How Does Denoising Guidance Encourage Creativity in Diffusion Models?

Anonymous authors
Paper under double-blind review

Abstract

Diffusion models can generate images that go beyond the examples seen during training, but where does this creativity come from? We show that an important part of it can emerge in the *tail of denoising* with denoising guidance. Classifier-free guidance after substantial object structure has formed can preserve pose and layout while guiding the semantic content of generated images. We analyze the tail of deterministic DDIM sampling and show that its local geometry is governed by a *conditional PCA* whose dominant directions depend jointly on the intermediate denoising state and the guidance class. Iterative denoising amplifies these directions through a power-iteration mechanism. We test these predictions using only post-training analysis of a pretrained DiT model trained on ImageNet, and observe the predicted class- and image-dependent geometry. Our results identify a simple geometric mechanism through which guidance can encourage compositional creativity in diffusion models.

1 Introduction

How can diffusion models, trained to reconstruct their training samples, generate novel images? Reconstruction is built into their design: the forward process gradually adds noise to training samples, and the reverse process is trained to remove it (Ho et al., 2020; Song et al., 2021b). Yet modern diffusion models do more than denoise. They generate images never seen during training, combine familiar concepts in new ways, and produce markedly different outcomes under different forms of conditioning (Ho & Salimans, 2021; Rombach et al., 2022; Liu et al., 2022; Blattmann et al., 2022). How, then, does denoising give rise to *creativity*?

A growing line of research studies when diffusion models sample from the underlying population distribution rather than reproducing their training data, often referred to as *generalization* (Biroli et al., 2024; Bonnaire et al., 2025). Under the exact score of the empirical distribution, reverse diffusion eventually recovers individual training samples (Biroli et al., 2024), and prolonged training can similarly drive learned models toward memorization (Bonnaire et al., 2025). Generalization can arise when the learned score is instead less tied to the empirical distribution. With sufficiently large datasets, models trained on different subsets learn similar score functions (Kadkhodaie et al., 2024), while architectural biases such as locality and equivariance can lead convolutional models to learn patch-based scores rather than exactly fit the empirical score (Kamb & Ganguli, 2025). These results address when a diffusion model reproduces its training data and when it can instead generate new samples from the same underlying distribution, but not how novelty emerges during denoising.

But creativity is beyond generalization. Consider the iconic example of prompting Stable Diffusion with "a photograph of an astronaut riding a horse." The interesting part is not simply that the generated image differs from the training samples. The model must bring together visual structure learned in different contexts and make these concepts coexist in a configuration that may never have appeared in the training data. We refer to this ability as *compositional creativity* (Liu et al., 2022). We study how such compositions emerge during denoising. In particular, Can one condition shape the canvas, while another takes the brush and determines what appears on it? A potential candidate for canvas-brush division is the guidance mechanism, capable of steering the reverse process toward a given condition (Dhariwal & Nichol, 2021; Ho & Salimans, 2021). Related work has shown that changing the condition partway through sampling can redirect or blend the semantic content of the generated image (Liew et al., 2022; Olearo et al., 2026; Choi et al., 2024; Görgün et al., 2026). We use this

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



Figure 1: **Signal vanishing in forward diffusion models.** Images from left to right are obtained by adding a gaussian noise to the original images with increasing variance starting from variance zero. Images on right can be interpreted as either dog or eagle.

intervention as a probe. By switching the condition during sampling, we can ask which aspects of the image have already been established and which can still respond to the new condition. The early denoising steps may shape the canvas by fixing the coarse spatial organization (Wang & Vastola, 2023), while later guidance acts more like a brush, deciding how that structure is interpreted and completed. The key question is whether the brush can only fill in what is already there, or whether it can still redraw the composition itself.

The forward diffusion process offers a useful clue. Figure 1 follows a dog and an eagle as increasing Gaussian noise is added. Noise does not erase all information at once. At first, both the spatial structure and semantic identity of each image are clear. As the noise increases, however, the features that distinguish a dog from an eagle begin to disappear while much of the spatial organization remains. This leaves an intermediate regime in which the image still carries information about *where* the object is and how it is arranged, but much less about *what* the object is.

Reading the same process backward suggests where guidance begins to turn structure into creativity. If forward diffusion forgets *what* before it forgets *where*, then denoising may recover spatial organization before semantic identity is fully settled. The intermediate image can then serve as a scaffold. Its location, pose, and coarse structure constrain what comes next, while guidance steers how that structure is interpreted and completed. Changing the condition at this stage may therefore change *what* the image becomes without rebuilding *where* everything is from scratch. We refer to these later steps, where structure is already visible but semantics may still be open to change, as the *tail of the denoising process*. This tail is the focus of our study.

We find that this late-stage role of guidance has a simple geometric interpretation in the denoising tail. At a late intermediate state, much of the spatial structure is already constrained, yet several clean completions may still be plausible. We show that the dominant directions of this remaining variation are captured by a *conditional PCA*. These directions do not form a fixed basis for each class. Instead, they depend jointly on the condition and on the particular intermediate image. The image anchors the spatial structure, while the condition determines how that structure can plausibly vary. Changing the condition therefore changes the directions of completion while remaining faithful to the structure already present in the image.

The denoising tail provides the second part of the mechanism. As sampling continues, successive denoising steps repeatedly act along these conditional directions in a manner analogous to *power iteration*. Dominant directions are reinforced from one step to the next and become increasingly pronounced toward the end of the trajectory. The intermediate image anchors the emerging structure, the condition determines its plausible completions, and the remaining denoising steps amplify the dominant directions until the final semantic structure emerges. This links the conditional geometry of an intermediate state directly to what emerges at the end of denoising.

Contributions. We study how guidance can produce new semantic completions of an image during the tail of denoising. On the theoretical side, we analyze the dynamics of deterministic DDIM sampling (Song et al., 2021a). We show that, in the late denoising regime, the DDIM trajectory can be understood through repeated clean-image estimation, whose local geometry is governed by

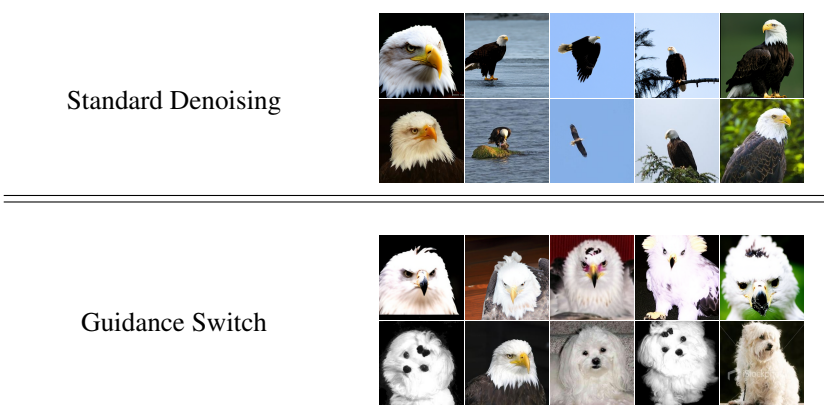


Figure 2: **Encouraging creativity through guidance switching. First block:** The first row shows images generated from the "bald eagle" class using standard DDIM denoising (Song et al., 2021a), and the second row shows their nearest neighbors in the training dataset under the LPIPS metric. **Second block:** The first row shows images generated by initially denoising toward the "Maltese dog" class and then switching the guidance to the "bald eagle" class. The second row shows the corresponding nearest-neighbor training images. Images generated with guidance switching are visually distinct from the displayed nearest neighbors; in particular, the guidance switch makes the "Maltese dog" images look more like "bald eagles".

a conditional PCA that depends jointly on the intermediate image and the conditioning class. This view further reveals a power-iteration mechanism through which repeated denoising progressively amplifies the dominant conditional directions. On the empirical side, we test these predictions directly on the pretrained DiT model of Peebles & Xie (2023) trained on ImageNet (Deng et al., 2009). We find that the class- and image-dependent directions predicted by the theory are already present in this large-scale pretrained model and become increasingly pronounced along the denoising tail.

2 Creativity via Guided Denoising

Diffusion models are prone to overfitting when the training dataset is small. In particular, Kadkhodaie et al. (2024) report that, in their CelebA experiments, at least 10,000 training samples were required for their diffusion model to generate images that are significantly different from the training data. This number is substantially larger than the number of training examples available for each class in ImageNet. For instance, the ImageNet training set contains only approximately 1,300 images from the "bald eagle" class. This per-class sample size is significantly smaller than the dataset sizes reported as necessary for generalization in such experiments (Kadkhodaie et al., 2024; Bonnaire et al., 2025). Nevertheless, diffusion models such as Facebook's DiT (Peebles & Xie, 2023) exhibit remarkable creativity in sample generation. How can these models generate creative images when each ImageNet class contains only approximately 1,300 training examples?

Some of this success can, of course, be attributed to representations learned jointly from images of different classes. We suspect, however, that a deeper source of creativity lies in how guidance brings these representations together during denoising. Our main insight is that the post-training denoising mechanism can substantially encourage creativity, generating images that differ significantly in visual appearance from the training examples. To illustrate this phenomenon, Figure 2 shows five images generated from the "bald eagle" class using a pretrained diffusion model, together with their nearest neighbors in the training dataset under the LPIPS metric (Zhang et al., 2018). Although the generated images are already visually distinct from their nearest neighbors, we can encourage an even higher degree of novelty by modifying the guided denoising mechanism. In particular, we use a guidance-switching mechanism that initially denoises the image toward the "Maltese dog" class and subsequently switches the guidance to the "bald eagle" class. This switch produces eagle-like images that are visually different from their nearest-neighbor training samples. Related switching mechanisms have previously been used in practical applications (Liew et al., 2022; Choi et al., 2024), but their underlying mechanisms and their role in creative generation remain largely unexplored.

