BEYOND SAMPLE COPYING: STRUCTURAL MEMORIZATION IN DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion models generalize well in practice. Paradoxically, an optimal diffusion model fully memorizes the training data and therefore fails to generalize, raising the question of what induces generalization in a real diffusion model. We show that diffusion models progressively overfit the denoising training objective, creating a generalization gap between validation and training performance at intermediate noise levels. In a fully analytic 2D toy model with a controlled denoising error, we trace this gap to the interaction between model error and the density of the data distribution's support. The optimal denoising flow field localizes sharply around individual training points, whereas model error suppresses exact recall of training points, yielding a smooth, generalizing flow field. Finally, we examine how training-time overfitting manifests along inference trajectories. We find that predictions made from intermediate trajectory states occupy a distinct feature-space regime from predictions made from noised training and validation images. As training progresses and model size increases, these predictions develop greater relative affinity to the training data, despite the absolute similarity to the validation data not decreasing. Together, these findings show that the denoising objective and inference trajectories express structural memorization differently.

All code will be released upon publication.

1 INTRODUCTION

The goal of generative modeling is to sample from an unknown data distribution given only a finite set of examples. Diffusion models and closely related flow models have achieved strong performance on continuous data, generating novel samples across image (Sohl-Dickstein et al., 2015; Ho et al., 2020; Karras et al., 2022; 2024b; Song et al., 2021; Kingma et al., 2023), video (Ho et al., 2022; Wang et al., 2023; Blattmann et al., 2023; Gupta et al., 2025; Melnik et al., 2024), and audio (Kong et al., 2021; Huang et al., 2023; 2025; Liu et al., 2023) domains, while achieving competitive results with autoregressive models for text (Popov et al., 2021; Xu et al., 2025; Ochs & Habernal, 2024; Yi et al., 2024).

Despite their impressive ability to generate high-quality, novel samples across different data modalities, diffusion models are known to occasionally replicate training samples (Carlini et al., 2023; Somepalli et al., 2023; Gu et al., 2025; Ren et al., 2025). Supporting this notion, Bayes-optimal diffusion models are unable to sample anything other than the training data (Scarvelis et al., 2025; Adaloglou et al., 2025; Gao & Li, 2024), Practical generalization must therefore arise from the fact that learned flow fields only approximate this optimum. Existing analyses describe this approximation as a smoothing of the optimal field (Farghly et al., 2025a; Achilli et al., 2024; Scarvelis et al., 2025) and identify regularization, inductive bias, and training dynamics as factors that influence the balance between generalization and memorization (Kamb & Ganguli, 2025; Shah et al., 2025; Bonnaire et al., 2025).

Analysing the noise schedule of diffusion models reveals that memorization risk is not uniform across noise levels (Dodson et al., 2026). Instead, the intermediate noise levels are the critical regime for determining both memorization and generalization. Memorization risk is not uniform in the input space either. Instead, better diffusion models selectively fit the score in some regions of the input space, while underfitting it in others (Song et al., 2025). Specifically, as models improve,

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.

their score approximation improves on the regions of input space the model is highly likely to see during training, while it worsens outside of those regions.

Memorization is not limited to verbatim copying, either. Garnier-Brun et al. recently described a form of biased generalization in which test loss continues to improve while generated samples become increasingly proximate to the training data (Garnier-Brun et al., 2026). This result shows that held-out loss and final-sample proximity capture different aspects of generalization. It also raises a more specific question: how is training-specific memorization on the denoising objective expressed along inference trajectories?

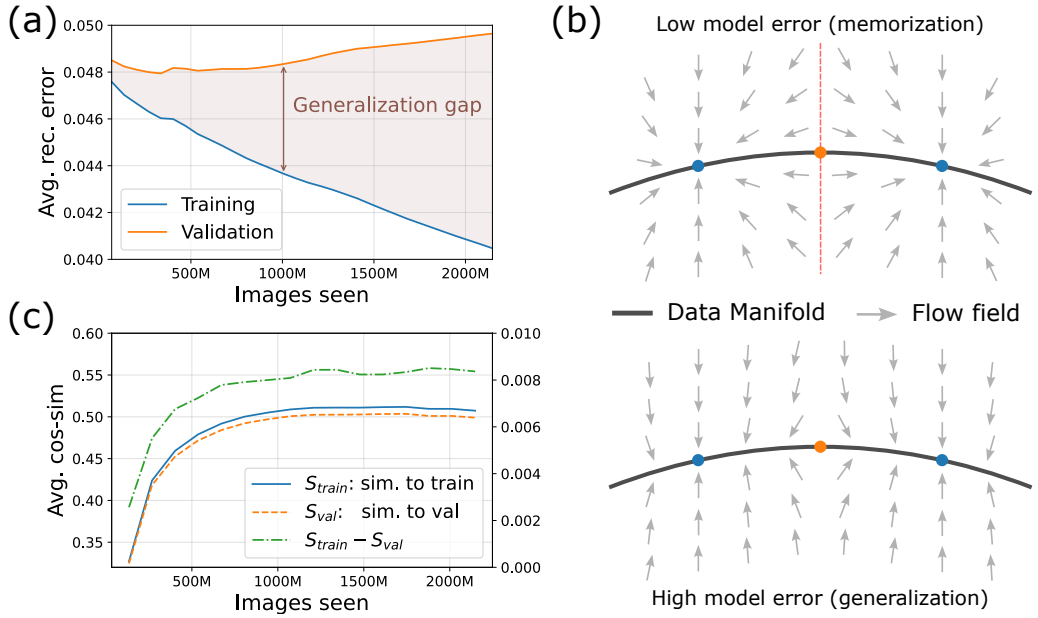


Figure 1: **Qualitative summary of contributions.** **(a)** State-of-the-art diffusion models exhibit a growing generalization gap: the difference in reconstruction error (denoising training objective) between training and validation data increases as training progresses. **(b)** This gap is modulated by model error: Increasing model error smoothes the learned flow field, suppressing the generalization gap. For low model error (top), the flow field localizes sharply around training points, while for high model error (bottom), it more smoothly approaches the data manifold. **(c)** Similarly, during inference, trajectories become increasingly biased towards the training data, relative to the validation data, despite increasing their absolute similarity to both.

In this work, we investigate this kind of structural memorization — a systematic difference in how diffusion models treat training versus unseen data — which persists independently of whether any given generated sample is a verbatim copy. We trace this behaviour from the denoising objective, through a mechanistic toy model, to inference trajectories.

Our starting point is the denoising objective used by training, where we compare the reconstruction performance on noised training and validation examples at matched noise levels. This reveals a substantial generalization gap that increases with training time, model size, and decreasing dataset size. The gap is concentrated at intermediate noise levels and, at sufficiently long training, is accompanied by a decrease in validation performance while training performance continues to improve.

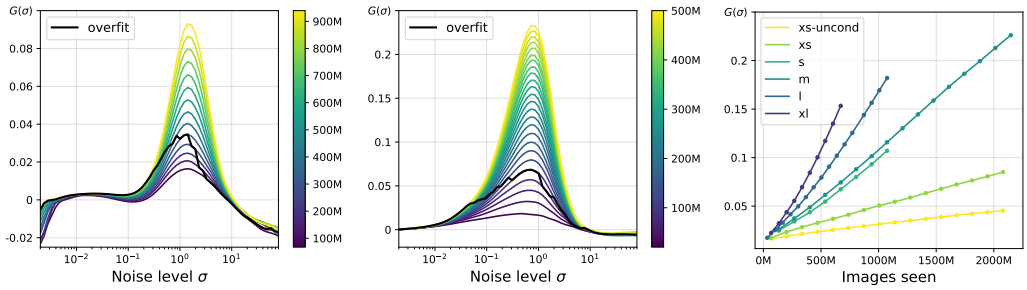
By simulating an error-prone denoiser in a fully analytic 2D toy model, we explain how this gap can arise from the interplay between model error and the density of the data distribution's support. When model error is small, or the data manifold is sparsely supported, the learned flow field localizes sharply around individual training samples, leading to a large gap between training and validation performance. Larger model error or denser support smooths the field around the data support, thereby suppressing this gap.

Finally, turning to inference-level memorization, we define the relative feature-space affinity of predictions made from intermediate states of inference trajectories to the training reference set relative

to the validation reference set. This measure differs from the reconstruction-based generalization gap, but it likewise probes differential treatment of training and validation data. We show that these predictions occupy a distinct feature-space regime from predictions made from directly noised training and validation images. As training progresses and model capacity increases, these predictions develop greater relative affinity to the training data, while their absolute similarity to the validation data does not decrease.

In summary, our analysis traces training-specific behaviour across levels of the diffusion process: from training objective denoising performance, through a mechanistic model of the learned flow field, to the feature-space geometry of inference trajectories.

2 Better performing diffusion models show a larger generalization gap



(a) EDM2-XXL, IN-512 (b) EDM, CIFAR-10 (c) EDM2, IN-64

Figure 2: The relative generalization gap $G(\sigma)$ increases with images seen during training (colorbar in millions) and with model size in relation to dataset size. The black line indicates the beginning of overfitting (validation error starts increasing) for each noise level σ.

