by data processing along the forward noising channel, and $I_t(S;ABC) \geq I_t(S;BC)$, by spatial data processing. Rearranging the same four mutual informations gives the second form,

$$\begin{aligned} I_t(A;C \mid B) - I_t(A;C \mid B,S) &= I_t(S;BC) - I_t(S;B) - I_t(S;ABC) + I_t(S;AB) \\ &\leq I_t(S;ABC). \end{aligned} \tag{17}$$

Here $I_t(S;B) \geq 0$ and $I_t(S;ABC) - I_t(S;AB) \geq 0$, while $I_t(S;BC) \leq I_t(S;ABC)$.

Before $t_{\text{spec}}^{\text{start},\alpha\delta}$, the definition of the speciation start gives $I_0(S;AB) - I_t(S;B) < \alpha\delta$. Equations 15 and equation 16 then give $I_t(A;C \mid B) < \delta$. After $t_{\text{spec}}^{\text{end},\alpha\delta}$, the definition of the speciation end gives $I_t(S;ABC) < \alpha\delta$; equations 15 and 17 again give $I_t(A;C \mid B) < \delta$. Thus every time at which CMI exceeds δ lies between the speciation endpoints. Since δ is below the CMI supremum, this set is nonempty. Taking its infimum and supremum proves the claim. □

3.3 System size scaling of two windows and phase transitions

The containment theorem 1 relates the two windows, but does not determine how their widths scale with system size. This scaling describes whether a finite-size crossover sharpens into a phase transition. To get a sense of how the windows scale with system size in simple distributions, we explicitly study the windows for a two-component Gaussian mixture. Relegating the calculational details to E, we only quote the analytical results. We study the distribution $p(x, s)$ where $p(x \mid s) \propto \exp(-(x-\mu_s)^2/2\sigma^2)$ and $p(s = \pm 1) = 1/2$, taking $\mu_s = s\vec{1} \in \mathbb{R}^d$. The scalings for the two windows for this distribution are

$$1 - t_{\text{spec}}^{\text{start},\alpha\delta} \asymp \frac{\sqrt{\log N}}{N}\,,\; 1 - t_{\text{nonloc}}^{\text{start},\delta} \asymp \frac{1}{N}\,,\; 1 - t_{\text{nonloc}}^{\text{end},\delta} \asymp \frac{1}{N}\,,\; 1 - t_{\text{spec}}^{\text{end},\alpha\delta} \asymp \frac{1}{N^2}, \tag{18}$$

which all converge to $t = 1$ in the large-N limit. The intuition behind these scalings follow from the simple fact that decoding success is determined by the signal-to-noise ratio $\|\mu_s(t)\|/\sigma(t)$. For μ_s we chose, the signal in B or BC is extensive $\|\mu_s\| \propto N(1-t)$, and thus decoding becomes ambiguous at $1 - t = O(1/N)$. Going left to right (choosing $\delta \sim 1/N^2$ to get a nontrivial CMI) the start of the speciation window is marked by when the classifier using B loses δ amount of information, whose error is given by a Gaussian tail; setting this equal to the loss yields a $\sqrt{\log 1/\delta}$ in addition to the signal to noise factor $1/N$. The locality windows are about a similar but a different question: when is decoding hard with B only but easy with BC? Since the signal is extensive in each, both locality windows directly inherit $1 - t = O(1/N)$. Lastly, in the end of the speciation window, we are in the weak signal regime, and even the global classifier is weak. From a quadratic approximation to the mutual information $I(S : ABC)$ integral as $N^2(1-t)^2$, which needs to be equal to δ, we get $1 - t = O(1/N^2)$. In summary, the mechanism behind this sharpening can be understood through the decoding transition of the Gaussian mixture model, which behaves like a soft repetition code, where each local patch carries partial information about the semantic label, and a growing buffer combines these clues.

Inspired by the Gaussian mixture results, we give sufficient conditions for both windows to sharpen to a common critical point at pure noise, $t = 1$, as system size grows. The theorem below makes this condition for local semantic identification precise and generalizes the calculation restricted to the Gaussian mixture model.

Theorem 2. *(informal) Let $m_s = \mathbb{E}[X_{B,0} \mid S = s]$ be the mean image on the buffer for label s, and write*

$$D_N = \min_{s \neq s'} \|m_s - m_{s'}\|^2.$$

Suppose within-class fluctuations have Gaussian-type tails with variance scale at most C_N along every unit direction. Thus D_N measures the separation of the class means, while C_N measures the noise that can obscure this separation. For a fixed finite label set, assume the common-cause hypothesis holds with size-independent $\alpha > 0$ and use $X_t = (1-t)X_0 + tZ$. If $D_N/(C_N+1) \to \infty$, both windows shrink toward $t = 1$, provided they exist and their defining information thresholds do not decrease too rapidly with system size relative to this growing separation-to-noise ratio. The precise regularity condition on the thresholds and the full assumptions are given in Appendix F.

When $D_N/(C_N + 1)$ grows, the semantic signal in the buffer increasingly dominates the within-class fluctuations and the added diffusion noise. As in a repetition code, the label can then be read from the buffer with vanishing error at any fixed $t < 1$; only at pure noise is the signal completely lost.

Reliable decoding makes the remaining label uncertainty $H(S \mid X_{B,t})$ small. The common-cause hypothesis and the chain rule then give

$$\alpha I_t(A; C \mid B) \leq I_t(S; A \mid B) - I_t(S; A \mid B, C) \leq H(S \mid X_{B,t}).$$

Thus, when the buffer already identifies the label, the CMI is small. Moreover, $I_0(S; AB) - I_t(S; B) \leq H(S \mid X_{B,t})$, so the speciation window cannot start until the buffer begins to lose the label. The quantitative decoding bound pushes this start toward $t = 1$ at the allowed thresholds, and Theorem 1 places the other three endpoints between that start and 1.

4 CONCLUSION

We connect semantic speciation and nonlocality through the information shared between distant regions of a sample. Under the common-cause hypothesis, we prove that the nonlocality window lies within the speciation window. This connection interprets semantics as shared global information and speciation as its decoding during generation. We further give a sufficient condition for both windows to sharpen to a common critical point at pure noise as system size grows, which we instantiate and analysze in gaussian mixtures.

These results open new questions about how semantic information shapes nonlocality and phase transitions in generative models. The common-cause and semantic-decoding assumptions can be tested in graphical models and real datasets. These studies can reveal how spatial structure affects when the transitions occur and how sharply they develop. The theory can also guide when denoisers use distant context and how their receptive fields change during generation. We leave these directions to future work.

AI USE STATEMENT

We used ChatGPT 6 Pro, GPT 6 Astra, and GPT 5.6 Sol to formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, and implement the software code for common-cause experiment D and local forward–backward experiment B.

Additionally, we used generative AI tools to create or modify scientific figures or images, suggest experimental parameters, creation of artifacts, draft parts of a research paper, summarize or analyse existing literature, brainstorming, sourcing/searching for information, and identify relevant literature.

We have not used generative AI tools to generate synthetic data sets, clean and reformat dataset, support qualitative and thematic data analysis, interpret results, and help develop theoretical models or conceptual frameworks.

We have reviewed all AI-assisted work. AI-generated proofs are manually checked and verified by 3 authors. AI-generated codes are reviewed and tested by 1 author. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide codes in supplementary materials for reproduction of common-cause experiment D and local forward–backward experiment B.

REFERENCES

Luca Ambrogioni. How out-of-equilibrium phase transitions can seed pattern formation in trained diffusion models, 2026. URL https://arxiv.org/abs/2603.20092. 2

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. 1, 4

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. URL https://web.stanford.edu/~boyd/cvxbook/. 17

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)*, pp. 11462–11471, 2022. doi: 10.1109/CVPR52688.2022.01118. 1

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009. 4

Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021. URL https://arxiv.org/abs/2105.05233. 18

Rick Durrett. *Probability: Theory and Examples*. Cambridge University Press, 5 edition, 2019. URL https://sites.math.duke.edu/~rtd/PTE/PTE5_011119.pdf. Linked author's draft dated January 11, 2019. 15, 16

Alex Dytso, H. Vincent Poor, and Shlomo Shamai (Shitz). A general derivative identity for the conditional mean estimator in Gaussian noise and some applications. arXiv preprint arXiv:2104.01883, 2021. URL https://arxiv.org/abs/2104.01883. 15

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. 1

Omar Fawzi and Renato Renner. Quantum conditional mutual information and approximate markov chains. *Communications in Mathematical Physics*, 340(2):575–611, September 2015. ISSN 1432-0916. doi: 10.1007/s00220-015-2466-x. URL http://dx.doi.org/10.1007/s00220-015-2466-x. 13

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org. 14

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. 2

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. 2

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. URL https://arxiv.org/abs/2207.12598. 18

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. 14

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. 1, 2, 3, 4, 5, 12, 13

Henry Hunt, Mason Kamb, and Surya Ganguli. An exact information theory of generalization phase transitions in bayesian diffusion models. *arXiv preprint arXiv:2607.08041*, 2026. 1, 2

Mason Kamb and Surya Ganguli. An analytic theory of creativity in convolutional diffusion models. *arXiv preprint arXiv:2412.20292*, 2024. 1, 2

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. 1

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. 2

Artem Lukoianov, Chenyang Yuan, Justin Solomon, and Vincent Sitzmann. Locality in image diffusion models emerges from data statistics. *arXiv preprint arXiv:2509.09672*, 2025. 1