3 THE GUIDED DENOISING MECHANISM

The main idea behind the guidance mechanism is biasing the denoising process of diffusion models towards a particular image class (Ho & Salimans, 2021). The denoiser is trained on data $x^{(c)}$ drawn from the conditional distribution $p(x \mid c)$ by minimizing the denoising objective

$$f^* \in \arg\min_{f \in \mathcal{F}} \mathbb{E} \left\| f_{t,c}(\alpha_t x^{(c)} + \sigma_t z) - z \right\|^2. \tag{3.1}$$

Here $\mathcal{F}$ is a parametric function class, such as a family of neural networks. The Gaussian noise $z \sim \mathcal{N}(0, I_d)$ is drawn independently of $x^{(c)}$. The expectation averages over the training images and their class labels, the sampled time t, and the noise z. The coefficients $\alpha_t = (1 + e^{-t})^{-1}$ and $\sigma_t^2 = 1 - \alpha_t^2$ determine the signal-to-noise ratio, with $t \in \mathbb{R}$.

Throughout the paper, generation proceeds in increasing time, so the signal coefficient α_t increases and the noise coefficient σ_t decreases. We use m denoising updates along the schedule

$$t_1 < \cdots < t_m < t_{m+1} = +\infty.$$

We use $t_{m+1} = +\infty$ to denote the final, noise-free output, where $\alpha_{t_{m+1}} = 1$ and $\sigma_{t_{m+1}} = 0$. The network is evaluated only at the finite times $t_1, \ldots, t_m$.

We use f for the noise predictor and g for the denoiser. Given f, define

$$g_{t,c,w}(x) = \frac{x - \sigma_t\big((1+w) f_{t,c}(x) - w f_{t,\varphi}(x)\big)}{\alpha_t}. \tag{3.2}$$

Here φ denotes unconditional prediction and $w \geq 0$ is the guidance weight (Ho & Salimans, 2021). We write f^* for a minimizer of (3.1) and g^* for the corresponding denoiser. We consider the case $w = 0$ and write

$$g_{t,c}(x) := g_{t,c,0}(x) = \frac{x - \sigma_t f_{t,c}(x)}{\alpha_t}.$$

For a fixed class c, deterministic DDIM sampling combines the denoiser and noise predictions to obtain the next sampling state (Song et al., 2021a),

$$x_{t_{j+1}} = \alpha_{t_{j+1}} g_{t_j,c}(x_{t_j}) + \sigma_{t_{j+1}} f_{t_j,c}(x_{t_j}). \tag{3.3}$$

Both predictions are evaluated at the current state x_{t_j}. At the final step, this update returns $g_{t_m,c}(x_{t_m})$.

We take a closer look at the final s steps of denoising, with $1 \leq s \leq m$. Let y be the intermediate image at their starting time $t = t_{m-s+1}$. In this tail, the schedule gives $1 - \alpha_t = O(e^{-t})$ and $\sigma_t = O(e^{-t/2})$. This motivates the simpler continuation

$$\big(g_{t_m,c} \circ \cdots \circ g_{t_{m-s+1},c}\big)(y), \tag{3.4}$$

in which each clean-image estimate becomes the input to the next denoiser. We compare this composition with the DDIM updates in (3.3), starting from the same y and using the same conditioning class and time schedule. For a fixed number of remaining steps and under the regularity assumptions on the predictors in $\mathcal{F}$ stated in Section B, their final outputs differ by $O(e^{-t/2})$ as $t \to +\infty$. Consequently, we use (3.4) to study these final steps.

We study a particular denoising mechanism based on guidance switching (Liew et al., 2022; Choi et al., 2024), where the denoising process is obtained by stitching two denoisers from different classes. This approach has two stages. Step (1) denoises under class c_1 to obtain an intermediate image y. Step (2) continues from this image under class c_i. With s steps remaining, the simplified model gives

$$G_s^{(i)}(y) := \overbrace{\big(g_{t_m,c_i} \circ \cdots \circ g_{t_{m-s+1},c_i}\big)}^{\text{step (2)}}\big(\underbrace{\quad y \quad}_{\text{output of step (1)}}\big). \tag{3.5}$$

Step (1) pushes the noise toward a distribution that can be viewed as consisting of indistinguishable noisy images resembling multiple classes, such as dogs and eagles. Step (2) then branches the denoising process toward different image classes. Our tail analysis starts from the shared image y and studies the remaining class-conditioned maps. The first branch continues with c_1, while the other branches use their respective classes c_i. Each branch receives the same y.

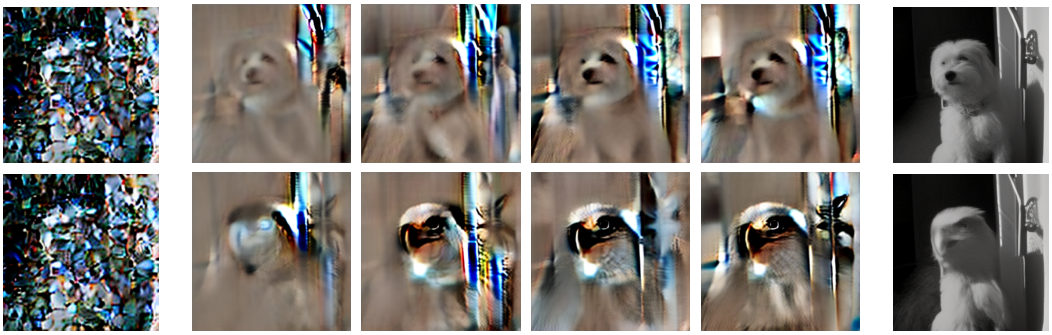


Figure 3: **Spectrum of guided denoising.** The first column shows the intermediate image y at the switching time $t = t_{m-s+1}$, corresponding to diffusion-model timestep 700 in the implementation, after seven updates toward the "Maltese dog" class. The next four columns show leading Jacobian eigendirections evaluated at this image, under "Maltese dog" conditioning in the first row and "bald eagle" conditioning in the second row. The final column shows the corresponding output obtained by continuing the denoising tail under the same conditioning class. The eigendirections already contain recognizable class-specific features that later appear in the output, while their spatial organization follows the structure present in the intermediate image. The directions are rescaled for visualization. The model is the pretrained DiT of Peebles & Xie (2023) on ImageNet (Deng et al., 2009).

4 Tail Denoising as Power Iteration

We observe that the outcome of the tail denoising steps can be predicted from the Jacobian of the denoising function $Jg^*_{t,c}(y)$, evaluated at the shared intermediate image y and at the first tail time $t = t_{m-s+1}$ in (3.5), where the switching mechanism begins. In particular, a small number of Jacobian eigendirections computed at this point already contain recognizable features of the final denoised images. Figure 3 shows these directions alongside the corresponding outputs and illustrates their close visual similarity.

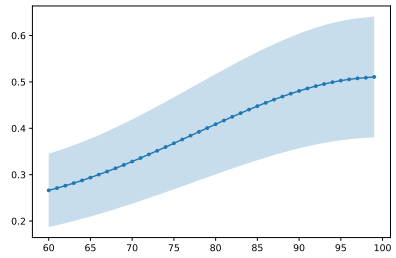


Figure 4: **Tail denoising projection.** Each image is first denoised unconditionally for 60 steps to obtain the intermediate image y at the switching point. A class c is then sampled uniformly from the ImageNet classes and used for the remaining 40 class-conditioned denoising steps. At the switching point, we compute a fixed 40-dimensional subspace spanned by the leading Jacobian directions. The plot shows the norm of the orthogonal projection of the denoising update at each subsequent step onto this fixed subspace, averaged over 100 generated images and 68% confidence interval. The model is the pretrained DiT of Peebles & Xie (2023) on ImageNet.

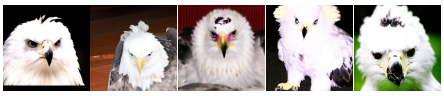
The Jacobian $Jg^*_{t,c}(y)$ also appears to depend on both the conditioning class c and the intermediate image. For the same intermediate image y, the leading directions look dog-like under conditioning on the "Maltese dog" class and eagle-like under conditioning on the "bald eagle" class. At the same time, their spatial organization follows the structure already present in the noisy image. For example, the eigendirections tend to add class-specific details to the body regions that are already partly visible in y. Thus, the directions reflect both the target class and the spatial information carried by the intermediate image.

This observation suggests a simple explanation for what happens during the tail: repeated denoising may progressively emphasize a small number of eigendirections, similar to the behavior of power iteration. To check whether this pattern extends beyond the visual examples in Figure 3, we conduct the same analysis on 100 randomly generated images. The first 60 denoising steps are performed using unconditional denoisers, while the final 40 inference steps are conditioned on a randomly selected ImageNet class. Figure 4 shows that the orthogonal projection of the denoised image onto the subspace spanned by the top 40 (out of 4096) eigenvectors of $Jf^*_{60,c}$ increases as the number of denoising steps grows.

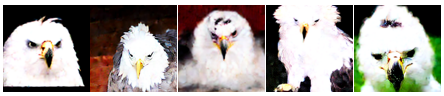
In this section, we study why this behavior arises. We first show that, under mild regularity conditions, the tail is well approximated by the time-frozen composition $(g_{t,c})^{\circ s}$. We then analyze repeated applications of this fixed denoiser and explain how they can favor a small number of eigendirections, much like power iteration. This provides an explanation for why the directions that already appear at the beginning of the tail become increasingly prominent in the final denoised output.

Freezing the denoising time. We first compare the time-varying composition with repeated use of the first denoiser $g_{t,c}$, where $t = t_{m-s+1}$. Starting from the same image y, after $1 \le n \le s$ steps the two procedures give

Standard denoising $\quad x_n = \big(g_{t_{m-s+n},c} \circ \cdots \circ g_{t,c}\big)(y)$



Time-frozen denoising $\quad \bar{x}_n = (g_{t,c})^{\circ n}(y)$



Here $x_0 = \bar{x}_0 = y$, and $(g_{t,c})^{\circ n}$ denotes n applications of the denoiser. Freezing the time turns the scheduled sequence into the power of one denoiser. The next proposition shows that, under the stated predictor bounds, the two trajectories become close late in generation. The accompanying comparison uses 50 inference steps. We freeze the time after 35 completed steps, leaving $s = 15$ denoising updates. The resulting images in the two rows are visually very similar, supporting the time-freezing approximation. The top row shows the outputs from standard denoising, while the bottom row shows the corresponding outputs from time-frozen denoising.

Proposition 4.1. *Start the scheduled and time-frozen compositions from the same image y at time $t = t_{m-s+1} \ge 0$. Under the mild regularity assumptions, for a fixed number of remaining steps,* $\max_{0 \le n \le s} \|x_n - \bar{x}_n\| = O(e^{-t/2})$.