To quantify the difference in denoising performance between training and validation data, we define the **relative generalization gap**

$$G(\sigma) = \frac{L_{\text{val}}(\sigma) - L_{\text{train}}(\sigma)}{L_{\text{train}}(\sigma)}, \tag{1}$$

where L_{train} and L_{val} denote the mean squared reconstruction errors at the noise level σ on training and validation examples, respectively. For a dataset split $\mathcal{D} \in \{\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{val}}\}$, we define

$$L(\mathcal{D}, \sigma) = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} \|\boldsymbol{y}_i - \boldsymbol{y}(\boldsymbol{x}_i, \sigma)\|_2^2 , \tag{2}$$

when reconstructing clean data points $\boldsymbol{y}_i$ from noisy data points $\boldsymbol{x}_i := \boldsymbol{x}_i(\sigma) = \boldsymbol{y}_i + \sigma \boldsymbol{n}_i$, $\boldsymbol{n}_i \sim \mathcal{N}(0, I)$, using the target predictor $\boldsymbol{y}(\cdot, \sigma)$. Denote $L_{\text{train}}(\sigma) := L(\mathcal{D}_{\text{train}}, \sigma)$ and $L_{\text{val}}(\sigma) := L(\mathcal{D}_{\text{val}}, \sigma)$.

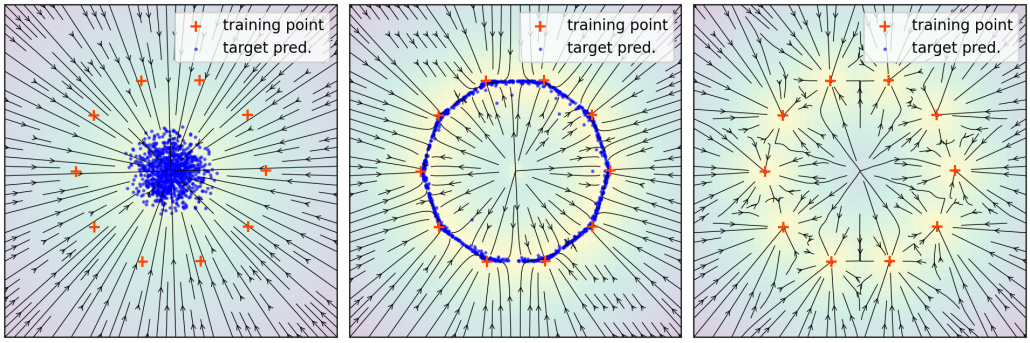
Perhaps unsurprisingly, we find that well-performing denoiser models show a significant relative generalization gap between training and validation data (Fig. 2), indicating that they exhibit different reconstruction performance on training and validation data for a given noise scale. This behavior is consistent across all models and datasets we tested. We use the pretrained EDM2 (Karras et al., 2024b) models to carry out experiments on ImageNet and both pretrained and re-trained EDM (Karras et al., 2022) models for experiments on CIFAR-10 and CIFAR-100. The hyperparameters are as reported in (Karras et al., 2022; 2024b), unless specified otherwise. Additional results for these models and datasets are available in the supplementary material.

We observe a positive correlation between the relative generalization gap and longer training times, larger model size, and smaller dataset size. This is consistent with findings of other memorization metrics, where the onset of memorization scales linearly with dataset size (Bonnaire et al., 2025; Favero et al., 2025). Additionally, we observe classical overfitting: during continued training, the

validation reconstruction performance degrades (Fig. 2, black line), while the training reconstruction performance improves. Crucially, both the relative generalization gap and the overfitting regime are most pronounced at intermediate noise levels and sharply drop outside of this range, consistent with previous findings about memorization being most pronounced at intermediate noise levels (Dodson et al., 2026). This raises the question of what causes the relative generalization gap to arise and how samples generated by standard diffusion sampling are affected by this.

3 Model error causes the generalization gap

To understand the mechanism behind the noise-dependent relative generalization gap observed in Section 2, we study a fully analytic 2D toy diffusion model, following (Adaloglou et al., 2025). Consider a distribution of N training points $\{\boldsymbol{y}_i^{train}\}_{i=1}^N$ and N validation points $\{\boldsymbol{y}_i^{\text{val}}\}_{i=1}^N$ along a circle, alternating between each other. The shape of a circle was chosen for two reasons: it mimics the isotropy of high-dimensional data distributions, and the superposition of samples lies far outside the data manifold. The equidistant distribution of data points is chosen for simplicity. The results of this section are qualitatively unchanged for a random distribution on the circle. See supplementary material for more details.



(a) High noise (b) Intermediate noise (c) Low noise

Figure 3: Flow field lines of the optimal target predictor $\boldsymbol{y}$ at different noise levels $\sigma = 28, 2.8, 0.63$. Blue dots show predictions $\boldsymbol{y}^*(\boldsymbol{x}, \sigma)$ for $\boldsymbol{x} = \boldsymbol{y} + \sigma\boldsymbol{n}, \boldsymbol{n} \sim \mathcal{N}(0, I), \boldsymbol{y} \sim D_{\text{train}}$. Color indicates the magnitude of $\boldsymbol{y}^*(\boldsymbol{x}, \sigma) - \boldsymbol{x}$. Field geometry is **(a)** global, predictions tend towards the superposition of all training points, **(b)** moderately localized around training points, predictions approximate the data manifold, and **(c)** highly localized around each training point, predictions replicate the training points.

Consider the time variable $t \in [0, 1]$ and the current noise level $\sigma := \sigma(t)$, with $\sigma(0) = 0$ and $\sigma(1) \gg std(\mathbf{y})$. We denote by $\boldsymbol{x} := \boldsymbol{x}(\sigma(t))$ a point along a denoising trajectory. The **Bayes-optimal target predictor** is the posterior mean under the empirical training distribution (Karras et al., 2022):

$$\boldsymbol{y}^*(\boldsymbol{x}, \sigma) = \frac{\sum_{i=1}^N \boldsymbol{y}_i^{\text{train}} \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{y}_i^{\text{train}}, \sigma^2 I)}{\sum_{i=1}^N \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{y}_i^{\text{train}}, \sigma^2 I)}. \tag{3}$$

This predictor defines a smooth vector field whose geometry depends on σ (Fig. 3). In the high-noise limit, all training points contribute equally, giving

$$\lim_{\sigma \to \infty} \boldsymbol{y}^*(\boldsymbol{x}, \sigma) = \frac{1}{N} \sum_{i=1}^N \boldsymbol{y}_i^{\text{train}}. \tag{4}$$

To see this, write

$$\boldsymbol{y}^*(\boldsymbol{x}, \sigma) = \frac{\sum_i \boldsymbol{y}_i^{\text{train}} w_i(\boldsymbol{x}, \sigma)}{\sum_i w_i(\boldsymbol{x}, \sigma)}, \tag{5}$$

with

$$w_i(\boldsymbol{x}, \sigma) := \exp\left(\frac{\boldsymbol{x}^\top \boldsymbol{y}_i^{\text{train}}}{\sigma^2} - \frac{\|\boldsymbol{y}_i^{\text{train}}\|_2^2}{2\sigma^2}\right). \tag{6}$$

As $\sigma \to \infty$, $w_i(\boldsymbol{x}, \sigma) \to 1$ for every i, which yields Eq. (4). In the low-noise limit, the posterior concentrates on the nearest training point, giving

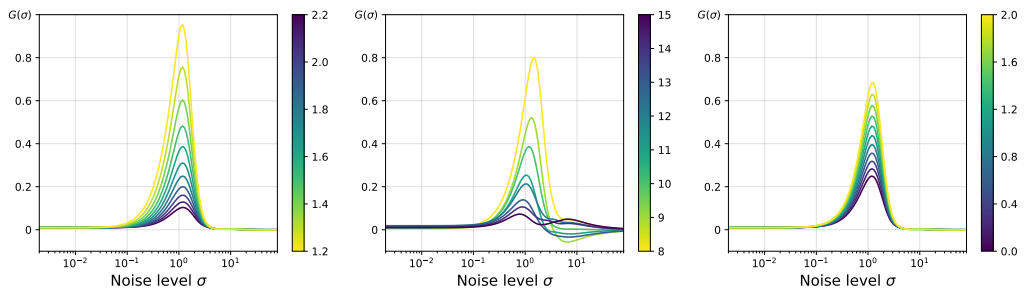
$$\lim_{\sigma \to 0} \boldsymbol{y}^*(\boldsymbol{x}, \sigma) = \arg\min_{\boldsymbol{y}_i^{\text{train}}} \|\boldsymbol{x} - \boldsymbol{y}_i^{\text{train}}\|_2. \tag{7}$$

For intermediate noise levels σ, a mixture of the above occurs, where noisy samples $\boldsymbol{x}$ are attracted by only their closest training samples. Sampling randomly initialized trajectories with this predictor yields perfect memorization of the training set. Note that the conditional optimal predictor is the same as $\boldsymbol{y}^*$, but the posterior mean includes only those training points satisfying the given condition.