David J. C. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, Cambridge, UK, 2003. ISBN 9780521642989. 13

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. 1

Matthew Niedoba, Berend Zwartsenberg, Kevin Murphy, and Frank Wood. Towards a mechanistic explanation of diffusion model generalization. *arXiv preprint arXiv:2411.19339*, 2024. 1, 2

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023. 1

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. 1

Ben Poole, Sherjil Ozair, Aaron van den Oord, Alexander A. Alemi, and George Tucker. On variational bounds of mutual information, 2019. URL https://arxiv.org/abs/1905.06922. 4

Akhil Premkumar. On the separability of information in diffusion models, 2026. URL https://arxiv.org/abs/2509.23937. 4

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. 1

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115 (3):211–252, 2015. doi: 10.1007/s11263-015-0816-y. 4

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. 3

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, 2025. doi: 10.1073/pnas.2408799121. URL https://www.pnas.org/doi/abs/10.1073/pnas.2408799121. 1, 4, 18

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. 1

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html. 14

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. 14

Tomoei Takahashi, Takashi Takahashi, and Yoshiyuki Kabashima. Dynamical regimes of discrete diffusion models, 2026. URL https://arxiv.org/abs/2604.10961. 1

Andre Wibisono and Varun Jog. Convexity of mutual information along the heat flow. arXiv preprint arXiv:1801.06968, 2018. URL https://arxiv.org/abs/1801.06968. 16

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. ConvNeXt V2: Co-designing and scaling ConvNets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16133–16142, 2023. 18

Longxuan Yu, Xing Shi, Xianghao Kong, Tong Jia, and Greg Ver Steeg. Mmg: Mutual information estimation via the mmse gap in diffusion, 2025. URL https://arxiv.org/abs/2509.20609. 12

Yifan Zhang and Sarang Gopalakrishnan. Conditional mutual information and information-theoretic phases of decohered gibbs states, 2025. URL https://arxiv.org/abs/2502.13210. 4

Yifan F. Zhang, Fangjun Hu, Guangkuo Liu, Mert Okyay, and Xun Gao. Concurrence of symmetry breaking and nonlocality phase transitions in diffusion models, 2026. URL https://arxiv.org/abs/2605.04830. 1, 3, 4, 12

Wei Zhu, Jian Zhang, Lixing Yu, Kun Yue, and Zhiwen Tang. Dissecting failure dynamics in large language model reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8893–8914. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.401. URL https://aclanthology.org/2026.acl-long.401/. 2

A Local Denoising Models and Error Bounds

In this appendix, we review and advance various aspects of local diffusion models. First, we review a sufficient condition for locality of a denoiser in terms of the conditional mutual information, proven already in Hu et al. (2025), for the reader's convenience. Second, specialising to Gaussian noise relevant for diffusion models, we prove that the same quantity also bounds the locality gap—a quantity introduced in Zhang et al. (2026) to probe the locality transition. This directly connects the observations therein to the CMI, and supports the notion that locality gap is sensitive to the locality transition. Noting that mutual informations are very hard to sample, training a diffusion model and measuring the locality gap provides a way of controlling CMI, which follows the same spirit as Yu et al. (2025).

A.1 Denoising Generative Models

Denoising generative models learn to recover clean data from corrupted observations. Let $p, q : \Omega \to \mathbb{R}$ be probability densities where Ω is the space of events. Let $\mathcal{N}(y \mid x)$ be a noise channel. It induces a noisy version of a data density p as

$$\mathcal{N}(p)(y) = \int_\Omega \mathrm{d}x\, \mathcal{N}(y \mid x) p(x)\,. \tag{19}$$

$\mathcal{B}_{\mathcal{N},q}(y \mid x)$ is its Bayes recovery channel with prior q, defined as

$$\mathcal{B}_{\mathcal{N},q}(y \mid x) = \frac{\mathcal{N}(y \mid x) q(x)}{\mathcal{N}(q)(y)}\,. \tag{20}$$

Generation begins from a tractable, highly corrupted distribution and successively applies learned approximations to such reverse channels. This framework includes autoregressive models, for which the forward channel progressively masks a suffix of the variables and the reverse process reveals X_i according to $p(X_i \mid X_{<i})$, as well as diffusion models, for which the corruption is gradual, typically Gaussian, and the reverse dynamics are parameterized through a denoiser or score function. The results below are first stated for a generic noise channel and then specialized to diffusion models.

A.2 Bounding Local Recovery Error with CMI (Generic Denoising Model)

In this section, we prove that CMI bounds the squared total variance (TV) between a distribution and its noised and then locally denoised version. We will not try be rigorous in the measure-theoretic sense, but the statements can be formalised as such if desired.

The nontrivial ingredient in proving the desired result is the classical Fawzi-Renner inequality from Fawzi & Renner (2015), which we now state.

Proposition 1 (Classical Fawzi-Renner Inequality)**.** *Let $\hat{p} : \Omega \to \mathbb{R}$ be the distribution denoised with prior q after applying the noise channel, i.e.,*

$$\hat{p}(x) \coloneqq [\mathcal{B}_{\mathcal{N},q}(\mathcal{N}(p))](x)\,.$$

then,

$$D_{\mathrm{KL}}(p\|q) - D_{\mathrm{KL}}(\mathcal{N}(p)\|\mathcal{N}(q)) \geq D_{\mathrm{KL}}(p\|\hat{p})\,, \tag{21}$$

where D_{KL} is the Kullback-Leibler divergence/relative entropy.

Now suppose our samples x can be partitioned into three sets, A, B, C with respective r.v.s x_A, x_B, x_C independent of any spatial geometry. Further suppose the noise channel acts only on A, i.e., $\mathcal{N}(p)(y) = \sum_{x_A \in \Omega_A} \mathcal{N}(y \mid x)p(x)$ where, in particular, $y_B = x_B$ and $y_C = x_C$. We also denote the variables as X and the noisy variables as $\tilde{X}$. Then, taking $p = p(x)$ and $q(x) = p(x_A, x_B)p(x_C)$, by definition of the mutual information $I(X;Y)$ between two r.v.s X and Y as the relative entropy between the joint and the product distribution, we have

$$D_{\mathrm{KL}}(p\|q) = I(AB;C)\,, \quad D_{\mathrm{KL}}(\mathcal{N}(p)\|\mathcal{N}(q)) = I(\tilde{A}B;C)\,. \tag{22}$$

The CMI is defined in terms of MI's, for our application, we write (for Z a generic r.v.)

$$I(ZB;C) = I(Z;C \mid B) + I(B;C)\,. \tag{23}$$

The term $I(B;C)$ cancels between the two MI's, giving

$$D_{\mathrm{KL}}(p\|q) - D_{\mathrm{KL}}(\mathcal{N}(p)\|\mathcal{N}(q)) = I(A;C \mid B) - I(\tilde{A};C \mid B) \leq I(A;C \mid B)\,,$$

by positivity of the CMI $I(\tilde{A};C \mid B) \geq 0$. Thus, we have obtained

$$I(A;C \mid B) \geq D_{\mathrm{KL}}(p\|\hat{p})\,. \tag{24}$$

Lastly, by Pinsker's inequality MacKay (2003), we have $D_{\mathrm{KL}}(p\|\hat{p}) \geq 2\,\mathrm{TV}(p\|\hat{p})^2$ where $\mathrm{TV}(p\|q) = \int_\Omega |p(x) - \hat{p}(x)|/2$. Chaining the inequalities together, we obtain the following theorem.

Theorem 3 (CMI bounds local reconstruction error Hu et al. (2025).)**.** *For a distribution p, a noise channel $\mathcal{N}$ acting only on A, with an ABC tripartition and with $\hat{p}(x) = [\mathcal{B}_{\mathcal{N},p(x_A,x_B)p(x_C)}(\mathcal{N}(p))](x)$ we have*

$$2\,\mathrm{TV}(p\|\hat{p})^2 \leq I(A;C \mid B)\,, \tag{25}$$

i.e., CMI of the distribution before the noise is added controls the locality of the denoiser.