We defer the formal statement and proof of this result to Section C.1. The proof provides an explicit finite-step bound that tracks how changes in the denoiser over the remaining time points accumulate through the subsequent updates.

Repeated denoising as power iteration. We now examine the time-frozen iteration $\bar{x}_{n+1} = g_{t,c}(\bar{x}_n)$. Let $d_n = \bar{x}_{n+1} - \bar{x}_n$ be the change in the image at step n. Two consecutive changes satisfy

$$d_{n+1} = \bar{x}_{n+2} - \bar{x}_{n+1} = g_{t,c}(\bar{x}_{n+1}) - g_{t,c}(\bar{x}_n).$$

When the Jacobian changes little along the trajectory, we can use Taylor's theorem and approximate this difference using its value $Jg_{t,c}(y)$ at the initial image,

$$\begin{aligned} d_{n+1} &\approx Jg_{t,c}(y)(\bar{x}_{n+1} - \bar{x}_n) = Jg_{t,c}(y)d_n, \\ d_n &\approx \big(Jg_{t,c}(y)\big)^n d_0, \qquad d_0 = g_{t,c}(y) - y. \end{aligned} \tag{4.1}$$

Each step approximately applies the same Jacobian to the preceding increment. This gives the power-iteration structure suggested by the images. In Section C.2, we make the power-iteration picture precise. Under a local second-order smoothness condition, the frozen tail is approximated by

$$(g_{t,c})^{\circ s}(y) \approx y + \sum_{n=0}^{s-1} \big(Jg_{t,c}(y)\big)^n \big(g_{t,c}(y) - y\big).$$

This expansion shows that repeated denoising naturally gives rise to successive powers of the Jacobian. For a fixed tail length s, the approximation error is second order in the largest increment, $O\big((\max_{0 \le r < s} \|d_r\|)^2\big)$, with the dependence on s given explicitly in Section C.2. Combining this with the $O(e^{-t/2})$ error between the DDIM tail and the time-frozen denoising from Section 3 and Proposition 4.1 shows that the DDIM tail admits the same Jacobian-power description when initialized from the same intermediate image y.

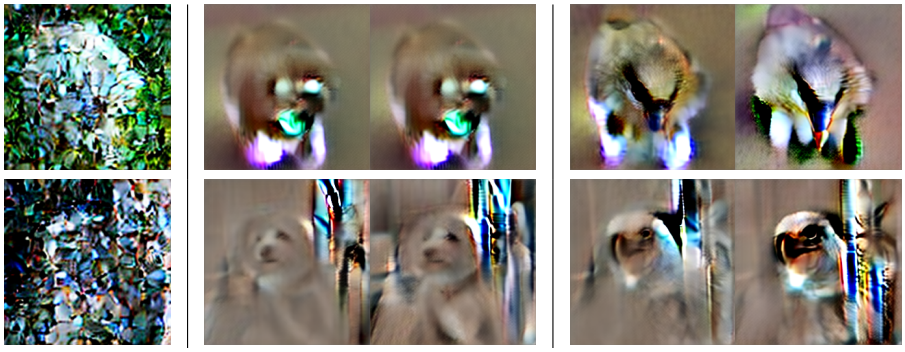


Figure 5: **Basis adaptation to the noisy observation and distribution class.** Each row starts from a different shared noisy image, shown in the first column. The next two columns show leading Jacobian eigendirections under "Maltese dog" conditioning, and the last two show the corresponding directions under "bald eagle" conditioning. Within each row, changing the conditioning class changes the class-specific features while preserving their alignment with the spatial structure of the shared image. Across the two rows, changes in the location and pose of the object are reflected in the corresponding eigendirections adapted to the noisy canvas.

5 Tail Denoising as Conditional PCA

So far, we prove that the tail denoising can be viewed as an approximate power iteration on denoising Jacobian $J_{g_t,c}(y)$. But, how the leading eigenvector of the Jacobian links to the data distribution? When the function class $\mathcal{F}$ is unrestricted, with $w = 0$ and $\mathbb{E}\left\|x^{(c)}\right\|^2 < \infty$, the denoiser is the conditional mean $g^*_{t,c}(y) = \mathbb{E}[x^{(c)} \mid y_t = y]$, where $y_t = \alpha_t x^{(c)} + \sigma_t z$, and its Jacobian, by the covariance identity of Kadkhodaie et al. (2024), satisfies

$$Jg^*_{t,c}(y) = \frac{\alpha_t}{\sigma_t^2} \operatorname{Cov}\left(x^{(c)} \mid y_t = y\right). \tag{5.1}$$

We include a derivation in Section C.3. Since α_t/σ_t^2 is positive, the Jacobian and the conditional covariance have the same eigenvectors, with the same ordering. These eigenvectors are therefore the principal directions of variation among clean images from class c, conditioned on the current noisy image y.

Let $\kappa_1 \geq \cdots \geq \kappa_d \geq 0$ be the eigenvalues of $Jg^*_{t,c}(y)$, with orthonormal eigenvectors $v_1, \ldots, v_d$. For the ideal frozen tail, $d_0 = g^*_{t,c}(y) - y$, and the power-iteration prediction becomes

$$\left(Jg^*_{t,c}(y)\right)^n d_0 = \sum_{\ell=1}^{d} \kappa_\ell^n \langle v_\ell, d_0 \rangle v_\ell.$$

After n stale denoising steps, the component of the initial update along v_ℓ is multiplied by κ_ℓ^n. When the leading eigenvalues are larger than the remaining ones, any nonzero component of d_0 along the corresponding eigendirections becomes increasingly dominant relative to the others. This is the standard power-iteration mechanism behind PCA. In our setting, it explains why a small number of Jacobian directions computed at the beginning of the tail can become more prominent as denoising proceeds as we observed in Figures 4 and 5. Section C.4 makes this statement quantitative and bounds the effect of the local approximation error.

The dependence of the conditional covariance on both y and c gives a concrete version of the canvas-and-brush picture from the introduction. The intermediate image carries the spatial organization that has already formed, while the condition determines how that organization can still be completed.

The intermediate image as the canvas. The basis depends on the particular noisy image at which the Jacobian is evaluated. Since the covariance in (5.1) is conditioned on $y_t = y$, changing y changes the posterior over plausible clean images and therefore changes $\operatorname{Cov}(x^{(c)} \mid y_t = y)$. By the switching time, y already contains coarse spatial information about the image, including the location, pose,

and shape of the main object. The principal directions follow this structure and place class-specific features where they are compatible with the image that has already formed. Figure 5 shows this directly. Within each row, the intermediate image is fixed while the conditioning class changes. The eigendirections become dog-like or eagle-like, but their spatial organization continues to follow the same object. Across rows, changing the location and pose in y causes the eigendirections to move with it. In this sense, the intermediate image acts as the canvas on which the rest of denoising takes place.

Conditional PCA as the brush. The conditioning class determines how this canvas can be completed. For a fixed intermediate image y, changing c changes the class law $p(x \mid c)$ and therefore the conditional covariance in (5.1). Its leading directions describe the dominant class-specific variations that remain plausible around the current image. In Figure 3, the same intermediate image gives dog-like directions under "Maltese dog" conditioning and eagle-like directions under "bald eagle" conditioning. These conditional directions play the role of possible brushstrokes. The remaining denoising steps repeatedly reinforce the dominant ones through the power-iteration mechanism described above. The image sets the canvas, the condition determines the brushstrokes, and denoising gradually turns those directions into visible features of the final image.

Connection to unguided denoising. Biroli et al. (2024) characterize three dynamical regimes in unconditional generation: (1) an initial regime in which the generated sample is not yet faithful to the data distribution; (2) a regime in which the trajectory has selected a data class but still moves within that class; and (3) a late regime in which the trajectory is attracted toward individual training samples, corresponding to memorization. Their analysis connects the transition into the second regime to the leading spectral structure of the data distribution and uses this picture to study when denoising moves from class-level generation toward memorization.

Our analysis focuses on a different question. We start from a particular intermediate image y, after its coarse spatial structure has already formed, and study how the remaining denoising steps respond when the conditioning class is changed. The relevant matrix is then the conditional covariance $\mathrm{Cov}(x^{(c)} \mid y_t = y)$, whose leading directions depend jointly on the target class c and on the specific intermediate image y. These directions therefore describe not only which class-specific features are emphasized, but also how those features are placed relative to the structure already present in the image. In this sense, the two analyses describe different aspects of spectral structure in diffusion dynamics. Biroli et al. (2024) use it to understand the progression from class formation to memorization in unconditional generation, whereas we use it to explain class adaptation in the guided denoising tail of a pretrained diffusion model. Our results show that, after a guidance switch, the relevant eigendirections are selected locally by both the conditioning class and the image being transformed.

6 Mixable Distributions

Denoising switches are effective for image classes that become indistinguishable at a sufficiently high noise level. As Figure 1 shows, images from different classes cannot be told apart beyond a certain noise level. Consequently, the same noisy image can be guided toward the distribution of either dogs or eagles, depending on which class-conditional denoiser is applied. By combining different denoisers on a single noisy image, guided denoising can therefore produce a variety of images resembling different objects.

We validate this hypothesis experimentally. Starting from the source class "golden retriever," we denoise for the first 150 of 200 inference steps and then switch to the denoiser of a target class ("goldfish," "sports car," or "bald eagle") for the remaining steps. To assess whether a switch succeeds, we classify each output with an ImageNet-pretrained ResNet-50 (He et al., 2016). Figure 6 reports the fraction of 100 generated images per target class that are classified as the target. Switching largely succeeds for "bald eagle" and "goldfish" but has a substantially lower success rate for "sports car," since sports cars and golden retrievers remain distinguishable at the noise level where the switch occurs.