To simulate real, erroneous diffusion models, we construct an **error-prone target predictor** $\boldsymbol{y}_\delta$ by substituting the real data distribution $P(\boldsymbol{y})$ by an "uncertain" data distribution $P_\delta(\boldsymbol{y}) = \int_{\mathbb{R}^d} \mathcal{N}(\boldsymbol{y}|\boldsymbol{y}', \delta^2) P(\boldsymbol{y}') d\boldsymbol{y}'$. The true data points are randomly shifted by isotropic Gaussian noise with variance δ^2, effectively broadening the hyperplane that the model fits. This construction yields the optimal denoiser for $\delta = 0$, the "pass-through" property defined in (Lukoianov et al., 2025), and The smoothing effect of the model error parameter δ on the flow field connects to several characterizations of how practical diffusion models deviate from the optimal score (Li et al., 2024; Vastola, 2025; Zhou et al., 2026), but provides a controlled setting in which the effect of this smoothing can be traced analytically. The corresponding posterior mean is given by

$$\boldsymbol{y}_\delta(x, \sigma) = \frac{\delta^2 \boldsymbol{x} + \sigma^2\, \boldsymbol{y}^*(\boldsymbol{x}, \tilde{\sigma})}{\tilde{\sigma}^2}, \qquad \tilde{\sigma}^2 = \sigma^2 + \delta^2. \tag{8}$$

Thus, model error both smooths the optimal predictor through the shift $\sigma \mapsto \tilde{\sigma}$ and mixes it with the identity map through an affine combination (Adaloglou et al., 2025). Whereas $\boldsymbol{y}^*$ can only reproduce training examples, $\boldsymbol{y}_\delta$ generalizes through random shifts induced by the initial noise of a trajectory. It is interesting to see that despite its simplicity, the toy model can qualitatively reproduce the generalization gap behaviour of EDM/EDM2 trained on ImageNet data (Fig. 4 vs Fig. 2), using the same protocol as in Section 2. Note that the toy model error parameter δ plays a role similar to training time in EDM/EDM2 in affecting the distance to the optimal denoiser.



(a) Error parameter δ (b) Training set size N (c) Guidance weight w

Figure 4: Relative generalization gap $G(\sigma)$ in our 2D toy model with circular data manifold **(a)** for different settings of the model error parameter δ (colorbar), **(b)** for different number of training points N (colorbar), and **(c)** computed using a single point on the data manifold as the validation set, parametrized by its angle ($\sigma \approx 1.1$). The resulting gap is zero at training points, and largest between training points, with magnitude related inversely to δ.

Analyzing the error-prone denoiser on noisy data points $\boldsymbol{x} = \boldsymbol{y}^{\mathcal{D}} + \sigma \boldsymbol{n}$, with $\mathcal{D} \in \{\text{train}, \text{val}\}$, $\boldsymbol{n} \sim \mathcal{N}(0, I)$, explains why the gap emerges and why it scales with the model error δ. We can write the *prediction error* as

$$\boldsymbol{y}_\delta(\boldsymbol{x}, \sigma) - \boldsymbol{y} = \frac{\sigma^2}{\tilde{\sigma}^2} \underbrace{(\boldsymbol{y}^*(\boldsymbol{x}, \tilde{\sigma}) - \boldsymbol{y})}_{\text{optimal prediction error}} + \frac{\delta^2}{\tilde{\sigma}^2} \sigma \boldsymbol{n}, \tag{9}$$

which corresponds to the reconstruction error we used in Eq. (2) to compute the relative generalization gap. At intermediate noise levels ($\sigma_{\text{min}} \ll \sigma \ll \sigma_{\text{max}}$), the optimal flow field localizes around training points (Fig. 3b), causing a gap in the optimal prediction error in Eq. (9) to arise between training and validation targets. This translates to a gap between the error-prone prediction errors (Fig. 4a) (the cross-term in the square of Eq. (9) contributes about half as much). However, several factors can suppress this mechanism, leading to the characteristic shape of the gap observed in Fig. 2: At high noise levels ($\sigma \approx \sigma_{\text{max}} \gg \delta$), the error-prone prediction collapses to the superposition of all training samples:

$$\boldsymbol{y}_\delta(\boldsymbol{x}, \sigma) = \boldsymbol{y}^*(\boldsymbol{x}, \sigma) + \mathcal{O}(\delta^2/\sigma^2) \overset{Eq.\,(4)}{\approx} \frac{1}{N} \sum_i \boldsymbol{y}_i^{train}, \tag{10}$$

resulting in similar errors to the training and validation samples. At low noise levels or high model error ($\sigma^2/\delta^2 \ll 1$), we have $\boldsymbol{y}_\delta - \boldsymbol{y} = \sigma \boldsymbol{n} + \mathcal{O}(\sigma^2/\delta^2)$, which causes the error to drop with the noise level σ, regardless of whether $\boldsymbol{y}$ is sampled from the training or validation set.

The scale of the generalization gap is determined by the geometry around training points. Hence, it is also influenced by the density of training points on the manifold. The results in Fig. 2 suggest this, with the much smaller CIFAR datasets showing significantly higher levels of overfitting than the larger ImageNet dataset, especially when comparing models with a similar number of parameters. Similarly, theoretical analyses show that the memorization-generalization crossover depends on the ratio of data structure dimensionality to dataset size (Achilli et al., 2024; Buchanan et al., 2025). Indeed, a denser distribution of training points on the manifold shows a significantly reduced gap in the toy model (Fig. 4b). Furthermore, this suggests that increasing the number of classes in class-conditional models (condition granularity) amplifies the gap, since in the toy model it's equivalent to reducing the dataset to specific subsets. Applying diffusion guidance (Karras et al., 2024a) to the toy model, it predicts a steady increase in the generalization gap as the guidance weight increases (Fig. 4c). We tested these predictions in Sections C and 5 against real diffusion models.

4 Feature-Space Geometry of Diffusion Inference

In this section, we ask how the structural memorization we investigated in Section 2 and Section 3 manifests in standard, class-conditional diffusion inference by examining the path inference trajectories take through space and the factors that influence it. Specifically, we want to answer the question of whether predictions made from inference trajectories are systematically biased toward the training set, resemble predictions made from unseen validation images, or occupy a distinct regime altogether. As intermediate trajectory states $\boldsymbol{x}(t)$ have no matched clean target, we cannot use the L2-reconstruction-based metric from Section 2 to analyse their relation to training and validation images.

We therefore compare predictions generated from three query populations: noisy training images, noisy validation images, and intermediate states of standard class-conditional inference. Consider a query set that determines the model input, and a reference set that supplies clean images to which the resulting predictions are compared. Denote the query sets as $Q \in \{\text{train}, \text{val}, \text{traj}\}$. Each query image is passed through the model at noise level σ to obtain a one-shot prediction $\boldsymbol{y}(\boldsymbol{q}, \sigma)$ of the corresponding clean image. For trajectory queries, σ is the noise level of the intermediate state, while for noisy training and validation queries, σ is chosen to match this same value. The reference can be either clean training or clean validation images, $R \in \{\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{val}}\}$.

To measure local similarity, we cannot use matched targets for noisy queries, and distributional metrics such as FID and FDD summarize differences between feature distributions and therefore do not directly measure whether individual query predictions are preferentially close to training rather than validation images. A viable alternative is to measure local similarity in feature space using cosine similarity. Specifically, consider the average cosine similarity between the predictions of each query image $\boldsymbol{q} \in Q$ and its k nearest reference images $\boldsymbol{r} \in R$:

$$S_k(Q, R; \sigma) = \frac{1}{|Q|} \sum_{\boldsymbol{q} \in Q} \frac{1}{k} \sum_{\boldsymbol{r} \in \text{kNN}(\boldsymbol{y}(\boldsymbol{q}, \sigma); R)} \cos(\phi(\boldsymbol{y}(\boldsymbol{q}, \sigma)), \phi(\boldsymbol{r})), \tag{11}$$

where $\boldsymbol{y}(\cdot, \sigma)$ is a one-shot model prediction, ϕ is a feature map, and $\text{kNN}(\boldsymbol{y}(\boldsymbol{q}, \sigma); R)$ are the k nearest neighbors of $\boldsymbol{y}(\boldsymbol{q}, \sigma)$ from R in feature space and restricted to reference images from the

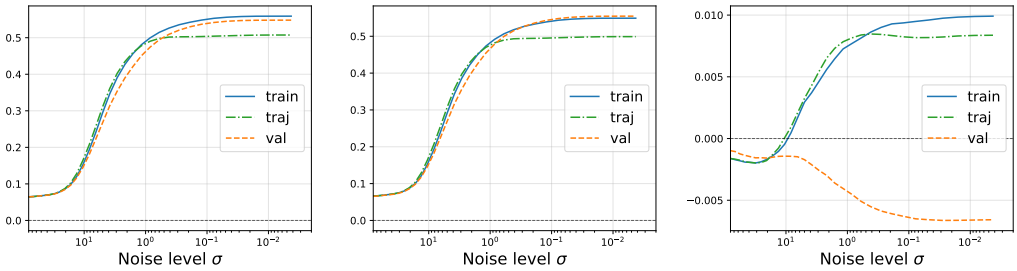
same class as $\boldsymbol{q}$. We use the well-established DINOv2 feature extractor because of its reliability and robustness when evaluating generative models Stein et al. (2023). When an image is part of both the query and the reference set, that image is excluded from the candidate neighbours, preventing the query's own reference prediction from dominating the score.

The hyperparameter k defines how many nearest neighbors are used in Eq. (11). For each $\boldsymbol{y}(\boldsymbol{q}, \sigma)$, small values of k give greater weight to the closest reference points and are therefore more sensitive to memorized images or near-duplicates. Larger k average over a wider local neighbourhood, making the statistic more representative of regional feature-space similarity. We use $k = 20$, balancing sensitivity to individual near-neighbours against the variance of a single-neighbour statistic. The qualitative trends remain unchanged across the range of values we tested for k (Fig. 6c, Fig. 20).