The last thing to clear up is to show that the denoiser in Theorem 3 is truly local. This follows by definition, where $q = p(x_A, x_B)p(x_C)$ and

$$\mathcal{B}_{\mathcal{N},q}(\mathcal{N}(p))(x) = \frac{\mathcal{N}(y_A \mid x_A)p(x_A x_B)p(x_C)}{\int \mathrm{d}x_A \mathcal{N}(y_A \mid x_A)p(x_A x_B)p(x_C)} = \frac{\mathcal{N}(y_A \mid x_A)p(x_A x_B)}{\int \mathrm{d}x_A \mathcal{N}(y_A \mid x_A)p(x_A x_B)}\,. \tag{26}$$

Suppose we break down a single round of diffusion noise pixel-by-pixel, and only consider a single step of this round. Taking A to be the pixel to whom the noise is added, the CMI bounds the total variance (by Pinsker's inequality and Fawzi-Renner, as applied in Hu et al., 2025, Section III.2 and Supplementary Material S2.A

$$\mathrm{TV}(p\|\hat{p})^2 \leq I(A;C \mid B)\,, \tag{27}$$

where $\mathrm{TV}(p\|q)$ is the total variation $\sum_{x\in\Omega}|p(x) - q(x)|/2$ and $\hat{p}(x) \coloneqq (\mathcal{B}_{\mathcal{N},p(x_A,x_B)}(\mathcal{N}_A(p))(x)$ is the denoised distribution after adding noise to P only on A via the noise channel $\mathcal{N}(y_A \mid x_A)$. Lastly, $\mathcal{B}_{\mathcal{N},\mathcal{N}_A}$ is the "Bayes recovery channel" given by Bayes' rule $\mathcal{B}_{\mathcal{N},q}(x \mid y) = \mathcal{N}(y \mid x)q(x)/\mathcal{N}(q)(y)$ for any probability distribution $q(x)$. If the CMI is zero, the total variance between the initial distribution p and the noised–locally-denoised $\hat{p}$ is zero—the local denoiser is exact. If instead, the CMI is not small but decays exponentially in the radius of B, i.e. if we have

$$I(A;C \mid B) \leq \gamma \exp(-r/\xi)\,, \tag{28}$$

where ξ is known as the *Markov length* (Hu et al., 2025, Section III.1), we can pick a r small enough to guarantee a total variation error of ϵ, i.e., demanding $\epsilon^2 \geq \gamma \exp(-r/\xi)$, we get

$$r \gtrsim \xi \log \frac{\sqrt{\gamma}}{\epsilon}\,. \tag{29}$$

Then, stitching together many single-step diffusions, we may derive a similar bound on the full diffusion path (Hu et al., 2025, Section III.3) where each individual step size must at least be $r \geq \xi \log(NK\sqrt{\gamma}/\epsilon)$, where K is the window size and N is the number of discretization steps. As such, a decaying CMI *guarantees a local denoiser*.

A.3 DIFFUSION MODELS

Diffusion models, as a special type of denoising generative models, propose a parametrisation of a data distribution based on the Langevin SDE Sohl-Dickstein et al. (2015); Ho et al. (2020); Song et al. (2021)

$$\mathrm{d}X_t = \mu(X_t, t)\mathrm{d}t + \sigma(t)\mathrm{d}\eta\,, \tag{30}$$

where η is a Gaussian random variable, drawn independently at each time step. Under suitable assumptions, there exists a corresponding Fokker-Planck equation (FPE)

$$\partial_t P = -\partial_x(\mu(x,t)P) + \partial_x(\sigma^2(t)\partial_x P)/2\,. \tag{31}$$

The FPE admits an exact time-reversed form Song et al. (2021). Let $Q(x,t) = P(x, T-t)$ with T the reversal time

$$\partial_t Q = -\partial_x(\tilde{\mu}(x, T-t)Q) + \partial_x(\sigma^2(T-t)\partial_x Q)/2\,, \tag{32}$$

with $\tilde{\mu}(x,t) = -\mu(x,t) + \sigma^2(t)\partial_x \log Q$. This implies the reversed SDE has a modified drift shifted by $-\sigma^2(t)\partial_x \log Q$, which we need to know to be able to reverse each trajectory. The hard part is the so-called *score* function $\partial_x \log Q$, which requires the knowledge of the full distribution $P(x,t)$. The application to generative modeling takes $P(x,0) = p_{\mathrm{data}}$, such that $P(x, t\to\infty) = \mathcal{N}(\mu(x), \sigma^2(\infty))$. Since sampling the late-time Gaussian is easy, the difficulty of sampling p_{data} is purely relegated to the reversal process. Interestingly, this reveals that we do not need all of p_{data}, we only need the gradient of its log, the score function. Therefore, we do not care about multiplicative constants in p_{data}, which are intractable to compute in many cases. To see this, take an energy based model $p_\theta(x) = \exp(-E_\theta(x))/Z_\theta$ (Goodfellow et al., 2016, Section 18.4). The score is

$$s_\theta(x) = -\partial_x \log Z_\theta - \partial_x E_\theta(x)\,, \tag{33}$$

and the first term is zero $\log Z_\theta$ does not depend on x, and as such, is zero.

A.4 BOUNDING LOCAL DENOISER ERROR WITH CMI (DIFFUSION MODEL)

While we have provided an information-theoretic guarantee on when a local score may be constructed, we have not operationalized what it means to have a local score function and how to interpret deviations from locality in terms of optimal denoising.

Define the expectation of a function $f(x)$ over the noisy data distribution be $p_t(x)$. We now will argue, in expectation, that the score function of the marginal distribution $p(x_{AB})$ (obtained by marginalizing over x_C) is the smallest over all t possible local approximations. Suppose we approximate the score $\nabla_A \log p_t(x)$ by a function $f(x_{AB})$. The error in the approximation satisfies

$$\begin{aligned}\mathbb{E}_{p_t}\|\nabla_A \log p_t(x) - f(x_{AB})\|^2 = \mathbb{E}_{p_t}\|\nabla_A \log p_t(x) - \nabla_A \log p_t(x_{AB})\|^2 \\ + \mathbb{E}_{p_t}\|\nabla_A \log p_t(x_{AB}) - f(x_{AB})\|^2\,,\end{aligned} \tag{34}$$

and as such, the best local approximation is the score of the AB marginal. Any other approximation incurs an extra cost $\|\nabla_A \log p_t(x_{AB}) - f(x_{AB})\|^2$ in expectation.

The proof follows directly, by adding and subtracting the local score marginal $\nabla_A \log p_t(x_{AB})$, which yields

$$\begin{aligned}&\mathbb{E}_{p_t(x)}\|\nabla_A \log p_t(x) - f(x_{AB})\|^2 = \\ &\quad \mathbb{E}_{p_t(x)}\|\nabla_A \log p_t(x) - \nabla_A \log p_t(x_{AB})\|^2 + \mathbb{E}_{p_t(x)}\|\nabla_A \log p_t(x_{AB}) - f(x_{AB})\|^2 \\ &\quad + 2\mathbb{E}_{p_t(x)}(\nabla_A \log p_t(x) - \nabla_A \log p_t(x_{AB})) \cdot (\nabla_A \log p_t(x_{AB}) - f(x_{AB}))\,.\end{aligned}$$

Noting the expectation is over $p_t(x)$, we have

$$\mathbb{E}_{p_t(x)}(\nabla_A \log p_t(x) - \nabla_A \log p_t(x_{AB})) \cdot (\nabla_A \log p_t(x_{AB}) - f(x_{AB})) \tag{35}$$

By definition, $\int \mathrm{d}x_C p_t(x_C \mid x_{AB}) \nabla_A \log p_t(x) = \nabla_A \log p_t(x_{AB})$, and the cross term is zero.

Thus, the error in any local approximation is the *locality* gap

$$\Delta_{\mathrm{loc}}(t)^2 = \mathbb{E}_{p_t(x)}\|\nabla_A \log p_t(x) - \nabla_A \log p_t(x_{AB})\|^2\,, \tag{36}$$

which is a fundamental probe of the locality of the true score function: it is zero, if and only if, the true score is local. While the locality gap is a so-called *instantenous* probe—it probes the score at a single time instance and does not determine directly the output 'quality'—it nevertheless describes the downsteam error in the generated image: using Tweedie's identity applied to the AB marginal, one can show

$$\mathbb{E}[x_A \mid x(t)] - \mathbb{E}[x_A \mid x_{AB}(t)] = \frac{t^2}{1-t}(\nabla_A \log p_t(x) - \nabla_A \log p_t(x_{AB}))\,, \tag{37}$$

i.e., the vector difference in the locality gap is the change in the optimal prediction x_A due to revealing C, and also

$$\mathbb{E}[|x_A - \mathbb{E}[x_A \mid x_{AB}(t)]|^2] - \mathbb{E}[|x_A - \mathbb{E}[x_A \mid x(t)]|^2] = \frac{t^4}{(1-t)^2}\Delta_{\mathrm{loc}}^2\,, \tag{38}$$

i.e., the locality gap is a direct probe of the extra denoising error on A incurred by hiding C in the optimal denoiser. In this sense, the norms of the instantenous probes control pixelwise differences/errors in the denoised output.

Furthermore, just like the total variation between the data distribution P and its noised-then-locally-denoised version $\hat{P}$, we can also show that CMI controls the size of the locality gap. In particular, it is possible to show

Proposition 2 (Controlling the locality gap by CMI)**.** *Assume*

$$X(t) = \alpha(t)X(0) + \sigma(t)Z, \qquad \sigma(t) > 0,$$

where Z is an independent standard Gaussian and $X(0)$ has finite second moment. Write $m = \dim X_A$, and define the locality gap by

$$\Delta_{\mathrm{loc}}(t) := \left(\mathbb{E}_{p_t}\, \|\nabla_A \log p_t(x(t)) - \nabla_A \log p_t(x_{AB}(t))\|^2\right)^{1/2},$$

where $\|\cdot\|$ is the Euclidean norm on the A coordinates. Then, with natural logarithms,

$$\Delta_{\mathrm{loc}}(t)^2 \le \frac{2\sqrt{m(m+3)}}{\sigma(t)^2}\sqrt{I_t(A; C \mid B)}.$$

Proof. Fix t. Tweedie's formula and the Hatsell–Nolte identity (Dytso et al., 2021, Eq. (3) and Proposition 1) give

$$\nabla_A^2 \log p_t(x(t)) = -\frac{I_m}{\sigma(t)^2} + \frac{\mathrm{Cov}(\alpha(t)X_A(0) \mid X(t))}{\sigma(t)^4} = \frac{\mathrm{Cov}(Z_A \mid X(t)) - I_m}{\sigma(t)^2}. \tag{39}$$

The second equality uses $\alpha(t)X_A(0) = X_A(t) - \sigma(t)Z_A$ at fixed $X(t)$.