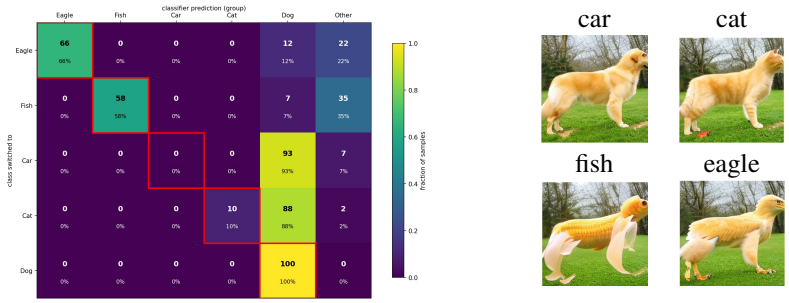


Figure 6: **Switching across classes.** DiT denoising (DDIM) starting from the source "golden retriever" for 150 denoising steps, then switching to target classes ("goldfish," "sports car," or "bald eagle"). *Left:* ResNet-50 classification rate for 100 randomly generated samples for each class. *Right:* Samples of switch outcomes with the shared noisy input y, where sampling from the car class was unsuccessful.

7 Discussions

We analyzed how denoising switches influence creativity, arguing that they can stitch together elements from different class distributions. Our key contribution is a proof that the tail of DDIM denoising is connected to power iteration on the denoising Jacobian, and thereby to principal component analysis (PCA) of the conditional covariance of the data. This connection explains how tail denoising adapts to both the class distribution and the intermediate denoising state.

Our work has several limitations that call for future research. (1) Although we discuss conditions that favor a successful denoising switch, we have not established mathematically when a switch generates images from the target distribution. (2) Our study does not explain the head of the denoising process and its role in creating a canvas that fixes object positions and poses. (3) We did not study the influence of the VAE decoder on denoising. (4) Our analysis treats the model as a black box, leaving aside the inner workings of transformers.

One observation that we found particularly interesting but could not explain is the fine-to-coarse loss of detail under sparsification of the denoising Jacobian. We sparsify the difference between the Jacobians of the eagle and dog denoisers, $Jg^*_{t_s,\text{eagle}}(y) - Jg^*_{t_s,\text{dog}}(y)$, retaining a progressively smaller fraction of its entries. As sparsity increases, the leading eigenvector gradually loses detail, from fine to coarse, as shown in Figure 7.

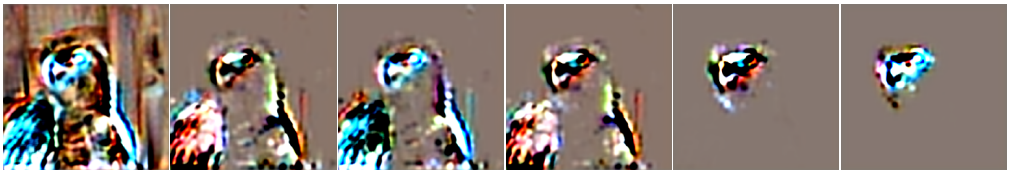


Figure 7: **Sparsification removes detail from fine to coarse.** Leading eigenvector of $\nabla f_{t_s,\text{eagle}}(y) - \nabla f_{t_s,\text{dog}}(y)$ when only the largest-magnitude entries are kept. From left to right, the retained fraction is 1, 0.01, 0.008, 0.004, 0.002, and 0.001.

8 AI Disclosure

Generative AI was used to assist with the experiments reported in Figure 2 and Figure 6; all other experiments were mainly implemented by the authors with AI assistance for leading pretrained models. Generative AI was also used to audit mathematical derivations and to assist with paraphrasing and improving the clarity of the writing. All mathematical arguments, experimental results, and final text were reviewed and verified by the authors.

REFERENCES

Giulio Biroli, Tony Bonnaire, Valentin De Bortoli, and Marc Mézard. Dynamical regimes of diffusion models. *Nature Communications*, 15(1):9957, 2024.

Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Müller, and Björn Ommer. Retrieval-augmented diffusion models. *Advances in Neural Information Processing Systems*, 35:15309–15324, 2022.

Tony Bonnaire, Raphaël Urfin, Giulio Biroli, and Marc Mézard. Why diffusion models don't memorize: The role of implicit dynamical regularization in training. *Advances in Neural Information Processing Systems*, 2025.

Benjamin Boys, Mark Girolami, Jakiw Pidstrigach, Sebastian Reich, Alan Mosca, and Omer Deniz Akyildiz. Tweedie moment projected diffusions for inverse problems. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX security symposium (USENIX Security 23)*, pp. 5253–5270, 2023.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM transactions on Graphics (TOG)*, 42(4):1–10, 2023.

Yujin Choi, Jinseong Park, Hoki Kim, Jaewook Lee, and Saerom Park. Fair sampling in diffusion models through switching mechanism. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 21995–22003, 2024.

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.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Alessandro Favero, Antonio Sclocchi, Francesco Cagnetta, Pascal Frossard, and Matthieu Wyart. How compositional generalization and creativity improve as diffusion models are trained. In *Forty-second International Conference on Machine Learning*, 2025.

Ada Görgün, Fawaz Sammani, Nikos Deligiannis, Bernt Schiele, and Jonas Fischer. Temporal concept dynamics in diffusion models via prompt-conditioned interventions. In *International Conference on Learning Representations*, 2026.

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

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2023.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

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

Zahra Kadkhodaie, Florentin Guth, Eero Simoncelli, and Stéphane Mallat. Generalization in diffusion models arises from geometry-adaptive harmonic representations. In *International Conference on Learning Representations*, 2024.

Mason Kamb and Surya Ganguli. An analytic theory of creativity in convolutional diffusion models. In *Forty-second International Conference on Machine Learning*, 2025.

Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. In *The Eleventh International Conference on Learning Representations*, 2023.

Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. *Advances in Neural Information Processing Systems*, 2024.

Hang Li, Chengzhi Shen, Philip Torr, Volker Tresp, and Jindong Gu. Self-discovering interpretable diffusion latent directions for responsible text-to-image generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12006–12016. IEEE, 2024.

Jun Hao Liew, Hanshu Yan, Daquan Zhou, and Jiashi Feng. Magicmix: Semantic mixing with diffusion models. *arXiv preprint arXiv:2210.16056*, 2022.

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *European conference on computer vision*, pp. 423–439. Springer, 2022.

Lorenzo Olearo, Giorgio Longari, Simone Melzi, Alessandro Raganato, and Rafael Peñaloza. How to blend concepts in diffusion models. *arXiv preprint arXiv:2407.14280*, 2024.

Lorenzo Olearo, Giorgio Longari, Alessandro Raganato, Rafael Peñaloza, and Simone Melzi. Blending concepts with text-to-image diffusion models. *Discover Artificial Intelligence*, 2026.

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

Manos Plitsis, Theodoros Kouzelis, Panagiotis Koromilas, Vassilis Katsouros, Mihalis A. Nicolaou, and Yannis Panagakis. Disentangling local and global semantics in diffusion models for image editing. *International Journal of Computer Vision*, 134(4):160, 2026.

Severi Rissanen, Markus Heinonen, and Arno Solin. Free hunch: Denoiser covariance estimation for diffusion models without extra costs. In *International Conference on Learning Representations*, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2022.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6048–6058. IEEE, 2023.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a.

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*, 2021b.

Berk Tinaz, Zalan Fabian, and Mahdi Soltanolkotabi. Emergence and evolution of interpretable concepts in diffusion models. *Advances in neural information processing systems*, 2025.

Enrico Ventura, Beatrice Achilli, Gianluigi Silvestri, Carlo Lucibello, and Luca Ambrogioni. Manifolds, random matrices and spectral gaps: The geometric phases of generative diffusion. In *International Conference on Learning Representations*, 2025.

Binxu Wang and John J Vastola. Diffusion models generate images like painters: an analytical theory of outline first, details later. *arXiv preprint arXiv:2303.02490*, 2023.

Qingsong Wang, Zhengchao Wan, Misha Belkin, and Yusu Wang. Seeds of structure: Patch pca reveals universal compositional cues in diffusion models. *Advances in Neural Information Processing Systems*, 2025.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

A Additional Related Work

One line of work studies generalization and memorization in diffusion models, asking why generated samples need not reproduce the training data. Empirical studies have shown that diffusion models can reproduce individual training examples under some training regimes (Somepalli et al., 2023; Carlini et al., 2023). More recent theory has sought to explain when memorization occurs and what allows generalization. Biroli et al. (2024) show that, under the exact score of the empirical distribution, reverse diffusion eventually becomes attracted toward individual training samples. In contrast, Kadkhodaie et al. (2024) find that, with sufficiently large datasets, networks trained on disjoint subsets learn similar score functions, suggesting that the learned score becomes less tied to the particular empirical sample. Bonnaire et al. (2025) identify a related effect over training time, where useful generation emerges before memorization and the gap between these two timescales grows with dataset size. Kamb & Ganguli (2025) provide a different explanation based on architectural inductive bias. They show that locality and equivariance in convolutional models can prevent exact fitting of the empirical score and allow new images to arise through the recombination of local structure. These works clarify when diffusion models reproduce their training data and when they generalize beyond it. Our focus is instead on creativity. We ask how guidance can take structure that has already formed and turn it into a new semantic composition.

Another line of work studies compositional generation, where diffusion models combine familiar concepts in new ways. Liu et al. (2022) show that models associated with different concepts can be composed to generate new combinations of objects, attributes, and relations. Liew et al. (2022) use the denoising trajectory itself, preserving structure formed early in sampling while introducing the semantics of another concept later on. Favero et al. (2025) study how compositional generalization improves with more data and longer training. Olearo et al. (2024) study a complementary setting at inference time, including methods that switch the prompt during denoising. Together, these works show that diffusion models can combine learned concepts beyond the examples seen during training. We focus on a different part of the question, namely how guidance shapes such combinations after the image has already begun to form.

Conditional diffusion models are commonly controlled through classifier guidance and classifier-free guidance (Dhariwal & Nichol, 2021; Ho & Salimans, 2021). Several works have shown that intervening during sampling can change the semantic content of the generated image. Hertz et al. (2023) manipulate cross-attention to edit a prompt while preserving much of the spatial layout, while Chefer et al. (2023) modify attention to strengthen concepts that might otherwise be missed. A more direct intervention is to change the condition itself during denoising. Liew et al. (2022) use such a switch to mix semantic concepts, and Choi et al. (2024) use attribute switching to control fairness at inference time. Olearo et al. (2026) study prompt switching as a method for zero-shot concept blending, while Görgün et al. (2026) use condition changes at different timesteps to study when a concept can still affect the generated image. The effect of guidance also depends on where it is applied along the trajectory (Kynkäänniemi et al., 2024). We build on this idea, not by proposing a new switching mechanism, but by using the switch as a probe to study creativity in diffusion models.