Using these absolute similarity measures, we define the relative train-affinity of Q:

$$B_k(Q; \sigma) = S_k(Q, \mathcal{D}_{\text{train}}; \sigma) - S_k(Q, \mathcal{D}_{\text{val}}; \sigma). \tag{12}$$

If $B_k(Q; \sigma) > 0$, then model predictions from Q are, on average, more similar to the training reference set than to the unseen validation reference set in feature space. For a full description of our experimental pipeline, see Section B. Similarity to the training reference set, $S_k(Q, \mathcal{D}_{\text{train}}; \sigma)$, mea-



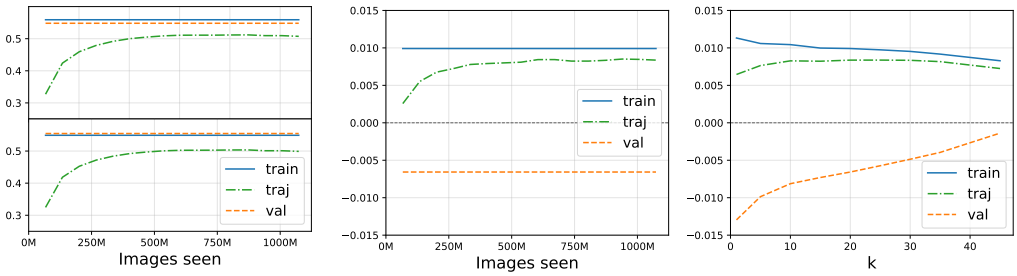
(a) $S_{20}(Q, \mathcal{D}_{\text{train}}; \sigma)$ (b) $S_{20}(Q, \mathcal{D}_{\text{val}}; \sigma)$ (c) $B_{20}(Q; \sigma)$

Figure 5: **(a) and (b)** Average cosine similarity between predictions of query images $\boldsymbol{q} \in Q$ and clean images from the reference sets $\mathcal{D}_{\text{train/val}}$ (Eq. (11)). As the noise level σ decreases, overall similarity to the reference sets increases. At low noise, both noisy training and noisy validation images as queries (blue and orange lines) exceed intermediate trajectory states (green line) in absolute similarity to both reference sets. **(c)** The relative train affinity (Eq. (12)) of intermediate trajectory states (green line) increases alongside the absolute similarities to the reference sets. All plots show EDM2-S on ImageNet64.

sures how closely the predictions are associated with clean training images in feature space, whereas similarity to the validation reference set, $S_k(Q, \mathcal{D}_{\text{val}}; \sigma)$, provides the corresponding comparison for unseen images from the true data distribution.

Fig. 5a and b show that, as soon as the noise level allows for discernible image features, intermediate trajectory states have lower absolute similarity to both reference sets than predictions from noised training or validation images. Assuming the reference sets sufficiently cover the data manifold and that DINOv2 similarity reflects proximity to it, this gap indicates that the trajectories cannot fully close the distance to the manifold of the true data distribution. Furthermore, trajectory-derived predictions exhibit a strong relative train-affinity (Fig. 5c). This affinity is most pronounced at low noise, even though trajectory-derived predictions remain less similar to both reference sets in absolute terms. This provides feature-space evidence from inference trajectories that is consistent with the structural memorization identified by the paired L2 analysis, while measuring a different quantity.

This geometric separation extends the supervision-region perspective of Dodson et al. Dodson et al. (2026): Our comparison shows that intermediate trajectory states are separated not only from the regions represented by noisy training queries, but also from the regions represented by directly noised validation queries. Thus, the relevant distinction here is not only between supervised and unsupervised inputs: trajectory states occupy a feature-space regime distinct from directly noised data from the entire data distribution.



(a) $S_{20}(Q, \mathcal{D}_{\text{train/val}}; \sigma)$ (b) $B_{20}(Q; \sigma)$ (c) $B_k(Q; \sigma)$

Figure 6: **(a)** Average cosine similarity between predictions of query images $\boldsymbol{q} \in Q$ and clean images from the reference sets $\mathcal{D}_{\text{train/val}}$ (Eq. (11)). As training time increases, the absolute similarity of intermediate trajectory states (green line) to both reference sets increases. **(b)** So does their relative train-affinity (Eq. (12), green line). **(c)** Their positive relative train-affinity is robust across choices of k. All plots show EDM2-S on ImageNet64.

With longer training, and likewise with larger model size, the absolute-similarity gap between trajectory-derived predictions and predictions from directly noised images progressively closes, closing the aforementioned distance to the manifold of the data distribution (Fig. 6a, Fig. 19, Fig. 6). This improvement comes at the cost of a larger relative train-affinity (Fig. 6b), consistent with the increasing relative generalization gap between training and validation performance in Section 2 and Section 3. Because similarity to the validation reference set, $S_k(\text{traj}, \mathcal{D}_{\text{val}}; \sigma)$, does not decrease with longer training or larger model size, this result reflects a relative shift toward the training reference set rather than degraded similarity to validation data.

Together with the reconstruction-based analysis from Section 2, these results show that the factors associated with a growing train–validation reconstruction gap—later checkpoints and larger models—are also associated with greater relative training affinity in trajectory-derived predictions. However, on the denoising training objective, they cause an eventual degradation in validation performance, while in inference, the absolute similarity to the validation set keeps increasing in all settings we tested.

5 Strong Conditioning Amplifies Memorization

Because diffusion models operate in high-dimensional spaces, they contain many regions where the data manifold is sparsely supported. The data samples in these regions tend to be memorized and are not significantly less likely under the model than other data samples (van den Burg & Williams, 2021). This effect is exacerbated when modeling a conditional distribution, since it is supported by an even smaller subset of the dataset. Thus, model conditioning should amplify memorization, and the more granular the conditioning, the stronger the memorization, with text-to-image models being an extreme case. To test this, we re-trained EDM (Karras et al., 2022) on CIFAR-10 for 100M images across varying numbers of classes. Classes were computed algorithmically using k-means clustering in DINOv2 feature space.

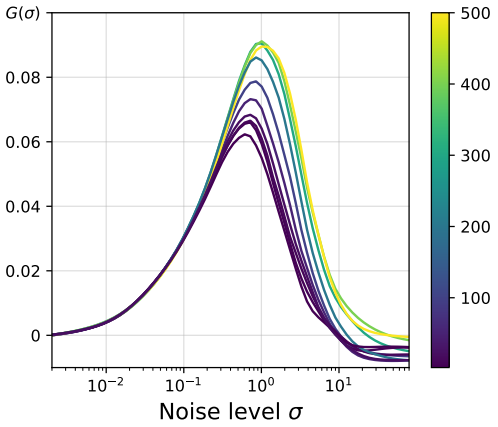


Figure 7: The relative generalization gap $G(\sigma)$ with average L2-reconstruction error increases and shifts towards higher noise levels when the number of classes (colorbar) for a class-conditioned model increases. Each line corresponds to a re-trained EDM model on CIFAR-10.

Under the reconstruction-based metric from Section 2, an increasingly granular model con-

dition indeed exhibits an increased generalization gap (Fig. 7a). This is consistent with prior work, which found that conditioning on even uninformative random labels triggers memorization in diffusion models (Gu et al., 2025). Additionally, we find that this increase is more pronounced at higher noise levels and disappears entirely at lower noise levels, indicating that the condition primarily guides the early stages of the generation process. This is consistent with the toy model (Fig. 4b), though the effect here is weaker, likely because, unlike in the toy model, the conditional predictors are not independent across conditions.

6 Conclusion, limitations and future work

In this work, we studied how structural memorization in diffusion models relates to both the denoising objective and the inference process. On the denoising objective, real diffusion models develop a growing train–validation reconstruction gap at intermediate noise levels that scales with longer training and larger model size, and is eventually accompanied by a decrease in validation performance. A fully analytic 2D toy model demonstrates how this gap arises from the interaction between model error and the density of the data distribution's support: the optimal denoising flow field localizes sharply around training samples, whereas model error smooths the learned flow field and can suppress exact recall. Along inference trajectories, we find a related but distinct effect. Predictions made from intermediate states occupy a different feature-space regime from predictions made from directly noised training and validation images, yet their relative affinity shifts toward the training reference set as training progresses and model capacity increases. Unlike the reconstruction gap, this shift does not cause a decrease in absolute similarity to validation references.

The reconstruction gap and trajectory affinity are therefore not interchangeable measures, and the trajectory affinity alone does not establish a target-paired inference generalization gap. The former uses matched clean targets and directly measures denoising performance; the latter remains meaningful even when inputs have no matched targets. The inference result therefore complements the training-level picture: predictions from trajectory states do not resemble predictions from noised training images, but they do carry a bias towards the training set.

Our experiments are limited to class-conditional image generation using EDM and EDM2 architectures on ImageNet and CIFAR benchmarks. The reconstruction gap follows from general properties of the score-matching objective, but the quantitative form of the gap and the trajectory-affinity result may depend on the data modality, sampler, feature representation, and reference-set construction.