We use two standard moment identities. Here $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm, whose square is the sum of squared matrix entrie. First, the *L^2-contraction of conditional expectation* states that, for any square-integrable scalar random variable V and side information S,

$$\mathbb{E}\Big[\big(\mathbb{E}[V \mid S]\big)^2\Big] \le \mathbb{E}[V^2].$$

This follows from conditional Jensen's inequality and the tower property (Durrett, 2019, Theorem 4.1.11).

By Wick's theorem for Gaussian moments we have,

$$\mathbb{E}\|Z_A\|^4 = \sum_{i,j=1}^{m} \mathbb{E}[Z_{A,i}^2 Z_{A,j}^2] = \sum_{i,j=1}^{m} (1 + 2\delta_{ij}) = m(m+2).$$

Since a covariance matrix is positive semidefinite,

$$\|\,\mathrm{Cov}(Z_A \mid X(t))\|_{\mathrm{F}} \le \mathrm{tr}\,\mathrm{Cov}(Z_A \mid X(t)) \le \mathbb{E}[\|Z_A\|^2 \mid X(t)].$$

Thus, expanding equation 39, dropping the nonpositive trace term, and applying the two moment identities above gives

$$\begin{aligned}\mathbb{E}_{p_t}\|\nabla_A^2 \log p_t(x(t))\|_{\mathrm{F}}^2 &= \frac{m - 2\mathbb{E}\,\mathrm{tr}\,\mathrm{Cov}(Z_A \mid X(t)) + \mathbb{E}\|\,\mathrm{Cov}(Z_A \mid X(t))\|_{\mathrm{F}}^2}{\sigma(t)^4} \\ &\le \frac{m + \mathbb{E}\Big[\big(\mathbb{E}[\|Z_A\|^2 \mid X(t)]\big)^2\Big]}{\sigma(t)^4} \\ &\le \frac{m + \mathbb{E}\|Z_A\|^4}{\sigma(t)^4} = \frac{m(m+3)}{\sigma(t)^4}.\end{aligned} \tag{40}$$

Now add a fictitious Gaussian noise of variance $u \ge 0$ to A alone. The resulting A variable has the same distribution as

$$\alpha(t)X_A(0) + \sqrt{\sigma(t)^2 + u}\, Z_A,$$

while $X_B(t)$ and $X_C(t)$ remain unchanged. The Gaussian noise Z_A is still independent of $(X(0), X_B(t), X_C(t))$. Therefore equation 40 remains valid with $\sigma(t)^2$ replaced by $\sigma(t)^2 + u$.

The *de Bruijn identity* states that, for a Gaussian-smoothed density $\rho_u(y)$ evolving according to

$$\partial_u \rho_u = \frac{1}{2}\sum_{i=1}^{m} \partial_{y_i}^2 \rho_u,$$

the entropy satisfies

$$\frac{d}{du}\left[-\int_{\mathbb{R}^m} \rho_u(y)\log\rho_u(y)\,dy\right] = \frac{1}{2}\int_{\mathbb{R}^m} \rho_u(y)\|\nabla_y \log\rho_u(y)\|^2\,dy$$

(Wibisono & Jog, 2018, Lemma 1). It also holds for conditional entropy when the conditioning variables are unchanged by the added noise: apply the identity to each conditional density and average.

Apply this identity to $I(A; C \mid B) = H(A \mid B) - h(A \mid B, C)$ where $H(A)$ is the differential entropy. The marginal score is the conditional expectation of the full score,

$$\mathbb{E}[\nabla_A \log p_t(X(t)) \,|\, X_{AB}(t)] = \nabla_A \log p_t(x_{AB}(t)).$$

The orthogonal-projection property of conditional expectation (Durrett, 2019, Theorem 4.1.15) therefore gives

$$\begin{aligned}&\frac{d}{du} I\Big(\alpha(t)X_A(0) + \sqrt{\sigma(t)^2 + u}\, Z_A : X_C(t) \mid X_B(t)\Big)\Big|_{u=0} \\ &\quad= \frac{1}{2}\mathbb{E}_{p_t}\left[\|\nabla_A \log p_t(x_{AB}(t))\|^2 - \|\nabla_A \log p_t(x(t))\|^2\right] \\ &\quad= -\frac{1}{2}\Delta_{\mathrm{loc}}(t)^2.\end{aligned} \tag{41}$$

The *Fisher-information dissipation identity* states, for the same Gaussian heat flow, that

$$\frac{d}{du}\int_{\mathbb{R}^m} \rho_u(y)\|\nabla_y \log\rho_u(y)\|^2\,dy = -\int_{\mathbb{R}^m} \rho_u(y)\|\nabla_y^2 \log\rho_u(y)\|_{\mathrm{F}}^2\,dy$$

(Wibisono & Jog, 2018, Lemma 1). Together with de Bruijn's identity, this says that the second entropy derivative is minus one half of the mean squared log-density Hessian.

Apply this to the two conditional entropies in the CMI, we obtain

$$\begin{aligned}
&\frac{d^2}{du^2} I\Big(\alpha(t)X_A(0) + \sqrt{\sigma(t)^2+u}\, Z_A : X_C(t) \mid X_B(t)\Big) \\
&\quad = \frac{1}{2}\mathbb{E}\left\|\nabla_A^2 \log p\Big(\alpha(t)X_A(0) + \sqrt{\sigma(t)^2+u}\, Z_A \,\Big|\, X_B(t), X_C(t)\Big)\right\|_{\mathrm{F}}^2 \\
&\qquad \underbrace{- \frac{1}{2}\mathbb{E}\left\|\nabla_A^2 \log p\Big(\alpha(t)X_A(0) + \sqrt{\sigma(t)^2+u}\, Z_A \,\Big|\, X_B(t)\Big)\right\|_{\mathrm{F}}^2}_{\geq 0} \\
&\quad \leq \frac{1}{2}\mathbb{E}\left\|\nabla_A^2 \log p\Big(\alpha(t)X_A(0) + \sqrt{\sigma(t)^2+u}\, Z_A \,\Big|\, X_B(t), X_C(t)\Big)\right\|_{\mathrm{F}}^2 \\
&\quad = \frac{1}{2}\mathbb{E}\left\|\nabla_A^2 \log p\Big(\alpha(t)X_A(0) + \sqrt{\sigma(t)^2+u}\, Z_A\,, X_B(t), X_C(t)\Big)\right\|_{\mathrm{F}}^2 \\
&\quad \leq \frac{m(m+3)}{2(\sigma(t)^2+u)^2} \leq \frac{m(m+3)}{2\sigma(t)^4},
\end{aligned} \tag{42}$$

where we used equation 40 for the second last inequality. Integrating equation 42 twice gives Taylor's quadratic upper bound Boyd & Vandenberghe (2004): if $f''(u) \leq L$ for $u \geq 0$, then

$$f(u) \leq f(0) + uf'(0) + \frac{L}{2}u^2.$$

Using equation 41 and nonnegativity of CMI, we obtain

$$\begin{aligned}
0 &\leq I\Big(\alpha(t)X_A(0) + \sqrt{\sigma(t)^2+u}\, Z_A : X_C(t) \mid X_B(t)\Big) \\
&\leq I_t(A; C \mid B) - \frac{u}{2}\Delta_{\mathrm{loc}}(t)^2 + \frac{m(m+3)}{4\sigma(t)^4}u^2.
\end{aligned}$$

Choose the nonnegative minimizer of this quadratic,

$$u = \frac{\sigma(t)^4 \Delta_{\mathrm{loc}}(t)^2}{m(m+3)}.$$

Substitution gives

$$\frac{\sigma(t)^4 \Delta_{\mathrm{loc}}(t)^4}{4m(m+3)} \leq I_t(A; C \mid B).$$

Rearranging and taking a square root proves the proposition. □

A.5 Conditional scores and Classifier-Free Guidance

An important aspect of diffusion models is that they produce images that are faithful to classes of semantic information; a good model trained on cat and dog images produces one animal at a time, it does not produce an amalgamation of the two. How does the model steer towards a specific semantic class? To model this behaviour, we assume the data distribution is a joint distribution $p(x, s)$ between images x and labels s where the diffusion noise only acts on the conditional image distribution, i.e., $p_t(x, s) = p(s)p_t(x \mid s)$. Since the model only has access to the image marginal score, we can write the score as using Bayes' rule

$$\nabla_A \log p_t(x) = \nabla_A \log p_t(x \mid s) - \nabla_A \log p_t(s \mid x)\,. \tag{43}$$

Rearranging this equation, we obtain the conditioning gap

$$\Delta_{\mathrm{cond}}(t)^2 \coloneqq \mathbb{E}\|\nabla_A \log p_t(s \mid x)\|^2 = \mathbb{E}\|\nabla_A \log p_t(x) - \nabla_A \log p_t(x \mid s)\|^2\,, \tag{44}$$

whose norm answers two questions: (i) how sensitive is a global classifier to a change in A or (ii) how much does conditioning change the global score function? Just like the locality gap, it also is the fundamental score error in approximating the conditional score by any unconditional function of the global image. Furthermore, by using Tweedie on the label-conditional distribution on the full image $p_t(x \mid s)$, we can show that the conditioning gap controls both the pixelwise difference in the denoised image, as well as the extra squared error in the downstream sample due to hiding/revealing

the semantic label s. Operationally, the score difference $\nabla_A \log p_t(s \mid x)$ is precisely the term added in diffusion models by classifier guidance Dhariwal & Nichol (2021) and the application of Bayes' rule to expand it in terms of the conditional and unconditional scores is the basis of classifier-gree guidance Ho & Salimans (2022) which drops the need to train an explicit classifier to obtain the gradient $\nabla_A \log p_t(s \mid x)$. Thus, the conditioning gap probes how strongly guidance can change the score on A at each time: a small gap means that even revealing the label provides little additional direction for denoising.