Another line of work studies how different parts of the denoising trajectory play different roles. Wang & Vastola (2023) describe a coarse-to-fine pattern in which global layout emerges before finer image details. Tinaz et al. (2025) study how interpretable concepts evolve during generation and find that different kinds of interventions are effective at different stages. Along similar lines, Görgün et al. (2026) show that the ability to introduce a new concept depends on when the condition is changed. Wang et al. (2025) take a complementary view and show that the initial noise already contains low-frequency cues related to the final composition. These results suggest that the denoising trajectory is not uniform in time. Our work focuses on its later stages, when much of the spatial structure has already formed but guidance can still shape the semantic content of the final image.

Our analysis is also related to work on the geometry of denoisers and score functions. Kadkhodaie et al. (2024) study image-adaptive bases learned by denoisers and relate them to the local geometry of natural images. Related work uses higher-order forms of Tweedie's identity to connect denoising with the conditional covariance of the clean image given a noisy observation (Boys et al., 2024; Rissanen et al., 2025), mainly for posterior and conditional sampling. A complementary line of work studies spectral and semantic directions in diffusion models. Ventura et al. (2025) analyze the spectrum of the score Jacobian and use spectral gaps to describe different geometric regimes of diffusion. Kwon

et al. (2023) identify semantic directions in internal diffusion representations, while Li et al. (2024) develop a self-supervised method for discovering such directions. More closely related to our setting, Plitsis et al. (2026) use the denoising-network Jacobian to construct semantic subspaces for image editing. Wang et al. (2025) use patch PCA to find compositional structure already present in the initial noise.

Our setting differs in a key respect. Rather than only identifying semantic directions, we characterize the geometry that is active at a particular intermediate image and show how it governs the rest of denoising. This geometry depends jointly on the intermediate image and the target class. We identify the leading directions of the denoiser Jacobian with the principal components of $\mathrm{Cov}(x^{(c)} \mid y_t = y)$ and show that repeated denoising progressively amplifies these directions along the tail. In this way, the conditional geometry at the switching point directly predicts how class-specific structure emerges in the final image.

B Approximating tail of DDIM by Denoiser Compositions

We compare the final s steps of DDIM with the composition of denoiser estimators introduced in Section 3. Replacing the tail of a DDIM update by its clean estimate introduces an error that becomes small as the noise decreases. We first bound the predictions and then track how these errors accumulate over the remaining steps. All assumptions below are stated in terms of the noise predictor f. We derive its relation to g, which later provides the bounds used in the proof.

B.1 The two tails and the assumptions

Fix a class c and a tail length $1 \le s \le m$, starting at $t = t_{m-s+1}$. Let n count the completed tail steps. At step n, the predictor is evaluated at t_j, where $j = m - s + n + 1$. The composition states x_n and DDIM states u_n satisfy

$$\begin{aligned} x_{n+1} &= g_{t_j,c}(x_n), \\ u_{n+1} &= \alpha_{t_{j+1}} g_{t_j,c}(u_n) + \sigma_{t_{j+1}} f_{t_j,c}(u_n), \end{aligned} \tag{B.1}$$

for $n = 0, \dots, s-1$. Both procedures use the same predictor, class, and time schedule, with $w = 0$. Their outputs are x_s and u_s. The approximation result starts both tails from the supplied switching image, $u_0 = x_0 = y$. For $c = c_i$, the composition output is $x_s = G_s^{(i)}(y)$. The general comparison also allows different initial images. All vector norms below are Euclidean.

Assumption B.1 (Lipschitzness of noise predictor)**.** At each finite time v, the noise predictor satisfies

$$\|f_{v,c}(x) - f_{v,c}(y)\| \le K_X(v)\,\|x - y\|$$

on the input pairs. The constant $K_X(v)$ may depend on the class and these pairs, as well as on v.

For the present comparison, the required pairs are (u_n, x_n) at time t_j.

Recall that g is obtained from f by subtracting the predicted noise. Our schedule gives

$$\begin{aligned} \frac{1}{\alpha_v} &= 1 + e^{-v}, \\ \frac{\sigma_v}{\alpha_v} &= \sqrt{\frac{1-\alpha_v^2}{\alpha_v^2}} = \sqrt{(1+e^{-v})^2 - 1} = \sqrt{2e^{-v} + e^{-2v}}. \end{aligned}$$

Hence

$$\begin{aligned} g_{v,c}(x) &= \frac{x - \sigma_v f_{v,c}(x)}{\alpha_v} \\ &= (1 + e^{-v})x - \sqrt{2e^{-v} + e^{-2v}}\, f_{v,c}(x). \end{aligned} \tag{B.2}$$

Subtracting the estimates at two inputs and applying Assumption B.1, we obtain

$$\|g_{v,c}(x) - g_{v,c}(y)\| = \frac{1}{\alpha_v} \left\| x - y - \sigma_v \big(f_{v,c}(x) - f_{v,c}(y)\big) \right\|$$

$$\leq \frac{\|x-y\| + \sigma_v \|f_{v,c}(x) - f_{v,c}(y)\|}{\alpha_v}$$
$$\leq \frac{1+\sigma_v K_X(v)}{\alpha_v} \|x-y\|.$$

We therefore use

$$L_X(v) := \frac{1+\sigma_v K_X(v)}{\alpha_v}, \qquad \|g_{v,c}(x) - g_{v,c}(y)\| \leq L_X(v)\|x-y\|. \tag{B.3}$$

Let $L = \max_{j=m-s+1,\dots,m} L_X(t_j)$. This bounds how much a clean estimator can enlarge a difference between the two states. In particular, $L \geq 1$.

The approximation rate also uses a bound on the prediction values along the DDIM trajectory.

Assumption B.2. Assume Assumption B.1 on the pairs (u_n, x_n) and suppose

$$\|u_0\| \leq R, \qquad \left\|f_{t_j,c}(u_n)\right\| \leq M_f, \qquad j = m-s+n+1, \quad 0 \leq n < s.$$

As the starting time t increases, R, M_f, and L remain bounded independently of t. The tail length s is fixed unless stated otherwise.

Since $L_X(v) = \alpha_v^{-1} + (\sigma_v/\alpha_v)K_X(v)$, this allows $K_X(v)$ to grow with v as long as its scaled contribution stays bounded.

We next obtain a bound on the clean estimates along DDIM. Substitute the definition of g into the DDIM update and divide by $\alpha_{t_{j+1}}$ to get

$$\begin{aligned}\frac{u_{n+1}}{\alpha_{t_{j+1}}} &= g_{t_j,c}(u_n) + \frac{\sigma_{t_{j+1}}}{\alpha_{t_{j+1}}} f_{t_j,c}(u_n) \\ &= \frac{u_n}{\alpha_{t_j}} - \frac{\sigma_{t_j}}{\alpha_{t_j}} f_{t_j,c}(u_n) + \frac{\sigma_{t_{j+1}}}{\alpha_{t_{j+1}}} f_{t_j,c}(u_n) \\ &= \frac{u_n}{\alpha_{t_j}} + \left(\frac{\sigma_{t_{j+1}}}{\alpha_{t_{j+1}}} - \frac{\sigma_{t_j}}{\alpha_{t_j}}\right) f_{t_j,c}(u_n).\end{aligned}$$

The ratio $\sigma_v/\alpha_v = \sqrt{2e^{-v} + e^{-2v}}$ decreases as v increases. Taking norms therefore gives

$$\left\|\frac{u_{n+1}}{\alpha_{t_{j+1}}}\right\| \leq \left\|\frac{u_n}{\alpha_{t_j}}\right\| + \left(\frac{\sigma_{t_j}}{\alpha_{t_j}} - \frac{\sigma_{t_{j+1}}}{\alpha_{t_{j+1}}}\right) M_f.$$

Apply this inequality successively from u_0 to u_n. For $j = m-s+n+1$, the resulting sum is

$$\begin{aligned}\left\|\frac{u_n}{\alpha_{t_j}}\right\| &\leq \frac{R}{\alpha_t} + M_f \sum_{k=m-s+1}^{j-1} \left(\frac{\sigma_{t_k}}{\alpha_{t_k}} - \frac{\sigma_{t_{k+1}}}{\alpha_{t_{k+1}}}\right) \\ &= \frac{R}{\alpha_t} + \left(\frac{\sigma_t}{\alpha_t} - \frac{\sigma_{t_j}}{\alpha_{t_j}}\right) M_f.\end{aligned}$$

The equality follows because consecutive terms in the sum cancel. This state bound holds for $0 \leq n \leq s$, including the endpoint $j = m+1$. For $0 \leq n < s$, it also bounds the clean estimate as follows,

$$\begin{aligned}\left\|g_{t_j,c}(u_n)\right\| &\leq \left\|\frac{u_n}{\alpha_{t_j}}\right\| + \frac{\sigma_{t_j}}{\alpha_{t_j}} M_f \\ &\leq \frac{R}{\alpha_t} + \left(\frac{\sigma_t}{\alpha_t} - \frac{\sigma_{t_j}}{\alpha_{t_j}}\right) M_f + \frac{\sigma_{t_j}}{\alpha_{t_j}} M_f \\ &= \frac{R + \sigma_t M_f}{\alpha_t} =: M_g(t).\end{aligned} \tag{B.4}$$

Thus the initial-state and noise-prediction bounds give

$$M_g(t) = (1+e^{-t})R + \sqrt{2e^{-t} + e^{-2t}}\, M_f.$$

The state bound above also gives $\|u_n\| \leq M_g(t)$, since $\alpha_{t_j} \leq 1$.

B.2 Accumulation of the one-step error

Lemma B.3 (Comparison of composition tail with DDIM). *Assume Assumption B.1 on the pairs (u_n, x_n). If the difference between a DDIM step and the clean estimate at the same input has norm at most ε at every step, then*

$$\max_{0\le n\le s} \|u_n - x_n\| = O(\|u_0 - x_0\| + \varepsilon)\,.$$

The constant depends only on s and L.