Beyond these methodological limitations, the results expose a measurement problem. Reconstruction-based metrics, verbatim-copy tests (Somepalli et al., 2023; Carlini et al., 2023; Wen et al., 2024; Hasegawa & Yasuda, 2025), distributional distances, and trajectory-affinity measures probe different aspects of memorization and generalization, and no single metric captures the full phenomenon (Stein et al., 2023). A recent concurrent study likewise shows that standard test loss can coexist with increasing training-specificity in generated samples (Garnier-Brun et al., 2026), complementing our analysis. An important open question is how these relative affinity shifts relate to final-sample copying, privacy leakage, and downstream sample quality, and whether they can be altered independently through training or inference interventions.

AI USE STATEMENT

In this work, we used generative AI tools for assistance with

- propose or refine hypotheses
- design or provide feedback on research
- methodology or experiments
- implement methods
- interpret results

We have not used generative AI tools for help with

- develop theoretical models or conceptual frameworks
- formulate mathematical claims
- provide critical ingredients for proving mathematical claims
- assist in the writing of proofs

and

- generate synthetic data sets
- assist with translation
- clean and reformat dataset
- support qualitative and thematic data analysis

are not applicable to this work. Additionally, we used generative AI tools for

- suggest experimental parameters
- create or edit software code
- summarize or analyse existing literature
- brainstorming
- sourcing/searching for information
- edit a research paper to improve readability
- identify relevant literature
- suggest a structure for a research paper
- propose a title or keywords for a research paper

We have reviewed all AI-assisted work. 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 want to ensure that our results are fully reproducible with minimal effort. The most direct measure we take to achieve that is to publish our codebase upon publication of the paper. Furthermore, Section B contains the technical details of all our experiments. The proofs and derivation of section Section 3 are referenced, unless they are trivial.

AUTHOR CONTRIBUTIONS

Tim Kaiser contributed all code and experiments, the literature research for the manuscript, the structure of the paper and its narrative, the writing of the manuscript and its final polish. Prof. Dr. Markus Kollmann contributed the initial idea of the relative generalization gap with average L2-reconstruction error, as well as guidance and discussion of the results, narrative, and manuscript.

REFERENCES

Beatrice Achilli, Enrico Ventura, Gianluigi Silvestri, Bao Pham, Gabriel Raya, Dmitry Krotov, Carlo Lucibello, and Luca Ambrogioni. Losing dimensions: Geometric memorization in generative diffusion, 2024. URL https://arxiv.org/abs/2410.08727.

Nikolas Adaloglou, Tim Kaiser, Damir Iagudin, and Markus Kollmann. Guiding a diffusion model using sliding windows, 2025. URL https://arxiv.org/abs/2411.10257.

Donghoon Ahn, Hyoungwon Cho, Jaewon Min, Wooseok Jang, Jungwoo Kim, SeonHwa Kim, Hyun Hee Park, Kyong Hwan Jin, and Seungryong Kim. Self-rectifying diffusion sampling with perturbed-attention guidance, 2025. URL https://arxiv.org/abs/2403.17377.

Sumukh K Aithal, Pratyush Maini, Zachary C. Lipton, and J. Zico Kolter. Understanding hallucinations in diffusion models through mode interpolation, 2024. URL https://arxiv.org/abs/2406.09358.

Jie An, De Wang, Pengsheng Guo, Jiebo Luo, and Alexander Schwing. On inductive biases that enable generalization of diffusion transformers, 2024. URL https://arxiv.org/abs/2410.21273.

Ricardo Baptista, Agnimitra Dasgupta, Nikola B. Kovachki, Assad Oberai, and Andrew M. Stuart. Memorization and regularization in generative diffusion models, 2025. URL https://arxiv.org/abs/2501.15785.

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

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023. URL https://arxiv.org/abs/2311.15127.

Tony Bonnaire, Raphaël Urfin, Giulio Biroli, and Marc Mézard. Why diffusion models don't memorize: The role of implicit dynamical regularization in training, 2025. URL https://arxiv.org/abs/2505.17638.

Sam Buchanan, Druv Pai, Yi Ma, and Valentin De Bortoli. On the edge of memorization in diffusion models, 2025. URL https://arxiv.org/abs/2508.17689.

Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models, 2023. URL https://arxiv.org/abs/2301.13188.

Zhengdao Chen. On the interpolation effect of score smoothing in diffusion models, 2026. URL https://arxiv.org/abs/2502.19499.

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

Nick Dodson, Xinyu Gao, Qingsong Wang, Yusu Wang, and Zhengchao Wan. Two calm ends and the wild middle: A geometric picture of memorization in diffusion models, 2026. URL https://arxiv.org/abs/2602.17846.

Mischa Dombrowski, Felix Nützel, and Bernhard Kainz. Lcmem: A universal model for robust image memorization detection, 2025. URL https://arxiv.org/abs/2512.14421.

Tyler Farghly, Peter Potaptchik, Samuel Howard, George Deligiannidis, and Jakiw Pidstrigach. Diffusion models and the manifold hypothesis: Log-domain smoothing is geometry adaptive, 2025a. URL https://arxiv.org/abs/2510.02305.

Tyler Farghly, Patrick Rebeschini, George Deligiannidis, and Arnaud Doucet. Implicit regularisation in diffusion models: An algorithm-dependent generalisation analysis, 2025b. URL https://arxiv.org/abs/2507.03756.

Alessandro Favero, Antonio Sclocchi, and Matthieu Wyart. Bigger isn't always memorizing: Early stopping overparameterized diffusion models, 2025. URL https://arxiv.org/abs/2505.16959.

Weiguo Gao and Ming Li. How do flow matching models memorize and generalize in sample data subspaces?, 2024. URL https://arxiv.org/abs/2410.23594.

Jerome Garnier-Brun, Luca Biggio, Davide Beltrame, Marc Mézard, and Luca Saglietti. Biased generalization in diffusion models, 2026. URL https://arxiv.org/abs/2603.03469.

Anand Jerry George, Rodrigo Veiga, and Nicolas Macris. Denoising score matching with random features: Insights on diffusion models from precise learning curves, 2025. URL https://arxiv.org/abs/2502.00336.

Xiangming Gu, Chao Du, Tianyu Pang, Chongxuan Li, Min Lin, and Ye Wang. On memorization in diffusion models, 2025. URL https://arxiv.org/abs/2310.02664.

Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Fei-Fei Li, Irfan Essa, Lu Jiang, and José Lezama. Photorealistic video generation with diffusion models. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 393–411, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-72986-7.

Indranil Halder. A solvable generative model with a linear, one-step denoiser, 2025. URL https://arxiv.org/abs/2411.17807.

Masaya Hasegawa and Koji Yasuda. Quantifying the ease of reproducing training data in unconditional diffusion models, 2025. URL https://arxiv.org/abs/2503.19429.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models, 2022.

Susung Hong. Smoothed energy guidance: Guiding diffusion models with reduced energy curvature of attention, 2024. URL https://arxiv.org/abs/2408.00760.

Kuan-Po Huang, Shu wen Yang, Huy Phan, Bo-Ru Lu, Byeonggeun Kim, Sashank Macha, Qingming Tang, Shalini Ghosh, Hung yi Lee, Chieh-Chi Kao, and Chao Wang. Impact: Iterative mask-based parallel decoding for text-to-audio generation with diffusion modeling, 2025. URL https://arxiv.org/abs/2506.00736.

Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models, 2023. URL https://arxiv.org/abs/2301.12661.

Zahra Kadkhodaie, Florentin Guth, Eero P. Simoncelli, and Stéphane Mallat. Generalization in diffusion models arises from geometry-adaptive harmonic representations, 2024. URL https://arxiv.org/abs/2310.02557.

Mason Kamb and Surya Ganguli. An analytic theory of creativity in convolutional diffusion models, 2025. URL https://arxiv.org/abs/2412.20292.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models, 2022. URL https://arxiv.org/abs/2206.00364.

Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself, 2024a. URL https://arxiv.org/abs/2406.02507.

Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models, 2024b. URL https://arxiv.org/abs/2312.02696.

Diederik P. Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models, 2023.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis, 2021.

Antoni Kowalczuk, Dominik Hintersdorf, Lukas Struppek, Kristian Kersting, Adam Dziedzic, and Franziska Boenisch. Finding dori: Memorization in text-to-image diffusion models is not local, 2025. URL https://arxiv.org/abs/2507.16880.

Xiang Li, Yixiang Dai, and Qing Qu. Understanding generalizability of diffusion models requires rethinking the hidden gaussian structure, 2024. URL https://arxiv.org/abs/2410.24060.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D. Plumbley. Audioldm: Text-to-audio generation with latent diffusion models, 2023. URL https://arxiv.org/abs/2301.12503.

Artem Lukoianov, Chenyang Yuan, Justin Solomon, and Vincent Sitzmann. Locality in image diffusion models emerges from data statistics, 2025. URL https://arxiv.org/abs/2509.09672.

Andrew Melnik, Michal Ljubljanac, Cong Lu, Qi Yan, Weiming Ren, and Helge Ritter. Video diffusion models: A survey, 2024. URL https://arxiv.org/abs/2405.03150.

Matthew Niedoba, Berend Zwartsenberg, Kevin Murphy, and Frank Wood. Towards a mechanistic explanation of diffusion model generalization, 2025. URL https://arxiv.org/abs/2411.19339.

Sebastian Ochs and Ivan Habernal. Private synthetic text generation with diffusion models, 2024. URL https://arxiv.org/abs/2410.22971.

Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech, 2021.