B Semantic Speciation for Local Region

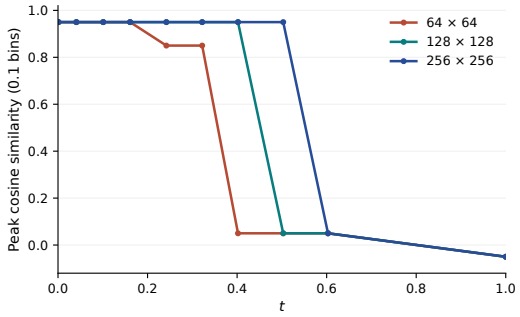


Figure 3: Label–reconstruction similarity remains high at early diffusion times and then falls toward zero. The sharp drop occurs first for the 64×64 crop, then for the 128×128 crop, and last for the full image.

We adapt the forward–backward ImageNet protocol of Sclocchi et al. Sclocchi et al. (2025), using the same unconditional 256-pixel diffusion checkpoint and 250-step respaced reverse sampler. Here $t \in [0, 1]$ denotes normalized diffusion time. We select one validation image from each of 100 classes in the ImageNet-1K (ILSVRC2012) validation set. A separate classifier chooses a class-bearing 64×64 crop from 25 candidate locations; a 128×128 crop uses the same center, and the global observation is the full image. Local crops are enlarged to 256 pixels before noising, so the reverse process receives no pixels outside the selected region. We draw one reverse sample per image and time.

Following Sclocchi et al., we measure cosine similarity between classifier logits of each source and reconstruction, see Figure 3. We standardize ConvNeXt V2 Large (Woo et al., 2023) logits using 1,000 clean reference images and plot the most populated bin of the 100 pairwise similarities. The figure uses fixed 0.10-wide bins throughout. This binned cosine peak is distinct from $P_{\mathrm{FB}}(R, t)$; enlarged crops also differ from the full images used to train the denoiser.

C Proof of Lemma 1

Proof. The mutual information is the average divergence between the posterior and the prior:

$$I_t(S; R) = \mathbb{E} \sum_s q_s \log \frac{q_s}{p(s)}.$$

For each observation, Pinsker's inequality bounds this divergence below by $\frac{1}{2}\mathbb{E}(\sum_s |q_s - p(s)|)^2$, which is at least $\frac{1}{2}\mathbb{E}\sum_s (q_s - p(s))^2$. The inequality $\log u \le u - 1$ bounds it above by $\sum_s (q_s - p(s))^2/p(s)$. Averaging proves equation 9.

For equation 10, Jensen's inequality gives $-\sum_s q_s \log q_s \ge -\log \sum_s q_s^2$. Average this inequality and use Jensen's inequality together with $P_{\mathrm{FB}}(R, t) = \mathbb{E}\sum_s q_s^2$ to get $H(S \mid X_{R,t}) \ge -\log P_{\mathrm{FB}}(R, t)$. This is the first bound because $I_t(S; R) - H(S) = -H(S \mid X_{R,t})$.

For the other bound, define

$$E = \begin{cases} 0, & S = \widehat{S}, \\ 1, & S \neq \widehat{S}, \end{cases} \qquad e := \Pr(E = 1) = 1 - P_{\mathrm{FB}}(R, t).$$

The original and returned labels are independent given $X_{R,t}$, so knowing the returned label does not reduce $H(S \mid X_{R,t})$. The entropy chain rule therefore gives

$$\begin{aligned} H(S \mid X_{R,t}) &= H(S \mid X_{R,t}, \widehat{S}) \\ &= H(S, E \mid X_{R,t}, \widehat{S}) \\ &= \underbrace{H(E \mid X_{R,t}, \widehat{S})}_{\text{Was the guess wrong?}} + \underbrace{H(S \mid E, X_{R,t}, \widehat{S})}_{\text{If wrong, which label?}} \end{aligned} \tag{45}$$

The uncertainty in this yes-or-no answer, without any additional information, is $H(E) = h_2(e)$. Knowing $X_{R,t}$ and $\widehat{S}$ can only reduce that uncertainty. Thus

$$H(E \mid X_{R,t}, \widehat{S}) \leq h_2(e).$$

Now suppose we have been told the value of E. When $E = 0$, the original label is exactly $\widehat{S}$. There is no remaining uncertainty:

$$H(S \mid E = 0, X_{R,t}, \widehat{S}) = 0.$$

When $E = 1$, the original label cannot equal $\widehat{S}$, so there are at most $L-1$ possibilities. A distribution over $L - 1$ possibilities has entropy at most $\log(L - 1)$:

$$H(S \mid E = 1, X_{R,t}, \widehat{S}) \leq \log(L - 1).$$

The second situation occurs with probability e. Averaging the two cases gives

$$H(S \mid E, X_{R,t}, \widehat{S}) \leq (1 - e) \cdot 0 + e \log(L - 1).$$

Substitute into equation 45 we obtain the second bound in equation 10. □

D EMPIRICAL PROBE OF THE COMMON-CAUSE HYPOTHESIS

We probe the common-cause interpretation in Stable Diffusion 3 Medium (Figure 2). Each scene has three nested descriptions: longer versions retain the shorter description and append semantic details. We generate trajectories at 1024×1024 resolution using 30 FlowMatch Euler steps (scheduler shift 3), global attention throughout, and classifier-free guidance of scale 4. At every pre-update latent state, we evaluate the same pretrained weights with global attention and with local attention implemented using FlexAttention. The local variant restricts image–image attention to a clipped 15×15 token neighborhood (Chebyshev radius 7) in all 24 transformer blocks, while retaining all text connections. For each description-length group, we measure

$$\Delta_b(t) = \mathbb{E}_{i,r}\left[\frac{1}{d}\left\| v_\theta^{\mathrm{loc}}(x_t^{i,r}, t, c_{i,b}) - v_\theta^{\mathrm{glob}}(x_t^{i,r}, t, c_{i,b})\right\|_2^2\right], \qquad b \in \{\mathrm{C}, \mathrm{U}\}, \tag{46}$$

where d is the number of latent coordinates, $c_{i,\mathrm{C}}$ is the scene description, $c_{i,\mathrm{U}} = \varnothing$ is the empty prompt, and the empirical expectation averages scenes i and seeds r. Both branches are evaluated on the same globally guided trajectory generated for the corresponding description; the predictions v_θ are measured before applying guidance. Because SD3 predicts flow velocity, Δ_b is a score-gap proxy.

In Figure 2, lines average over seeds and then scenes; shading denotes pointwise 95% confidence intervals from 20,000 paired scene-cluster bootstrap resamples, stratified by subject category. The time-weighted reduction $1 - \sum_k w_k \Delta_{\mathrm{cond}}(t_k) / \sum_k w_k \Delta_{\mathrm{uncond}}(t_k)$, with $w_k = t_k - t_{k+1}$, is 20.7%, 23.4%, and 24.4% for short, medium, and long descriptions. Diffusion time increases from clean ($t = 0$) to noise ($t = 1$); generation proceeds right to left. These results provide noise-dependent evidence consistent with semantic common causes. Attention truncation remains an operational probe rather than an exact marginal-score construction, so the experiment does not directly establish the mutual-information inequality in Hypothesis 1.

E GAUSSIAN MIXTURE CALCULATIONS

In this Appendix we provide details of the calculations of the score gaps and the CMI for Gaussian mixtures of various kinds.

E.1 SCORES FOR GAUSSIAN MIXTURES

In this subsection we compute various scores and score gaps for Gaussian mixtures, for which almost all results are analytic.

We consider $N \times N$ images as a random variable X_0, whose instances are flattened into vectors $\vec{x}_0 \in \mathbb{R}^{N^2}$ and use an interpolation $X_t = (1-t)X_0 + tZ$ where $Z \sim \mathcal{N}(0, \mathbb{I}_{N^2})$. Take the joint distribution of labels $S = \pm 1$ and images X to be an equal-weight ferromagnetic Gaussian mixture where $p(S = s) = 1/2$, $p(x_0 \mid S = s) \sim \mathcal{N}(s\mu\vec{1}, \sigma^2\mathbb{I}_{N^2})$, where $\vec{1} \in \mathbb{R}^{N^2}$ is the vector of ones, $\mu \in \mathbb{R}$ is a parameter setting the $O(1)$ separation scale of the means, and σ^2 is the intrinsic variance of the distribution. The observed distribution is the image marginal

$$p(\vec{x}_0) = \frac{1}{Z_{\mu,\sigma^2}} \sum_{s=\pm 1} \left[\exp(-\frac{1}{2\sigma^2}(\vec{x} - s\mu\vec{1})^2) \right] . \tag{47}$$

Since sum of Gaussian random variables remains a Gaussian, the noised distribution itself is a two-component mixture with time-dependent parameters $\mu_t = \mu(1-t)$ and $\sigma_t^2 = (1-t)^2\sigma^2 + t^2$. We thus perform calculations suppressing the time dependence and restore as needed.