Proof. At step n, with $j = m - s + n + 1$, write the one-step difference as

$$\delta_n := u_{n+1} - g_{t_j,c}(u_n).$$

Subtracting $g_{t_j,c}(u_n)$ from the DDIM update gives

$$\delta_n = (\alpha_{t_{j+1}} - 1) g_{t_j,c}(u_n) + \sigma_{t_{j+1}} f_{t_j,c}(u_n). \tag{B.5}$$

Since $x_{n+1} = g_{t_j,c}(x_n)$, the difference between the next states is

$$u_{n+1} - x_{n+1} = g_{t_j,c}(u_n) - g_{t_j,c}(x_n) + \delta_n.$$

Taking norms and using (B.3), we obtain

$$\begin{aligned}\|u_{n+1} - x_{n+1}\| &\le \left\|g_{t_j,c}(u_n) - g_{t_j,c}(x_n)\right\| + \|\delta_n\| \\ &\le L\,\|u_n - x_n\| + \|\delta_n\|\,.\end{aligned}$$

Each step therefore propagates the existing difference and adds the error from the current replacement. Iterating this inequality gives

$$\|u_n - x_n\| \le L^n \|u_0 - x_0\| + \sum_{r=0}^{n-1} L^{n-1-r} \|\delta_r\|\,, \qquad 0 \le n \le s. \tag{B.6}$$

To verify the induction, the case $n = 0$ is equality, with an empty sum. If the bound holds at n, the next step satisfies

$$\begin{aligned}\|u_{n+1} - x_{n+1}\| &\le L\left(L^n \|u_0 - x_0\| + \sum_{r=0}^{n-1} L^{n-1-r} \|\delta_r\|\right) + \|\delta_n\| \\ &= L^{n+1} \|u_0 - x_0\| + \sum_{r=0}^{n} L^{n-r} \|\delta_r\|\,.\end{aligned}$$

This is the same bound at $n + 1$. Finally, substituting $\|\delta_r\| \le \varepsilon$ and using $L \ge 1$ gives

$$\max_{0\le n\le s} \|u_n - x_n\| \le L^s \|u_0 - x_0\| + \varepsilon \sum_{r=0}^{s-1} L^r.$$

For fixed s and L, this proves the claimed order. □

B.3 The rate for our noise schedule

Corollary B.4 (Tail approximation). *Under Assumption B.2, start both tails from the same image, $u_0 = x_0 = y$. For fixed s and uniform bounds R, M_f, L,*

$$\max_{0\le n\le s} \|u_n - x_n\| = O(e^{-t/2}), \qquad t = t_{m-s+1} \to +\infty.$$

For $s = 1$, their outputs agree exactly.

Proof. We first calculate the size of the schedule coefficients. Since $\alpha_v = (1 + e^{-v})^{-1}$,

$$1 - \alpha_v = \frac{e^{-v}}{1 + e^{-v}},$$

$$\sigma_v^2 = 1 - \frac{1}{(1+e^{-v})^2} = \frac{2e^{-v} + e^{-2v}}{(1+e^{-v})^2}$$
$$= e^{-v}\frac{2+e^{-v}}{(1+e^{-v})^2} \le 2e^{-v}.$$

The last inequality uses $2 + e^{-v} \le 2(1+e^{-v})^2$. Taking square roots gives

$$1 - \alpha_v \le e^{-v},$$
$$\sigma_v \le \sqrt{2}\, e^{-v/2}.$$

We now bound the one-step error using (B.5) and the prediction bounds. For each nonfinal step, with $j = m - s + r + 1$,

$$\begin{aligned}\|\delta_r\| &\le (1-\alpha_{t_{j+1}})\left\|g_{t_j,c}(u_r)\right\| + \sigma_{t_{j+1}}\left\|f_{t_j,c}(u_r)\right\| \\ &\le M_g(t)(1-\alpha_{t_{j+1}}) + M_f\sigma_{t_{j+1}} \\ &\le M_g(t)e^{-t} + \sqrt{2}M_f e^{-t/2},\end{aligned}$$

where the last inequality follows from $t_{j+1} \ge t$. At the final step, $\alpha_{t_{m+1}} = 1$ and $\sigma_{t_{m+1}} = 0$, so $\delta_{s-1} = 0$.

Thus every step satisfies the error bound in Lemma B.3 with $\varepsilon = M_g(t)e^{-t} + \sqrt{2}M_f e^{-t/2}$. The expression for $M_g(t)$ in (B.4) remains bounded as $t \to +\infty$, so $\varepsilon = O(e^{-t/2})$. Since $u_0 = x_0$, applying the lemma with fixed s and uniformly bounded L gives

$$\max_{0\le n\le s}\|u_n - x_n\| = O(e^{-t/2}).$$

For later use, we also record the explicit bounds supplied by the lemma. Substituting the schedule-dependent estimate for $\|\delta_r\|$ into (B.6) gives

$$\begin{aligned}\|u_n - x_n\| \le M_g(t)\sum_{r=0}^{n-1} L^{n-1-r}\left(1-\alpha_{t_{m-s+r+2}}\right) \\ + M_f\sum_{r=0}^{n-1} L^{n-1-r}\sigma_{t_{m-s+r+2}}, \qquad 0 \le n \le s.\end{aligned}$$

For $s \ge 2$, using $\delta_{s-1} = 0$ yields

$$\begin{aligned}\|u_s - x_s\| &\le \sum_{r=0}^{s-2} L^{s-1-r}\|\delta_r\| \\ &\le \left(M_g(t)e^{-t} + \sqrt{2}M_f e^{-t/2}\right)\sum_{r=1}^{s-1} L^r. \qquad \text{(B.7)}\end{aligned}$$

For $s = 1$, both outputs equal $g_{t_m,c}(y)$, so they agree exactly. □

The final DDIM update is exactly a denoiser estimate. Any difference between the two final outputs comes from the inputs produced by earlier steps. The powers of L in (B.7) account for how those earlier errors pass through the subsequent estimators.

C Power Iteration and PCA in the Denoising Tail

We follow the explanation in Section 4. First, we compare the scheduled composition with repeated use of the denoiser at the initial time. We then show that its increments approximately follow powers of $Jg_{t,c}(y)$. For an ideal predictor, identifying this Jacobian with a conditional covariance gives the principal-component interpretation. The proofs below state the approximation errors and their assumptions.

C.1 FREEZING THE TIME ARGUMENT

Put $t = t_{m-s+1}$. The two compositions in Section 4 start from $x_0 = \bar{x}_0 = y$ and satisfy

$$\begin{aligned} x_{n+1} &= g_{t_{m-s+n+1},c}(x_n), \\ \bar{x}_{n+1} &= g_{t,c}(\bar{x}_n), \qquad 0 \le n < s. \end{aligned} \tag{C.1}$$

We use the input bound on f from Assumption B.1, and add bounds on its values and time variation at the inputs used below.

Assumption C.1. Assume $t \ge 0$ and Assumption B.1 on $(x_n, \bar{x}_n)$ at the frozen time t. On the fixed inputs x_n used to compare times $u, v \in [t, t_m]$, suppose

$$\begin{aligned} \|x_n\| \le R, \qquad \|f_{v,c}(x_n)\| \le M_f, \\ \|f_{v,c}(x_n) - f_{u,c}(x_n)\| \le K_T|v-u|. \end{aligned} \tag{C.2}$$

The adjacent finite input times have gaps at most h. For the asymptotic rate, R, M_f, K_T, h and the derived input bound $L = \max_{j=m-s+1,\dots,m} L_X(t_j)$ are uniform in t. The tail length s is fixed.

Bounds at the compared inputs suffice. The estimate for g follows by subtracting the two instances of (B.2). For $v \ge u \ge t$,

$$\begin{aligned} g_{v,c}(x) - g_{u,c}(x) = (e^{-v} - e^{-u})x \\ - \big(\sqrt{2e^{-v} + e^{-2v}} - \sqrt{2e^{-u} + e^{-2u}}\big) f_{v,c}(x) \\ - \sqrt{2e^{-u} + e^{-2u}}\big(f_{v,c}(x) - f_{u,c}(x)\big). \end{aligned}$$

The coefficient derivatives satisfy

$$\begin{aligned} \left|\frac{d}{dv} e^{-v}\right| &= e^{-v}, \\ \left|\frac{d}{dv}\sqrt{2e^{-v} + e^{-2v}}\right| &= \frac{e^{-v} + e^{-2v}}{\sqrt{2e^{-v} + e^{-2v}}} \le \sqrt{2e^{-v} + e^{-2v}}. \end{aligned}$$

The mean-value theorem therefore gives the derived time bound

$$D(t) := Re^{-t} + \sqrt{2e^{-t} + e^{-2t}}\,(M_f + K_T),$$

with

$$\begin{aligned} \|g_{v,c}(x) - g_{u,c}(x)\| &\le D(t)|v-u|, \\ \|g_{t,c}(x_n) - g_{t,c}(\bar{x}_n)\| &\le L_X(t)\,\|x_n - \bar{x}_n\|. \end{aligned}$$

The second inequality is (B.3).

Proposition 4.1 (Time freezing)**.** *Start the two compositions in* (C.1) *from the same image* y *at* $t = t_{m-s+1} \ge 0$*. Under Assumption C.1, for a fixed number of remaining steps,*

$$\max_{0 \le n \le s} \|x_n - \bar{x}_n\| = O(e^{-t/2}).$$

Proof. At step n, add and subtract $g_{t,c}(x_n)$ to separate the change of time from the change of input. This gives

$$\begin{aligned} x_{n+1} - \bar{x}_{n+1} = g_{t_{m-s+n+1},c}(x_n) - g_{t,c}(x_n) \\ + g_{t,c}(x_n) - g_{t,c}(\bar{x}_n), \\ \|x_{n+1} - \bar{x}_{n+1}\| \le D(t)(t_{m-s+n+1} - t) + L\,\|x_n - \bar{x}_n\|. \end{aligned}$$

Iterating from $x_0 = \bar{x}_0$ yields

$$\|x_n - \bar{x}_n\| \le D(t) \sum_{r=0}^{n-1} L^{n-1-r}(t_{m-s+r+1} - t), \qquad 0 \le n \le s.$$

An empty sum is zero. Since $t_{m-s+r+1} - t \le rh$, this also gives

$$\begin{aligned}\|x_n - \bar{x}_n\| &\le hD(t) \sum_{r=0}^{n-1} rL^{n-1-r} \\ &\le \frac{hn(n-1)}{2} D(t) L^{n-1}, \qquad n \ge 1,\end{aligned}$$

where $L \ge 1$ follows from its definition. Finally, for $t \ge 0$,

$$D(t) \le Re^{-t} + \sqrt{3}(M_f + K_T)e^{-t/2} = O(e^{-t/2}).$$

The remaining factors are uniformly bounded for a fixed tail length. This proves the maximum bound. For $s = 1$, both outputs are $g_{t,c}(y)$. □

C.2 How a frozen denoiser produces matrix powers

Keep the time $t = t_{m-s+1}$ fixed and write

$$\bar{x}_0 = y, \qquad \bar{x}_{n+1} = g_{t,c}(\bar{x}_n), \qquad d_n = \bar{x}_{n+1} - \bar{x}_n.$$

In particular, $d_0 = g_{t,c}(y) - y$. Differentiating the clean-estimator definition gives

$$Jg_{v,c}(x) = \frac{1}{\alpha_v} I - \frac{\sigma_v}{\alpha_v} Jf_{v,c}(x). \tag{C.3}$$

We use a standard local smoothness condition to control the Taylor error.