Jie Ren, Yaxin Li, Shenglai Zeng, Han Xu, Lingjuan Lyu, Yue Xing, and Jiliang Tang. Unveiling and mitigating memorization in text-to-image diffusion models through cross attention, 2025. URL https://arxiv.org/abs/2403.11052.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752.

Christopher Scarvelis, Haitz Sáez de Ocáriz Borde, and Justin Solomon. Closed-form diffusion models, 2025. URL https://arxiv.org/abs/2310.12395.

Kulin Shah, Alkis Kalavasis, Adam R. Klivans, and Giannis Daras. Does generation require memorization? creative diffusion models using ambient diffusion, 2025. URL https://arxiv.org/abs/2502.21278.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics, 2015.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models, 2022. URL https://arxiv.org/abs/2212.03860.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and mitigating copying in diffusion models, 2023. URL https://arxiv.org/abs/2305.20086.

Kiwhan Song, Jaeyeon Kim, Sitan Chen, Yilun Du, Sham Kakade, and Vincent Sitzmann. Selective underfitting in diffusion models, 2025. URL https://arxiv.org/abs/2510.01378.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution, 2020. URL https://arxiv.org/abs/1907.05600.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations, 2021.

George Stein, Jesse C. Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Leigh Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L. Caterini, J. Eric T. Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models, 2023. URL https://arxiv.org/abs/2306.04675.

Gerrit J. J. van den Burg and Christopher K. I. Williams. On memorization in probabilistic deep generative models, 2021. URL https://arxiv.org/abs/2106.03216.

John J. Vastola. Generalization through variance: how noise shapes inductive biases in diffusion models, 2025. URL https://arxiv.org/abs/2504.12532.

Enrico Ventura, Beatrice Achilli, Gianluigi Silvestri, Carlo Lucibello, and Luca Ambrogioni. Manifolds, random matrices and spectral gaps: The geometric phases of generative diffusion, 2025. URL https://arxiv.org/abs/2410.05898.

Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, Yuwei Guo, Tianxing Wu, Chenyang Si, Yuming Jiang, Cunjian Chen, Chen Change Loy, Bo Dai, Dahua Lin, Yu Qiao, and Ziwei Liu. Lavie: High-quality video generation with cascaded latent diffusion models, 2023. URL https://arxiv.org/abs/2309.15103.

Yuxin Wen, Yuchen Liu, Chen Chen, and Lingjuan Lyu. Detecting, explaining, and mitigating memorization in diffusion models, 2024. URL https://arxiv.org/abs/2407.21720.

Minkai Xu, Tomas Geffner, Karsten Kreis, Weili Nie, Yilun Xu, Jure Leskovec, Stefano Ermon, and Arash Vahdat. Energy-based diffusion language models for text generation, 2025. URL https://arxiv.org/abs/2410.21357.

Qiuhua Yi, Xiangfan Chen, Chenwei Zhang, Zehai Zhou, Linan Zhu, and Xiangjie Kong. Diffusion models in text generation: a survey. *PeerJ Computer Science*, 10:e1905, 2024. doi: 10.7717/peerj-cs.1905. URL https://doi.org/10.7717/peerj-cs.1905.

Xinyu Zhou, Jiawei Zhang, and Stephen J. Wright. Smoothing the score function for generalization in diffusion models: An optimization-based explanation framework, 2026. URL https://arxiv.org/abs/2601.19285.

Beyond Sample Copying: Structural Memorization in Diffusion Models

Supplementary Material

This document provides details and additional experiments supporting the main text. In Section A, we review the related literature in detail. In Section B, we provide the necessary details to reproduce our results. In Section C, we investigate how the relative generalization gap behaves when using diffusion guidance. In Section D, we provide additional experiments about the impact of the modelling of long- and short-range dependencies on memorization. In section Section E, we report results for the toy model with random instead of symmetric data. In Section F and Section G, we provide additional results supporting the claims we made in Section 2 and Section 4.

A Related work

Background. Diffusion models were developed in their current form from two different approaches: via variational inference with an objective resembling denoising score matching (Ho et al., 2020), and by time-reversed stochastic differential equations with direct application of score matching (Song & Ermon, 2020; Song et al., 2021). Several improvements, including architectural advances (Dhariwal & Nichol, 2021), sampling techniques such as guidance (Dhariwal & Nichol, 2021), and latent diffusion (Rombach et al., 2022), have established diffusion models as the dominant class of generative models. This work uses the EDM framework developed by Karas et al (Karras et al., 2022), which achieved state-of-the-art performance in the class-conditional setting, owing to a comprehensive suite of analyses and improvements to the training and sampling of diffusion models (Karras et al., 2022; 2024b).

Understanding generalization in diffusion models. Recently, analytical solutions for denoising trajectories have been established for simple or small-sized datasets, showing that optimal diffusion models are unable to sample anything but the training data (Scarvelis et al., 2025; Adaloglou et al., 2025; Gao & Li, 2024). This is in stark contrast to the impressive ability of real diffusion models to generate high-quality, novel samples across different data modalities. Tackling this paradox, a series of works showed that the optimal flow field has a sharp geometry around the training samples (Gao & Li, 2024; Ventura et al., 2025) and that smoothing this geometry plays a key role for successful generalization (Farghly et al., 2025a; Achilli et al., 2024; Scarvelis et al., 2025). Several works have further characterized this smoothing through different lenses: as convergence toward the Gaussian structure of the training data (Li et al., 2024), as an effect of variance in the score matching target (Vastola, 2025), as an effect of using a neural network to approximate the score (Zhou et al., 2026), and as a regularization problem (Baptista et al., 2025). Possible causes of this smoothing include locality (Lukoianov et al., 2025) and equivariance (Kamb & Ganguli, 2025) biases from both data statistics and the architecture, a bias towards geometry-adaptive harmonic bases adapted to the underlying images (Kadkhodaie et al., 2024), and implicit regularization (Farghly et al., 2025b). Finally, it has been shown that a smoothed flow field causes diffusion models to interpolate between modes of the training data (Chen, 2026; Aithal et al., 2024). Training dynamics further modulate this: training time (Bonnaire et al., 2025), dataset size (Halder, 2025; Favero et al., 2025), and model complexity all affect when and how strongly memorization appears, alongside less obvious factors such as differences in training loss (Buchanan et al., 2025), the number of noise samples per data point (George et al., 2025), and model conditioning (Gu et al., 2025).

Detecting memorization in diffusion models. Prior work on detecting memorization has largely focused on sample-level detection at inference: Nearest-neighbor and copy-detection approaches (Somepalli et al., 2022; 2023; Kowalczuk et al., 2025; Dombrowski et al., 2025) flag generated images that closely match a specific training sample. Extraction attacks (Carlini et al., 2023; Wen et al., 2024) recover memorized images by adversarially probing the model. Model-centric scores such as Effective Model Memorization (Gu et al., 2025) and ODE-volume-based quantification (Hasegawa & Yasuda, 2025) summarize a model's overall tendency toward verbatim reproduction.

Training-bias beyond standard generalization metrics. A recent line of work has questioned whether decreasing held-out loss is sufficient to characterize generalization in diffusion models.

Garnier-Brun et al. identify a phase of biased generalization in which test loss continues to decrease while generated samples acquire anomalously high proximity to the training data, and attribute this effect to the progressive learning of finer, more training-dependent features (Garnier-Brun et al., 2026). Their analysis operates primarily at the level of final generated samples and their proximity to the training set. We study the relationship between this kind of training-specificity and the denoising process itself, at all noise levels, and in three regimes of the input space: Noisy training images, noisy held-out validation images, and the intermediate states of inference trajectories. The resulting trajectory measure is complementary to, rather than a substitute for, final-sample proximity or held-out loss.

The role of the noise level. The noise level at which memorization manifests has received particular attention. Statistical-physics analyses of the optimal score predict a "collapse" regime at small noise, where trajectories are pulled into individual training points (Biroli et al., 2024; Ventura et al., 2025). Closely related to our analysis of the relative generalization gap, a complementary geometric account argues that memorization risk is highly non-uniform across noise levels, and identifies a danger zone at intermediate noise levels where memorization is most pronounced (Dodson et al., 2026). Our analysis agrees that the reconstruction gap is concentrated at intermediate noise, but asks a different question at inference: how do predictions from trajectory states relate to training and validation references? We find that they develop greater relative affinity to training references as training progresses and model capacity increases, without a corresponding decrease in their absolute similarity to validation references.

The role of the denoising objective Closely related to our finding about the interplay between sampling trajectories and the likely locations of noisy training images in input space, Song et al. refine the perspective of a globally smoothed score function (Song et al., 2025). They find that better diffusion models selectively fit the score in some regions of the input space, while underfitting it in others. As models improve, their score approximation improves in the regions of input space the model is highly likely to see during training, while it worsens outside of those regions. As we do, they observe that inference trajectories and noisy training images have distinct geometries in feature space. Unlike Song et al., we quantify this asymmetry at the reconstruction level and its expression along inference trajectories.

B Reproducibility

B.1 Technical details

- **Feature extractors.** We used the official implementation from (Karras et al., 2024a) for the feature detectors DINOv2 ViT-L/14 and Inception-V3.
- **Noise schedule.** We used the official hyperparameters from (Karras et al., 2024b), i.e., $\sigma \in [2e-3, 80]$.
- **Sampler.** We used the official EDM/EDM2-sampler from (Karras et al., 2024b) with default hyperparameters.
- **EDM preconditioning.** For all experiments, we use the prediction after preconditioning, i.e., the full target prediction. Any other choice would lead to distortions based on the preconditioning parameters.