We first rewrite the s-conditional Gaussian of the marginal image distribution on some region R as

$$p(x_R \mid s) = \frac{\alpha(x_R)}{Z_{\mu,\sigma^2}} \exp(sB(x_R)) , \tag{48}$$

where $B(x_R) = (\mu/\sigma^2) \sum_{i \in R} x_i$ is a soft majority vote of all pixels in region R, named B as it is effectively a 'magnetic field' bias on region R given label s. s-independent terms are grouped into the prefactor $\alpha(x_R)$ which will cancel out of relevant scores. Since the label distribution is uniform, this is all we need: the reverse conditional distributions are given by the Bayes' formula corollary

$$p(s \mid x_R) = \frac{p(x_R \mid s)}{\sum_{s'=\pm 1} p(x_R \mid s')} = \frac{\exp(sB(x_R))}{2\cosh B(x_R)} . \tag{49}$$

As a result, Bayes optimal inference of the global label is $\mathbb{E}[S \mid X_R] \equiv \sum_{s \in \pm 1} sp(s \mid x_R) = \tanh B(x_R)$. Thus the task of inferring which Gaussian the observation x_R comes from is equivalent to studying the magnetisation of a non-interacting Ising magnet.

From the two conditional distributions, we can obtain the scores directly. For example,

$$s_A(x_R \mid s) \coloneqq \nabla_{x_A} \log p(x_R \mid s) = -\frac{1}{\sigma^2}(x_A - s\mu) , \tag{50}$$

which implies

$$s_A(x_R) \coloneqq \nabla_{x_A} \log p(x_R) = -\frac{x_A}{\sigma^2} + \frac{\mu}{\sigma^2} \tanh B(x_R) , \tag{51}$$

via the trick $\nabla_{x_A} \log p(x_R) = \sum_s p(s \mid x_R) \nabla_A \log p(x_R \mid s)$. In the score s_A, the first term only includes A, and the rest of the image enters through the $\tanh$. Then we have the locality gap

$$\Delta^U_{\text{loc}} = \frac{\mu}{\sigma^2} [\tanh B(x) - \tanh B(x_{AB})] . \tag{52}$$

This gives an explicit characterisation of what the locality gap is comparing: it asks how important $\sum_{i \in C} x_i$ is in inferring the label, softened through the $\tanh$. Its peak in t is directly given by the shift induced by C.

The same results also let us calculate the global conditioning gap

$$\Delta^G_{\text{cond}} = \frac{\mu}{\sigma^2}(\tanh(B(x)) - s) . \tag{53}$$

If we look at their difference,

$$\Delta^U_{\text{loc}} - \Delta^G_{\text{cond}} = \frac{\mu}{\sigma^2}(s - \tanh(B(x_{AB})) . \tag{54}$$

E.2 CMI FOR GAUSSIAN MIXTURES

Just like the scores, the conditional mutual information for the two component Gaussian mixture is analytically reducible to a single integral.

Exact formula for CMI. To start, we recall the definition of the CMI

$$I(A;C \mid B) \coloneqq \mathbb{E}_{x_B} D_{\mathrm{KL}}(p(x_A, x_C \mid x_B) \| p(x_A \mid x_B) p(x_C \mid x_B)) \,, \tag{55}$$

$$\equiv \int \mathrm{d}x p(x) \log \frac{p(x_A, x_C \mid x_B)}{p(x_A \mid x_B) p(x_C \mid x_B)} \,, \tag{56}$$

and get rid of conditioning terms by restoring B marginals. The result is

$$I(A;C \mid B) \equiv \int \mathrm{d}x p(x) \log \frac{p(x_A, x_C, x_B) p(x_B)}{p(x_A, x_B) p(x_C, x_B)} \tag{57}$$

$$= H(AB) - H(B) + H(BC) - H(ABC) \,, \tag{58}$$

$$= I(A;C|B,S) + [I(S;AB) - I(S;B)] - [I(S;ABC) - I(S;BC)] \,. \tag{59}$$

For the joint Gaussian mixture, $I(A;C|B,S) = 0$ since conditional on the label, the distribution completely factorizes, and we only need to compute mutual informations/differential entropies associated to a region R where $R \in \{AB, BC, B, ABC\}$.

The entropy $H(R)$ is

$$H(R) = -\int \mathrm{d}x_R p(x_R) \log p(x_R) = \frac{1}{2\sigma^2} \int \mathrm{d}x_R p(x_R) (x_R - \mu_R)^2 \\ - \int \mathrm{d}x_R p(x_R) \log(1 + \exp(-\frac{2}{\sigma^2} x_R \cdot \mu_R)) \,, \tag{60}$$

obtained by factoring out one of the Gaussians inside the $\log$. The first term gives two trivial gaussian integrals by expanding $p(x_R)$ as a sum of two Gaussian integrals (which we do not write out, they will cancel out between all the terms in the CMI equation 57). The second term is the nontrivial one because it includes a sum inside the $\log$, as well as integrals over all the pixels in R. However, since the integrand only depends on the dot product $x_R \cdot \mu_R$, we may write $x_R = x_R^{\parallel} + x_R^{\perp}$ where $x_R^{\perp} \cdot \mu_R = 0$ and $\mathrm{d}x_R = \mathrm{d}x_R^{\parallel} \mathrm{d}x_R^{\perp}$, and $\mathrm{d}x_R^{\perp}$ integrates out to give a constant that cancels with some of the normalisation in $p(x_R)$. The result is a one-dimensional integral

$$H(R) = W(R) - \frac{1}{Z_R^{\parallel}} \int \mathrm{d}x_R^{\parallel} \exp\left(-\frac{(x_R^{\parallel} - \|\mu_R\|)^2}{2\sigma^2} \right) \log\left(1 + \exp(-\frac{2}{\sigma^2} \|\mu_R\| x_R^{\parallel}) \right) \,. \tag{61}$$

Lastly, we perform a change of variable by writing the parallel component as 'mean + fluctuations', i.e. $x_R^{\parallel} = \mu_R + \sigma z$, which gives

$$H(R) = W(R) - \frac{1}{\sqrt{2\pi}} \int \mathrm{d}z \exp\left(-z^2/2\right) \log\left(1 + \exp(-2K^2 - 2Kz)\right) \,, \tag{62}$$

in terms of one free parameter, the signal-to-noise ratio $K_R \coloneqq \|\mu_R\|/\sigma$. Defining

$$F(K) \coloneqq \frac{1}{\sqrt{2\pi}} \int \mathrm{d}z \exp\left(-z^2/2\right) \log\left(1 + \exp(-2K^2 - 2Kz)\right) \,, \tag{63}$$

we can write the CMI as

$$I(A;C \mid B) = F(K_B) - F(K_{AB}) - [F(K_{BC}) - F(K_{ABC})] \,. \tag{64}$$

Asymptotic expansion for the F integral. The integral $F(K)$ can be exactly evaluated in the large signal limit $K \gg 1$. Performing another change of variables $y = 2K(K+z)$ yields

$$F(K) = \frac{1}{2K} \exp(-K^2/2) \int \frac{\mathrm{d}y}{\sqrt{2\pi}} \exp\left(-y^2/8K^2\right) \exp(y/2) \log\left(1 + \exp(-y)\right) \,, \tag{65}$$

and as $K \to \infty$, $\exp\left(-y^2/8K^2\right) \to 1$, and

$$F(K) \asymp \sqrt{\frac{\pi}{2K^2}} \exp(-K^2/2) \,, \tag{66}$$

where $\int \mathrm{d}y \exp(y/2) \log\left(1 + \exp(-y)\right) = 2\pi$ by elementary methods.

Diverging Markov Length. Now send the image size to infinity, growing both $|B|$ and $|C|$ as $O(N^2)$ while $|A| = O(1)$, a 'thermodynamic limit' relevant for denoising a small patch. This yields, in the $K \to \infty$ limit, a simple form for the CMI which reads

$$I(A;C \mid B) \asymp \sqrt{\frac{\pi}{2K_B^2}} \exp\left[-\frac{\mu^2}{2\sigma^2}|B|\right]\left(1-\exp\left[-\frac{\mu^2}{2\sigma^2}|A|\right]\right), \tag{67}$$

up to exponentially small corrections in $|C|$. Since this decays faster than exponential in $r \propto \sqrt{B}$, the Markov length $\xi = 0$, and there is no locality transition whenever K is large.