Assumption C.2 (Lipschitz Jacobian)**.** For the time t and class c under consideration, the Jacobian $Jf_{t,c}$ is globally Lipschitz in its input. That is, there exists a finite constant H such that

$$\|Jf_{t,c}(x) - Jf_{t,c}(x')\| \le H \|x - x'\|$$

for all $x, x' \in \mathbb{R}^d$.

A sufficient condition for Assumption C.2 is that $f_{t,c}$ is twice differentiable with a uniformly bounded second derivative. By (C.3),

$$\|Jg_{t,c}(x) - Jg_{t,c}(x')\| \le \frac{\sigma_t}{\alpha_t} H \|x - x'\|. \tag{C.4}$$

In particular, taking $x' = y$ gives

$$\|Jg_{t,c}(x)\| \le \|Jg_{t,c}(y)\| + \frac{\sigma_t}{\alpha_t} H \|x - y\|.$$

Hence $Jg_{t,c}$ is also bounded.

Theorem C.3. *Fix a class c, a tail length $s \ge 1$, and $t = t_{m-s+1}$. Let $\bar{x}_0 = y$, let $\bar{x}_{n+1} = g_{t,c}(\bar{x}_n)$, and write $d_n = \bar{x}_{n+1} - \bar{x}_n$. Under the local second-order smoothness condition in Assumption C.2, for a fixed tail length,*

$$\left\| d_n - \big(Jg_{t,c}(y)\big)^n d_0 \right\| = O\left(\left(\max_{0 \le r < s} \|d_r\| \right)^2 \right), \qquad 0 \le n < s.$$

Moreover,

$$\left\| \bar{x}_s - y - \sum_{n=0}^{s-1} \big(Jg_{t,c}(y)\big)^n d_0 \right\| = O\left(\left(\max_{0 \le r < s} \|d_r\| \right)^2 \right).$$

The hidden constants depend on the fixed tail length, the local smoothness bound, and a bound on the size of the Jacobian $\|Jg_{t,c}(y)\|$.

Proof. Taylor's theorem in integral form, applied on the segment from $\bar{x}_n$ to $\bar{x}_{n+1} = \bar{x}_n + d_n$, gives

$$d_{n+1} = g_{t,c}(\bar{x}_{n+1}) - g_{t,c}(\bar{x}_n)$$

$$= \int_0^1 J g_{t,c}(\bar{x}_n + \theta d_n) d_n \, d\theta.$$

Adding and subtracting $J g_{t,c}(y) d_n$ yields

$$\begin{aligned} d_{n+1} &= J g_{t,c}(y) d_n + e_n, \\ e_n &= \int_0^1 \left[J g_{t,c}(\bar{x}_n + \theta d_n) - J g_{t,c}(y) \right] d_n \, d\theta. \end{aligned} \tag{C.5}$$

By (C.4),

$$\begin{aligned} \|e_n\| &\le \frac{\sigma_t}{\alpha_t} H \int_0^1 \|\bar{x}_n + \theta d_n - y\| \, \|d_n\| \, d\theta \\ &\le \frac{\sigma_t}{\alpha_t} H \left(\|\bar{x}_n - y\| + \frac{1}{2} \|d_n\| \right) \|d_n\| . \end{aligned}$$

Since

$$\bar{x}_n - y = \sum_{q=0}^{n-1} d_q,$$

we have

$$\|\bar{x}_n - y\| \le n \max_{0 \le q < s} \|d_q\| .$$

Consequently,

$$\|e_n\| \le \frac{\sigma_t}{\alpha_t} H \left(n + \frac{1}{2} \right) \left(\max_{0 \le q < s} \|d_q\| \right)^2 . \tag{C.6}$$

Iterating the recurrence in (C.5) gives

$$d_n = \left(J g_{t,c}(y) \right)^n d_0 + \sum_{r=0}^{n-1} \left(J g_{t,c}(y) \right)^{n-1-r} e_r. \tag{C.7}$$

Combining this identity with (C.6) yields

$$\left\| d_n - \left(J g_{t,c}(y) \right)^n d_0 \right\| \le \frac{\sigma_t}{\alpha_t} H \left(\max_{0 \le q < s} \|d_q\| \right)^2 \times \sum_{r=0}^{n-1} \left(r + \frac{1}{2} \right) \|J g_{t,c}(y)\|^{n-1-r} .$$

Finally, using $\bar{x}_s - y = \sum_{n=0}^{s-1} d_n$, we obtain

$$\begin{aligned} \left\| \bar{x}_s - y - \sum_{n=0}^{s-1} \left(J g_{t,c}(y) \right)^n d_0 \right\| &\le \frac{\sigma_t}{\alpha_t} H \left(\max_{0 \le q < s} \|d_q\| \right)^2 \\ &\quad \times \sum_{n=0}^{s-1} \sum_{r=0}^{n-1} \left(r + \frac{1}{2} \right) \|J g_{t,c}(y)\|^{n-1-r} . \end{aligned} \tag{C.8}$$

For fixed s and under the stated smoothness and Jacobian bounds, the sums on the right are bounded independently of the increment size. Both approximation errors are therefore

$$O\left(\left(\max_{0 \le q < s} \|d_q\| \right)^2 \right).$$

□

Returning to the scheduled and DDIM tails. The frozen and scheduled tails start from the same image and have the same first increment d_0. For the scheduled output, the triangle inequality gives

$$\left\| x_s - y - \sum_{n=0}^{s-1} (Jg_{t,c}(y))^n d_0 \right\| \le \|x_s - \bar{x}_s\| + \left\| \bar{x}_s - y - \sum_{n=0}^{s-1} (Jg_{t,c}(y))^n d_0 \right\|.$$

The first term is the time-freezing error from Proposition 4.1; the second is the Taylor error controlled by Theorem C.3. Hence, under their combined assumptions,

$$\left\| x_s - y - \sum_{n=0}^{s-1} (Jg_{t,c}(y))^n d_0 \right\| = O\left(\left(\max_{0 \le r < s} \|d_r\| \right)^2 + e^{-t/2} \right).$$

Now let u_s be the DDIM output from the same starting image y, using the same class and remaining schedule as x_s. A second application of the triangle inequality gives

$$\begin{aligned} \left\| u_s - y - \sum_{n=0}^{s-1} (Jg_{t,c}(y))^n d_0 \right\| \le \|u_s - x_s\| + \|x_s - \bar{x}_s\| \\ + \left\| \bar{x}_s - y - \sum_{n=0}^{s-1} (Jg_{t,c}(y))^n d_0 \right\|. \end{aligned}$$

By Corollary B.4, the first term is $O(e^{-t/2})$ for a fixed-length tail. By Proposition 4.1, the second term has the same order. The third term is second order in the largest frozen-tail increment by Theorem C.3. Therefore,

$$\left\| u_s - y - \sum_{n=0}^{s-1} (Jg_{t,c}(y))^n d_0 \right\| = O\left(\left(\max_{0 \le r < s} \|d_r\| \right)^2 + e^{-t/2} \right).$$

The $O(e^{-t/2})$ term accounts for the DDIM-to-composition and time-freezing errors in the late tail of denoising. Thus, for a fixed number of remaining steps, DDIM admits the same local Jacobian-power description up to this error and the Taylor remainder.

C.3 The denoiser Jacobian as a conditional covariance

This subsection identifies the matrix that appears in the local power-iteration approximation of (4.1). Suppose f^* is the unrestricted population minimizer and $\mathbb{E}\left\|x^{(c)}\right\|^2 < \infty$. Then

$$g^*_{t,c}(y) = \mathbb{E}[x^{(c)} \mid y_t = y], \qquad y_t = \alpha_t x^{(c)} + \sigma_t z.$$

Thus, the ideal denoiser estimator is the conditional mean of the clean image given the noisy observation y. We show that the Jacobian of this conditional mean is a scaled conditional covariance.

Under the Gaussian noise model, Bayes' rule gives the following expression for the conditional mean

$$g^*_{t,c}(y) = \frac{\mathbb{E}\left[x^{(c)} \exp\left(-\frac{\|y - \alpha_t x^{(c)}\|^2}{2\sigma_t^2} \right) \right]}{\mathbb{E}\left[\exp\left(-\frac{\|y - \alpha_t x^{(c)}\|^2}{2\sigma_t^2} \right) \right]}.$$

This is a weighted average over clean images from class c. Images for which $\alpha_t x^{(c)}$ is closer to the observed image y receive larger weight. We differentiate this expression with respect to y in an arbitrary direction $v \in \mathbb{R}^d$. The quotient rule gives

$$\begin{aligned} Jg^*_{t,c}(y)v &= \frac{\alpha_t}{\sigma_t^2} \left(\mathbb{E}\left[x^{(c)} \langle x^{(c)}, v \rangle \mid y_t = y \right] - \mathbb{E}[x^{(c)} \mid y_t = y] \left\langle \mathbb{E}[x^{(c)} \mid y_t = y], v \right\rangle \right) \\ &= \frac{\alpha_t}{\sigma_t^2} \operatorname{Cov}\left(x^{(c)} \mid y_t = y \right) v. \end{aligned}$$

The expression in parentheses is the conditional covariance applied to v. Since this holds for every v,

$$Jg^*_{t,c}(y) = J_y \mathbb{E}[x^{(c)} \mid y_t = y] = \frac{\alpha_t}{\sigma_t^2} \operatorname{Cov}\left(x^{(c)} \mid y_t = y \right). \tag{C.9}$$

This proves (5.1). At every finite t, the Gaussian factor allows differentiation under the expectation under the stated second-moment condition, including when the class distribution is discrete.