B.2 Average reconstruction error details

Input construction We construct all denoiser inputs by directly noising a data point $\boldsymbol{y} \in \mathcal{D}_{\text{train/val}}$: $\boldsymbol{x}(\sigma) = \boldsymbol{y} + \sigma\eta$. We match the noise realization $\eta \sim \mathcal{N}(0, I)$ to each image $\boldsymbol{y}$ via its random seed, so it remains fixed across noise levels σ, model sizes, and model checkpoints. We sweep 64 noise levels on a log scale within $\sigma \in [2e-3, 80]$, with 8192 samples each.

B.3 Average cosine similarity details

Query construction. For both metrics, each query is first passed through the denoiser at the specified noise level. Let $\phi(\cdot)$ denote the feature extractor and let $\hat{\boldsymbol{q}}_i$ denote the resulting one-step prediction. For directly noised queries, the denoiser input is $\boldsymbol{x}_i(\sigma) = \boldsymbol{q}_i + \sigma\eta_i$. For trajectory queries,

Algorithm 1 $L(\mathcal{D}, \sigma)$ (Eq. (2))

Require: Denoiser D, reference set with class-labels $\mathcal{D} = \{(\boldsymbol{y}_i, c_i)\}$, noise level σ
1: **for** $\boldsymbol{y}_i \in \mathcal{R}$ **do**
2: $\quad \boldsymbol{x}_i(\sigma) \leftarrow \boldsymbol{y}_i + \sigma\eta, \eta \sim N(0, I)$
3: $\quad$ Compute one-step denoiser prediction $\hat{\boldsymbol{y}}_i \leftarrow D(\boldsymbol{x}_i(\sigma),\, \sigma,\, c_i)$
4: **end for**
5: $L(\mathcal{D}, \sigma) \leftarrow \frac{1}{N} \sum_i ||\hat{\boldsymbol{y}}_i - \boldsymbol{y}_i||_2^2$
6: **return** $L(\mathcal{D}, \sigma)$

$\boldsymbol{x}_i(\sigma)$ is the saved diffusion trajectory state at the corresponding noise level. The denoiser and feature-extraction steps are therefore identical across query populations; only the construction of the denoiser input differs.

kNN feature similarity. For each query prediction $\hat{\boldsymbol{q}}_i$, we compute cosine similarities to reference features from the same class,

$$a_{ij} = \cos(\phi(\hat{\boldsymbol{q}}_i), \phi(\boldsymbol{r}_j)), \qquad \mathcal{J}_i = \{j : c_j = c_i\}.$$

Let $\mathcal{N}_k(i; R)$ denote the indices of the k largest values of a_{ij} among $j \in \mathcal{J}_i$. The kNN similarity is

$$S_k(Q, R; \sigma) = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{k} \sum_{j \in \mathcal{N}_k(i;R)} a_{ij}.$$

Thus, $S_k(Q, R; \sigma)$ is the mean cosine similarity between each query prediction and its k most similar same-class reference images, averaged first over neighbors and then over queries. The reference features are computed once and reused across query populations, noise levels, checkpoints, and model sizes.

For the train–validation affinity analysis, we evaluate each query population against both reference sets and define

$$B_k(Q; \sigma) = S_k(Q, R_{\text{train}}; \sigma) - S_k(Q, R_{\text{val}}; \sigma).$$

Positive values indicate greater feature-space affinity to the training reference set than to the validation reference set. Exact self-matches are excluded from $S_k(Q, R_{\text{train}})$ for training queries and from $S_k(Q, R_{\text{val}})$ for validation queries.

Algorithm 2 $S_k(Q, R; \sigma)$ (Eq. (11))

Require: Denoiser D, query set $Q = \{(\boldsymbol{q}_i, c_i)\}_{i=1}^{N_Q}$, reference set $R = \{(\boldsymbol{r}_j, c_j)\}_{j=1}^{N_R}$, noise level σ, feature extractor ϕ, number of neighbors k
1: **for** $j = 1, \ldots, N_R$ **do**
2: $\quad v_j \leftarrow \phi(\boldsymbol{r}_j)$
3: **end for**
4: **for** $i = 1, \ldots, N_Q$ **do**
5: $\quad$ Construct the denoiser input $\boldsymbol{x}_i(\sigma)$
6: $\quad \hat{\boldsymbol{q}}_i \leftarrow D(\boldsymbol{x}_i(\sigma), \sigma, c_i)$
7: $\quad \boldsymbol{u}_i \leftarrow \phi(\hat{\boldsymbol{q}}_i)$
8: $\quad \mathcal{J}_i \leftarrow \{j : c_j = c_i\}$
9: $\quad a_{ij} \leftarrow \cos(\boldsymbol{u}_i, \boldsymbol{v}_j)$ for all $j \in \mathcal{J}_i$
10: $\quad$ Select $\mathcal{N}_k(i; R)$, the indices of the k largest a_{ij}
11: $\quad s_i \leftarrow \frac{1}{k} \sum_{j \in \mathcal{N}_k(i;R)} a_{ij}$
12: **end for**
13: $S_k(Q, R; \sigma) \leftarrow \frac{1}{N_Q} \sum_{i=1}^{N_Q} s_i$
14: **return** $S_k(Q, R; \sigma)$

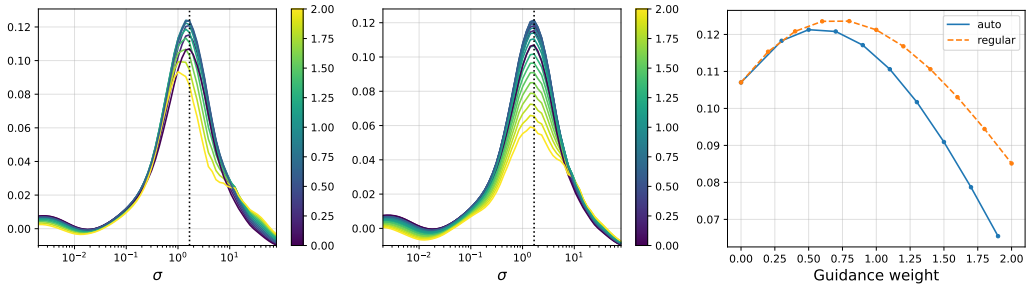
B.4 Compute resources

Below is a table detailing the compute resources required to run the experiments of this paper. We used 4 NVIDIA A100 GPUs (45GB) for all experiments. FDD and FID are computed in parallel, as they share everything but their feature extractor. Full experiments consist of runs across many model sizes, checkpoints, noise levels, and possibly additional parameters, such as guidance weights. Due to the large number of forward passes per sample, standard FD computation is highly sensitive to model size. In contrast, the reconstruction-based metrics are much less sensitive, since there, feature extraction, FD computation, and decoding share a relatively larger part of the total cost.

Table 1: Compute requirements for metric evaluation on ImageNet-64, for one model (checkpoint) and a single noise level σ. GPU-h and wall-clock time for ImageNet range from smallest to largest model size. They are approximate and refer to a server with 4xNVIDIA-A100 (45GB).

Dataset	**Metric**	**Sample size**	**GPU-h**	**Wall-clock**
ImageNet-64	$L(\mathcal{D}, \sigma)$	8192	4m - 4m	1m - 1m
	Standard FD	50.000	56m - 4h20m	14m - 1h5m
	$S_k(Q, R; \sigma)$	10.000	4m - 4m	1m - 1m
ImageNet-512	$L(\mathcal{D}, \sigma)$	8192	16m - 16m	4m - 4m
	Standard FD	50.000	2h28m - 7h44m	37m - 1h56m
	$S_k(Q, R; \sigma)$	10.000	-	-
CIFAR-10/100	$L(\mathcal{D}, \sigma)$	8192	4m	1m
	Standard FD	50.000	8m	2m
	$S_k(Q, R; \sigma)$	10.000	8m	2m

C Diffusion guidance



(a) CFG (b) Autoguidance (c) $\sigma \approx 1.67$

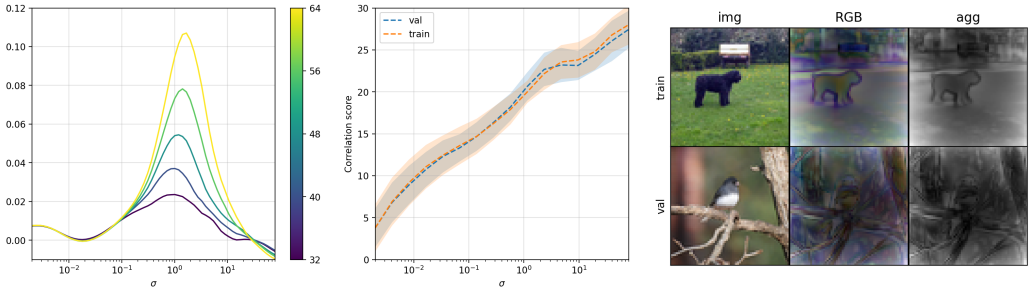
Figure 8: Relative generalization gaps (Eq. (1)) with EDM2-S on ImageNet-64. $w = 0$ corresponds to no guidance. An increasing guidance weight amplifies the gap, but this trend reverses for moderate to high guidance weights. (c) fixes σ close to the peak of the generalization gaps in (a) and (b), indicated by the dotted black line.