If we recall the noise path $X_t = (1-t)X_0 + tZ$ with Z standard Gaussian noise, we have a Gaussian mixture for every $t = 1$ with effective parameters $\mu_t = \mu(1-t)$ and $\sigma_t^2 = (1-t)^2\sigma^2 + t^2$. As the noise is i.i.d. on every patch, the signal to noise for any R is

$$K_R^2(t) = \frac{(1-t)^2\mu^2}{(1-t)^2\sigma^2 + t^2}|R|\,. \tag{68}$$

Is the signal-to-noise large as $N \to \infty$ for every $0 \le t \le 1$? Taking $t = 1 - k/N$ yields

$$K_R^2(t) \asymp \mu^2 k^2 |R|/N^2 = O(1)\,, \tag{69}$$

for $R = B$ and $R = BC$, and the asymptotic expansion fails. In fact, in this limit, we are able to show Markov length $\xi \to \infty$. Noticing that $K_{R_1R_2}^2 = K_{R_1}^2 + K_{R_2}^2$ for disjoint R_1 and R_2, we can write

$$F(K_R(t)) - F(K_{AR}(t)) = -K_A(t)^2 F'(K_R(t))\,, \tag{70}$$

where we use $K_A \to 0$ and the $'$ denotes differentiation with respect to K^2. The CMI takes the form

$$I_t(A;C \mid B) \asymp K_A^2(t)[F'(K_{BC}(t)) - F'(K_B(t))]\,, \tag{71}$$

which we try to maximize over t. In particular, setting

$$I_{\text{peak}} = \sup_t I_t(A;C \mid B)\,, \quad t_* = \arg\max_t I_t(A;C \mid B) \tag{72}$$

we are able to show no exponential decay exists. First, define an effective variance parameter

$$\sigma_{\text{eff}}^2(t) = \sigma^2 + \frac{t^2}{(1-t)^2}\,, \tag{73}$$

such that we may interpret the noise path as purely expanding the variances. Then, plug in the definition of $K_R(t)$, which gives

$$\sup_t K_A^2(t)[F'(K_{BC}(t)) - F'(K_B(t))] = \sup_t \frac{\mu^2|A|}{\sigma_{\text{eff}}^2(t)}\left[F'\left(\frac{\mu^2(|B|+|C|)}{\sigma_{\text{eff}}^2(t)}\right) - F'\left(\frac{\mu^2|B|}{\sigma_{\text{eff}}^2(t)}\right)\right]. \tag{74}$$

The limit keeps $|B|/|C| = O(1)$, so let's isolate that parameter, setting $\lambda = 1 + |C|/|B|$, we get

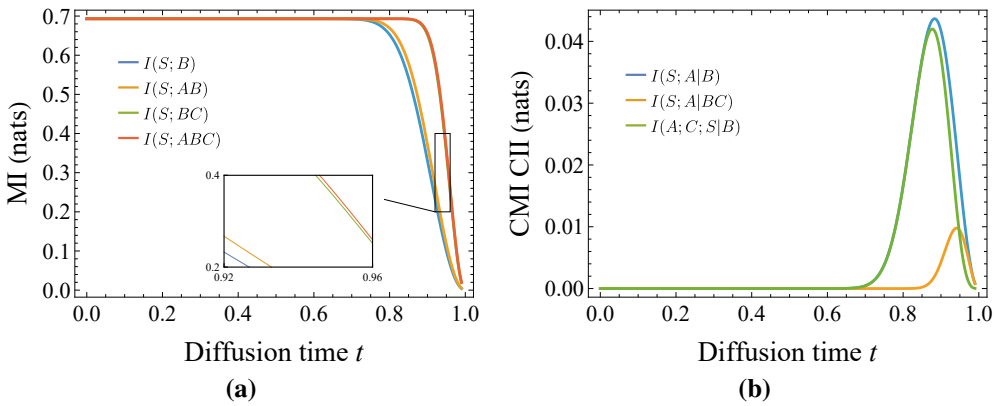
$$\sup_t K_A^2(t)[F'(K_{BC}(t)) - F'(K_B(t))] = \frac{|A|}{|B|}\sup_{K_B} K_B^2\left[F'\left(\lambda K_B^2\right) - F'\left(K_B^2\right)\right], \tag{75}$$

where we also set $\mu^2|B|/\sigma_{\text{eff}}^2(t) = K_B^2$. Also note that the supremum over t is equivalent to supremum over K_B^2, which is the only t dependent parameter, which is the main point of this substitution. Crucially, the supremum over K_B^2 yields just a number and does not depend on $|B|$ by itself anywhere. As a result, $I_{t_*}(A;C \mid B) \asymp 1/|B|$, and the Markov length is divergent.

Lastly, we can obtain k explicitly. Noting that we need $K_B^2(t_*) \asymp O(1)$ at the supremum, we need to obtain

$$\frac{\mu^2|B|}{\sigma_{\text{eff}}^2(t_*)} = O(1) \implies 1 - t_* = \frac{1}{\mu}\sqrt{\frac{K_B^2(t_*)}{|B|}}\,. \tag{76}$$

Since $|B| = O(N^2)$, the peak time t_* scales as $1 - 1/N$, as expected.



(a) **(b)**

Figure 4: Exact numerical evaluation of the Gaussian-mixture information measures with specific values given in E.3. **(a)** Mutual information curves for the four spatial regions. The cliff diagnoses the symmetry-breaking phase transition. Inset: zoomed-in region near critical window. **(b)** Local and global conditional information gaps $I_t(S;A|B), I_t(S;A|BC)$, and their difference which gives conditional interaction information $I_t(A;C;S|B)$. The peak diagnoses the non-locality transition.

The time windows. Since we can control exactly the CMI in the Gaussian mixture, let us try to give asymptotic formulas for each, demonstrating that the windows need not be exactly identical.

First, recall the nonlocality times $t_{\text{nonloc}}^{\text{start/end},\delta}$ are defined by finding times such that the CMI $I_t(A;C\mid B) \approx \delta$ by Defn. 1. We assume we pick δ as a fixed fraction of the peak CMI value. Since we already evaluated the CMI in the transition window, we immediately know scale identically as the peak, i.e. $t_{\text{nonloc}}^{\text{start/end},\delta} \sim 1 - O(1/N)$.

Second, recall that $t_{\text{spec}}^{\text{start/end},\delta}$ is defined *mutual informations*. For the start, we want to find the first t such that $I_0(S;AB) - I_t(S;B) = \delta$. Assuming δ is smaller than the peak value, this is equivalent to calculating $F(K_B(t)) - F(K_{AB}(0)) = \delta$. Since B occupies a large fraction of the image, the second term is exponentially suppressed in $|B|$ and we may drop it. For large signal to noise, we then just need to solve

$$\frac{\exp(-K_B^2(t_{\text{spec}}^{\text{start},\delta})/2)}{K_B} = \delta \implies K_B^2(t_{\text{spec}}^{\text{start},\delta}) = 2\log\frac{1}{\delta} + O(\log\frac{1}{K_B})$$
$$\implies 1 - t_{\text{spec}}^{\text{start},\delta} = \frac{1}{\mu}\sqrt{\frac{K_B^2(t_{\text{spec}}^{\text{start},\delta})}{|B|}} \sim \sqrt{\frac{\log 1/\delta}{|B|}}\,. \quad (77)$$

Focusing near $t = 1$, near the critical time $\delta \sim 1/N^2$, we thus get $1 - t_{\text{spec}}^{\text{start},\delta} \sim \sqrt{\log(N)}/N$.

Lastly, the speciation end time is defined as the largest time such that $I_t(S;ABC) = \delta)$. Near $t \to 1$, when the global label is degraded, the signal to noise is very small, and instead we evaluate the the integral in equation 63 in the small K limit. This yields $I(S;ABC) = \log 2 - F(K_{ABC}) = K_{ABC}^2/2$ which we set equal to δ. Using the formula for the exponent being $O(1)$, we have

$$1 - t_{\text{spec}}^{\text{end},\delta} = \frac{\sqrt{\delta}}{\sqrt{|ABC|}} = O(1/N^2)\,, \quad (78)$$

where we again took $\delta = O(1/N^2)$. This proves that the locality window has a size proportional to $1/N$, while the semantic window has the scalings $1 - O(\sqrt{\log N}/N)$ and $1 - O(1/N^2)$.

E.3 Numerical evaluation of the exact GMM formulas

We numerically evaluate the exact information-theoretic formulas derived above for the ultra-local two-component Gaussian mixture model. We take the homogeneous mean vector $\mu = (1,\ldots,1)$, unit bare variance $\sigma^2 = 1$, and a 20×20 flattened system. The spatial partition is chosen as

$$a = \|\mu_A\|^2 = 16, \qquad b = \|\mu_B\|^2 = 80, \qquad c = \|\mu_C\|^2 = 304,$$

corresponding to a local patch A, a buffer B occupying 20% of the system, and the remaining context C. The forward noising process is

$$X_t = (1-t)X_0 + tZ, \qquad Z \sim \mathcal{N}(0, I).$$

We numerically evaluate the core integral $F(W)$ by adaptive numerical quadrature and the resulting curves are interpolated on a uniform grid $t \in [0, 0.99]$ with spacing $\Delta t = 0.005$.

Fig. 4(a) tracks the four mutual informations $I_t(S;B), I_t(S;AB), I_t(S;BC), I_t(S;ABC)$. These curves are monotone decreasing under the forward diffusion dynamics and develop sharp late-time cliffs. The smaller regions lose semantic information earlier, while the global regions remain informative until later times. This is the finite-size numerical signature of the forward-backward semantic transition.

Fig. 4(b) plots the two conditional gaps $I_t(S;A|B) = I_t(S;AB) - I_t(S;B)$, and $I_t(S;A|BC) = I_t(S;ABC) - I_t(S;BC)$. These gaps peak near the steepest parts of the corresponding MI cliffs. For the parameters used here, G_{loc} peaks near $t \simeq 0.88$, whereas G_{glob} is smaller and delayed, peaking near $t \simeq 0.94$. This delay reflects the greater robustness of the global context ABC to the forward noise.