Since $\alpha_t/\sigma_t^2 > 0$, the ideal denoiser Jacobian is symmetric positive semidefinite and has the same eigenvectors, with the same ordering, as the conditional covariance. Its leading eigendirections are therefore the principal directions of variation of clean images from class c conditioned on the noisy observation y. The class c determines the class distribution, while y determines the conditional distribution within that class. The resulting eigendirections therefore depend on both the conditioning class and the intermediate image.

Tweedie's identity gives

$$I + \sigma_t^2 \nabla^2 \log q_{t,c}(y) = \frac{\alpha_t^2}{\sigma_t^2} \operatorname{Cov}\Big(x^{(c)} \mid y_t = y\Big). \tag{C.10}$$

Equivalently,

$$g^*_{t,c}(y) = \nabla_y \left(\frac{\|y\|^2}{2\alpha_t} + \frac{\sigma_t^2}{\alpha_t} \log q_{t,c}(y) \right).$$

The Hessian of this potential is $Jg^*_{t,c}(y) \succeq 0$, so the potential is convex. This establishes that single denoiser estimator is a gradient of a convex function.

Relation to the noise predictor and a DDIM step. For any differentiable predictor, (C.3) is exact. In the ideal case, combining it with (C.9) gives

$$Jf^*_{t,c}(y) = \frac{1}{\sigma_t} I - \frac{\alpha_t^2}{\sigma_t^3} \operatorname{Cov}\Big(x^{(c)} \mid y_t = y\Big).$$

Thus $Jf^*_{t,c}(y)$ and $Jg^*_{t,c}(y)$ have the same eigenvectors. Their eigenvalue order is reversed because the covariance enters $Jf^*_{t,c}(y)$ with a negative coefficient.

The same eigendirections also appear in a single DDIM update. For an update from input time t to output time $u > t$, eliminate the noise prediction using $f_{t,c}(x) = (x - \alpha_t g_{t,c}(x))/\sigma_t$. The update becomes

$$x_u = \frac{\sigma_u}{\sigma_t} x_t + \left(\alpha_u - \frac{\sigma_u}{\sigma_t} \alpha_t \right) g_{t,c}(x_t),$$

with input Jacobian

$$J_{x_t} x_u = \frac{\sigma_u}{\sigma_t} I + \left(\alpha_u - \frac{\sigma_u}{\sigma_t} \alpha_t \right) Jg_{t,c}(x_t). \tag{C.11}$$

Both coefficients are nonnegative along the increasing-time schedule. Therefore, for the ideal estimator, the Jacobian of a single DDIM step has the same eigendirections as the conditional covariance, while its eigenvalues are shifted and rescaled.

C.4 How power iteration selects principal components

By Section C.3, the Jacobian $Jg_{t,c}(y)$ of the ideal denoiser estimator is symmetric positive semidefinite. Let $\kappa_1 \geq \cdots \geq \kappa_d \geq 0$ denote its eigenvalues, with orthonormal eigenvectors $v_1, \ldots, v_d$. Expanding the initial increment in this basis gives

$$d_0 = \sum_{\ell=1}^{d} \langle v_\ell, d_0 \rangle v_\ell,$$

$$\big(Jg_{t,c}(y)\big)^n d_0 = \sum_{\ell=1}^{d} \kappa_\ell^n \langle v_\ell, d_0 \rangle v_\ell.$$

The effect of repeated application is therefore simple. After n applications, the component along v_ℓ is weighted by κ_ℓ^n. Directions with larger eigenvalues consequently gain relative weight over

directions with smaller eigenvalues. For example, if $0 < \kappa_\ell < \kappa_1$ and the initial increment has nonzero components along both directions, their relative weighting changes by $(\kappa_1/\kappa_\ell)^n$.

The frozen denoising trajectory only approximately follows these Jacobian powers because the Jacobian changes as the image moves away from y. To account for this error, let P_r denote the projection onto the eigenspace corresponding to the leading r eigenvalues and write

$$E_n := \sum_{j=0}^{n-1} \left\| \left(Jg_{t,c}(y)\right)^{n-1-j} \right\| \|e_j\| .$$

By (C.6), E_n is second order in the largest frozen-time increment for fixed n. Projecting (C.7) onto the leading eigenspace and its orthogonal complement gives

$$\|(I - P_r)d_n\| \le \kappa_{r+1}^n \|(I - P_r)d_0\| + E_n,$$
$$\|P_r d_n\| \ge \kappa_r^n \|P_r d_0\| - E_n.$$

It follows that, for $1 \le n < s$,

$$\frac{\|(I - P_r)d_n\|}{\|P_r d_n\|} \le \frac{\kappa_{r+1}^n \|(I - P_r)d_0\| + E_n}{\kappa_r^n \|P_r d_0\| - E_n},$$

whenever the denominator is positive. If the Taylor error vanishes, the ratio between the components outside and inside the leading eigenspace decreases by a factor $(\kappa_{r+1}/\kappa_r)^n$. With a nonzero error, the same behavior holds as long as E_n remains small compared with the leading component. The increments themselves may decrease in magnitude while becoming increasingly concentrated in the leading eigenspace.

This effect also carries over to the final output of the frozen tail. Since $\bar{x}_s - y = \sum_{n=0}^{s-1} d_n$, the linear approximation weights the contribution of v_ℓ by $\sum_{n=0}^{s-1} \kappa_\ell^n$. Applying the same argument to (C.8) gives

$$\frac{\|(I - P_r)(\bar{x}_s - y)\|}{\|P_r(\bar{x}_s - y)\|} \le \frac{\left(\sum_{n=0}^{s-1} \kappa_{r+1}^n\right) \|(I - P_r)d_0\| + \sum_{n=0}^{s-1} E_n}{\left(\sum_{n=0}^{s-1} \kappa_r^n\right) \|P_r d_0\| - \sum_{n=0}^{s-1} E_n},$$

whenever the denominator is positive. Thus the eigendirections favored in the individual denoising increments are also favored in the accumulated change from y to the final image. In the simplest case $\kappa_1 > \kappa_2$, the usual power-iteration relation is

$$\left\| \frac{\left(Jg_{t,c}(y)\right)^s}{\kappa_1^s} - v_1 v_1^\top \right\| = \left(\frac{\kappa_2}{\kappa_1}\right)^s .$$

The ratio κ_2/κ_1 determines how quickly the leading direction separates from the others. More generally, concentration toward a leading r-dimensional subspace is controlled by the ratio κ_{r+1}/κ_r. When this ratio is close to one, a short tail produces little separation; when it is smaller, the preference for the leading eigenspace becomes stronger over repeated denoising steps.

The discussion so far concerns the time-frozen trajectory. We finally transfer the same picture to scheduled denoising and DDIM. The scheduled increment differs from the frozen increment by at most $\|x_{n+1} - \bar{x}_{n+1}\| + \|x_n - \bar{x}_n\|$. For the DDIM states u_n in Lemma B.3,

$$\left\|(u_{n+1} - u_n) - \left(Jg_{t,c}(y)\right)^n d_0\right\| \le E_n + \sum_{r=n}^{n+1} \left(\|u_r - x_r\| + \|x_r - \bar{x}_r\|\right).$$

The state-comparison terms are controlled by Section B and proposition 4.1. Hence, when these errors and the local Taylor error are small, the scheduled and DDIM increments follow the same spectral weighting predicted by the frozen iteration. The scheduled case follows by setting $u_r = x_r$.

C.5 Memorization Impact.

We find that the spectral view of tail denoising established in the previous two sections no longer holds once the model memorizes its training data. Following Bonnaire et al. (2025), we train a DiT from scratch on only two ImageNet classes, "bald eagle" and "Maltese dog," for 4000 epochs to encourage memorization. As Figure 8 shows, the generated samples closely match their nearest neighbors in the training set. Moreover, the leading eigenvectors of the denoising Jacobian of this memorizing model show no class-relevant structure (Figure 8, bottom row).

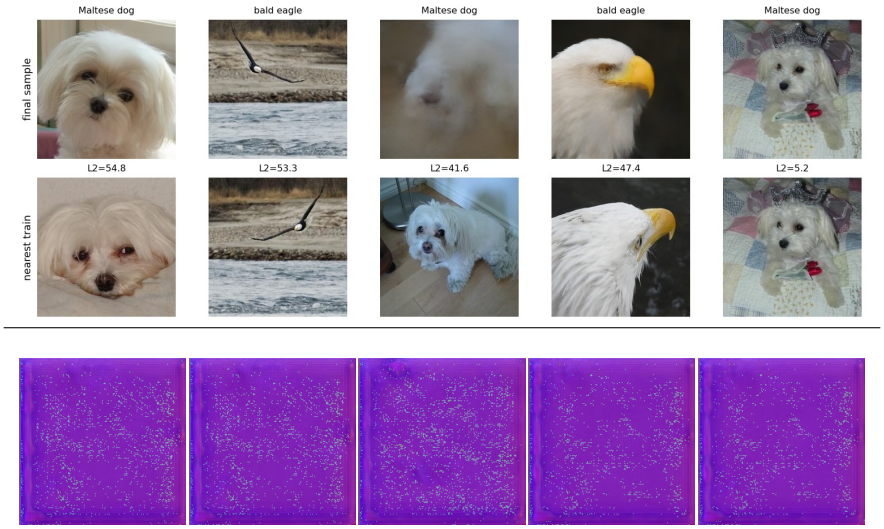


Figure 8: **No PCA structure after memorization.** *Top row:* samples generated by a DiT trained for 4000 epochs on "bald eagle" and "Maltese dog." *Middle row:* the nearest training image to each sample. *Bottom row:* the five leading eigenvectors of $Jg^*_{t_{40},\text{bald eagle}}(y)$, where y is obtained by 40 DDIM denoising steps with guidance toward "Maltese dog."