The conditional interaction information $I_t(A;S;C|B) = I_t(S;A|B) - I_t(S;A|BC)$ is also illustrated in Fig. 4(b). For the identity-covariance ultra-local GMM, conditioning on the semantic label removes all residual spatial dependence, so $I_t(A;S;C|B) = I_t(A;C|B)$. Thus the same curve is also the unconditional CMI diagnosing non-locality. The CII/CMI curve has a single sharp peak, numerically located near $t \simeq 0.88$ for the present parameters, coinciding with the local semantic information cliff. This provides an exact-calculation check of the proposed concurrence between the symmetry-breaking and non-locality transitions.

F Convergence from separated class means

We state and prove the formal version of the convergence result discussed in the main text.

Theorem 4 (Convergence from separated class means)**.** *(Formal version of Theorem 2)Let S have a fixed finite set of labels with positive prior probabilities. Suppose that the common-cause hypothesis holds with a size-independent $\alpha > 0$. For each N, let B be annulus with radius N, and use $X_t = (1-t)X_0 + tZ$, where $Z \sim \mathcal{N}(0, I)$ is independent of (X_0, S). Write $m_s = \mathbb{E}[X_{B,0} \mid S = s]$. Suppose there are constants $D_N, C_N > 0$, potentially scaling with N, such that, for all $s \neq s'$ and all $u \in \mathbb{R}^{|B|}$,*

$$\|m_s - m_{s'}\|_2^2 \geq D_N, \tag{79}$$

$$\mathbb{E}\Big[e^{u^\top (X_{B,0} - m_s)} \mid S = s\Big] \leq e^{C_N \|u\|_2^2/2}. \tag{80}$$

Suppose the separation between semantic means D_N scales faster than the tail bound within class C_N so that they admit a choice of δ_N below the CMI peak at each size, $0 < \alpha\delta_N < H(S)/2$ and

$$\log \frac{H(S)}{\alpha\delta_N} = o(D_N/(C_N+1)), \tag{81}$$

then both speciation endpoints and both nonlocality endpoints converge to $t = 1$. In particular,

$$1 - t_{\mathrm{spec}}^{\mathrm{start},\alpha\delta_N} = O\left(\sqrt{\frac{\log[H(S)/(\alpha\delta_N)]}{D_N/(C_N+1)}}\right). \tag{82}$$

Proof. First, we bound how often a noisy observation of B is assigned the wrong label. In class s, the mean observation is $(1-t)m_s$. Use the rule that picks the closest class mean. It can mistake s for s' only if the observation crosses half the gap between these means.

Let unit vector $e = (m_{s'} - m_s)/\|m_{s'} - m_s\|_2$ point toward that competing mean. Comparing the squared distances to $(1-t)m_s$ and $(1-t)m_{s'}$ shows that the displacement $Y = X_{B,t} - (1-t)m_s$ must satisfy

$$e^\top Y \geq z, \qquad z = \frac{1-t}{2}\|m_{s'} - m_s\|_2 \tag{83}$$

for an error to occur.

Conditional on $S = s$, we have $Y = (1-t)(X_{B,0} - m_s) + tZ_B$. To bound the chance of crossing the midpoint, we use a Chernoff bound on the conditional r.v. $e^\top Y \mid S$. In particular, we have

$$\begin{aligned}\Pr(\text{pairwise crossing} \mid S = s) &= \Pr(e^\top Y \geq z \mid S = s), \\ &= \Pr(e^{\lambda e^\top Y} \geq e^{\lambda z} \mid S = s), \\ &\leq e^{-\lambda z}\mathbb{E}[e^{\lambda e^\top Y} \mid S = s],\end{aligned} \tag{84}$$

where we take any $\lambda > 0$, and the last step uses Markov's inequality. Plugging $u = \lambda(1-t)e$ into 80 yields

$$\mathbb{E}\Big[e^{\lambda e^\top (1-t)(X_{B,0} - m_s)} \mid S = s\Big] \leq e^{C_N \lambda^2 (1-t)^2/2}. \tag{85}$$

Together with $\mathbb{E}\Big[e^{\lambda e^\top tZ_B} \mid S = s\Big] = e^{\lambda^2 t^2/2}$ by Gaussianity, we obtain

$$\Pr(\text{pairwise crossing} \mid S = s) \leq \exp\left[-\lambda z + \frac{\lambda^2}{2}\big(C_N(1-t)^2 + t^2\big)\right]. \tag{86}$$

For $t < 1$, right hand side is minimized at $\lambda = z/[C_N(1-t)^2 + t^2]$, yielding $\exp[-z^2/(2(C(1-t)^2 + t^2))]$. By 79, $z \geq (1-t)\sqrt{D_N}/2$; also $C_N(1-t)^2 + t^2 \leq C_N + 1$. A union bound over the $L-1$ competing labels gives the following estimate for $t < 1$:

$$\Pr(\text{nearest-mean error}) \leq (L-1)e^{-D_N/(C_N+1)(1-t)^2/8}. \tag{87}$$

Next, we connect classification error to the forward–backward success rate. Given $X_{B,t}$, write $q_s = p_t(s \mid X_{B,t})$. The best classifier chooses the label with largest q_s, so its error probability is $\mathbb{E}[1 - \max_s q_s]$. Since $1 - \sum_s q_s^2 \leq 2(1 - \max_s q_s)$, equation 5 and 87 give

$$1 - P_{\mathrm{FB}}(B,t) \leq 2\mathbb{E}[1 - \max_s q_s] \leq 2(L-1)e^{-D_N/(C_N+1)(1-t)^2/8}. \tag{88}$$

The two-sided sandwich bounds in 10 now give $H(S \mid X_{B,t}) \leq h_2(P_{\mathrm{FB}}(B,t)) + (1 - P_{\mathrm{FB}}(B,t))\log(L-1)$. For fixed L, the right-hand side is at most a constant times $\sqrt{1 - P_{\mathrm{FB}}(B,t)}$. Combining this with the error bound above yields

$$H(S \mid X_{B,t}) \leq Ke^{-kD_N/(C_N+1)(1-t)^2}, \tag{89}$$

where $K, k > 0$ do not depend on N or t.

Now set $t_N = 1 - b\sqrt{\log[H(S)/(\alpha\delta_N)]/(D_N/(C_N+1))}$, with b a fixed large constant. The threshold condition makes $t_N \to 1$, so $t_N \in [0,1]$ for large N. For every $t \leq t_N$, equation 89 is at most $K[\alpha\delta_N/H(S)]^{kb^2}$. Choose b so that $kb^2 > 1$ and $K2^{-(kb^2-1)} < H(S)$. Since $\alpha\delta_N/H(S) < 1/2$, this bound is smaller than $\alpha\delta_N$. Finally,

$$I_0(S; AB) - I_t(S; B) \leq H(S) - I_t(S; B) = H(S \mid X_{B,t}). \tag{90}$$

Thus the speciation start is no earlier than t_N, which proves 82. Theorem 1 places the other three endpoints between that start and 1. All four endpoints therefore converge to 1. □

G LIMITATIONS

We identify some limitations of our work below. First, under the common-cause hypothesis, we establish that the nonlocality window lies within the speciation window. We do not establish the converse: speciation need not produce appreciable nonlocal dependence. In particular, Theorem 1 assumes that the nonlocality window is nonempty rather than proving its existence. Containment also permits the nonlocality window to be much narrower than the speciation window and does not imply that their peaks coincide. We also leave the predicted relationship in autoregressive models to future work.

Moreover, our empirical test of the common-cause hypothesis is indirect. Directly estimating the conditional mutual information gap $I_t(A; C \mid B) - I_t(A; C \mid B, S)$ from finite samples is challenging for high-dimensional image distributions. We therefore probe the common-cause hypothesis using a score-gap proxy in Stable Diffusion 3. The observed reduction is consistent with the hypothesis but does not establish its required inequality.

In addition, our information-theoretic bounds constrain the relationship between the two windows, but do not predict their locations or widths without further properties of the data distribution. These quantities depend on how semantic information is distributed across regions and how it survives noise. A sharp phase transition additionally requires a specified sequence of growing systems, a noise schedule, and a scaling of the observation regions and information thresholds. Our separated-class-means criterion gives sufficient conditions for the windows to close, and Gaussian mixtures provide an explicit solvable example. Determining the windows for broader distributions requires additional control of their semantic information along the noise trajectory.

Another limitation is that our calculations herein only reflect distributions which are uncorrelated within each class and thus lack detailed spatial locality structure (our Gaussian mixtures are in some sense extremely local). While we have not explicitly discussed incorporating a non-identity covariance matrix in this manuscipt, the calculation is not much harder and does not change the behaviour of the CMI. For example, gaussian mixture with a shared covariance matrix with exponentially-decaying inverse matrix has same transition windows as the ones we presented. Thus we expect the soft-majority vote features of the Gaussian mixture to survive mild but reasonable local correlations.

Lastly, we remark that our results are about Bayes-optimal denoising on the data distribution. Details about real models, such as training errors, numerical sampling effects, model architecture can influence the phenomenology of the windows which we cannot control under our theory. We leave building more fine-grained connections between our information-theoretic approach and features of real models to future work.