Diffusion guidance is a sampling technique that enhances sample quality by correcting biases in the diffusion model using an auxiliary diffusion model that exhibits these biases even more strongly (Adaloglou et al., 2025). Classifier-Free Guidance (CFG) (Ho & Salimans, 2022) uses an unconditional model to amplify conditioning, whereas other methods use auxiliary models to correct or accentuate particular biases (Ahn et al., 2025; Adaloglou et al., 2025; Hong, 2024). The best results (on FDD) are obtained when the auxiliary model is a worse version of the primary model, e.g., by having fewer parameters or less training time (Autoguidance (Karras et al., 2024a)). In our toy model, this corresponds to combining predictors with different levels of model error. The resulting guided predictor more closely approximates the optimal denoiser and thereby sharpens the learned flow field (Adaloglou et al., 2025). Since an increased condition granularity amplifies the generalization gap (Section 5), the same might hold for CFG, because it accentuates the model condition. The

toy model predicts a moderate yet steady increase in the generalization gap as the guidance weight increases, given that the primary and auxiliary models differ only in their model error δ (Fig. 4c). However, at larger guidance weights, guidance amplifies small differences between the models too much, like color saturation, which degrades overall performance. This is not reflected in the toy model, due to its simplicity.

In practice, we see exactly these two trends play out: For small to moderate guidance weights, the gap increases, whereas larger guidance weights significantly reduce it (Fig. 8). This trend is similar for both CFG and Autoguidance. Note that optimal performance, as measured by FDD, is achieved around $w = 0.9$ for CFG and $w = 1.4$ for Autoguidance. When increasing the weight further, the degradation due to the overcorrection of small differences between the models outweighs the improvements due to the correction of the right biases Adaloglou et al. (2025).

D LONG-RANGE DEPENDENCIES TEND TO BE MEMORIZED



(a) Reduced receptive field (b) Avg. correlation score (c) Pixel-wise correlation score

Figure 9: **(a)** Decreasing the receptive field mitigates the relative generalization gap (Eq. (1)) with M_{L2}. Colorbar shows the sliding window size. 64 corresponds to regular inference. **(b)** During inference, the model shifts from long-range correlations to short-range correlations. The correlation score is averaged over pixels and images. **(c)** Edge pixels tend to have longer-range dependencies. $\sigma \approx 1.2$. RGB shows the pixel-wise correlation score, agg is aggregated over channels. All plots show ImageNet-64, EDM2-S.

Globally coherent images are more difficult to generate because features that require correct long-range pixel dependencies are highly image-specific and thus harder to generalize than local pixel patterns. Similar to training samples in sparsely supported regions of the input space, this should promote memorization of said features. Similar hypotheses were formulated by Niedoba et al. (Niedoba et al., 2025), who showed that aggregating local empirical denoisers replicates network generalization behavior, and An et al. who showed that restricting attention locality in diffusion transformers improves generalization (An et al., 2024). To test this, we constrain the range of pixel dependencies the denoiser can model, following (Adaloglou et al., 2025), by denoising smaller image patches of equal size and then composing them back together. Overlapping pixels are averaged. We find that this dramatically shrinks the generalization gap (Fig. 9a), supporting the hypothesis.

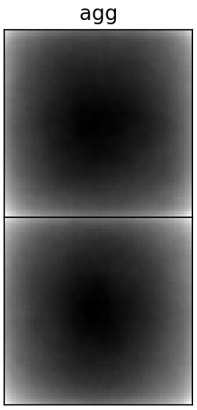


Figure 10: Edge pixels tend to have longer-range dependencies. Shown is the pixel-wise correlation score, averaged over 128 images and aggregated over channels, computed with EDM2-S on ImageNet-64 at $\sigma = 1$.

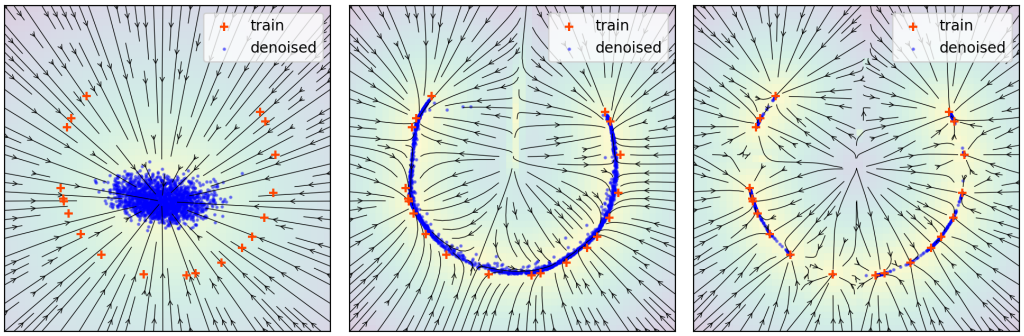
Further, we investigate how the denoising of each pixel is affected by other pixels, how these dependencies are distributed within an image, and how this relates to the noise level σ(t). We define a pixel-level correlation score by computing the average gradient of a pixel in the denoised image with respect to all pixels in the input image, weighted by the L2 distance to those pixels. Specifically, given an image $\boldsymbol{y}$ with N pixels, a denoiser $D(.)$, noise level σ and noise $\boldsymbol{n} \sim \mathcal{N}(0, 1)$, we compute the correlation score of the ij-th

pixel as

$$CS_{ij} = \frac{\frac{1}{N}\sum_{k,l} d(i,j,k,l)\frac{\partial D(\boldsymbol{x})_{ij}}{\partial \boldsymbol{y}_{kl}}}{\frac{1}{N}\sum_{k,l}\frac{\partial D(\boldsymbol{x})_{ij}}{\partial \boldsymbol{y}_{kl}}}, \quad \boldsymbol{x} = \boldsymbol{y} + \sigma\boldsymbol{n}, \boldsymbol{n} \sim \mathcal{N}(0, I)$$

with $d(i,j,k,l) = \left\| \begin{pmatrix} i \\ j \end{pmatrix} - \begin{pmatrix} k \\ l \end{pmatrix} \right\|_2$ being the L2 distance between the ij-th and kl-th pixels. A larger correlation score for a given pixel indicates that the gradient activations are high for distant pixels. We find that, on average, the farther a pixel in the output image is from the center, the longer its dependencies extend across the input images (Fig. 9c, Fig. 10). In other words, the object determines the background, but the background does not determine the object. It is unclear how the dataset's object-centric nature influences this result. When averaging the correlation scores across images, we see that as the noise level $\sigma(t)$ decreases, the distance of the average gradient signal across an image also decreases (Fig. 9b), i.e., the focus shifts from long-range dependencies to short-range dependencies during inference.

E TOY MODEL WITH RANDOM DATA



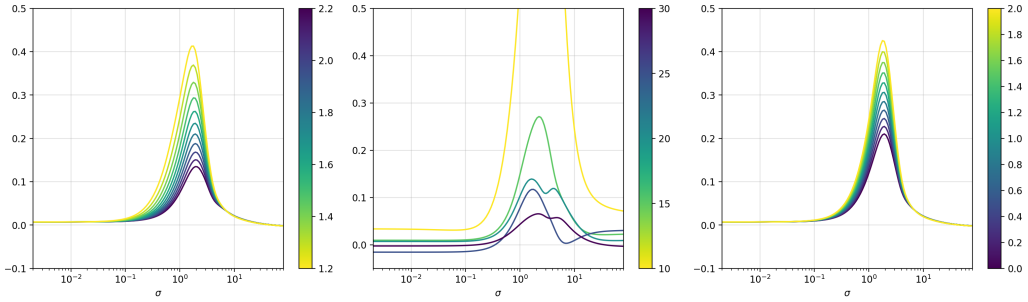
(a) High noise (b) Intermediate noise (c) Low noise

Figure 11: Flow field lines and predictions of the optimal target predictor $\boldsymbol{y}^*(\boldsymbol{x}, \sigma)$ at different noise levels $\sigma = 28, 2.8, 0.63$. Color indicates the magnitude of the prediction error $\boldsymbol{y}^*(\boldsymbol{x}, \sigma) - x$. Field geometry is **(a)** global, predictions tend towards the superposition of all training points. **(b)** moderately localized around training points, predictions approximate the data manifold **(c)** highly localized around each training point, predictions replicate the training points.

Instead of considering a symmetric distribution of N training points $\{\boldsymbol{y}_i^{train}\}_{i=1}^N$ and N validation points $\{\boldsymbol{y}_i^{val}\}_{i=1}^N$, we consider a random distribution of those points on a circle. The results are qualitatively the same, although considerably noisier, depending on the random seed. Fig. 11 shows the flow field geometry for this setup. Note that, especially for small N, the superposition of all training points does not necessarily lie at the center of the circle (Fig. 11a), possibly causing the right tail of the generalization gap to be different from 0 (Fig. 12)b. The manifold approximated by reconstructions at intermediate noise can have disconnected regions (Fig. 11b), where support is sparse, but densely populated regions collapse onto the training points later (Fig. 11c). The supplementary material includes a Jupyter Notebook that reproduces all figures related to the toy model and more.

F ADDITIONAL RESULTS FOR THE RELATIVE GENERALIZATION GAP

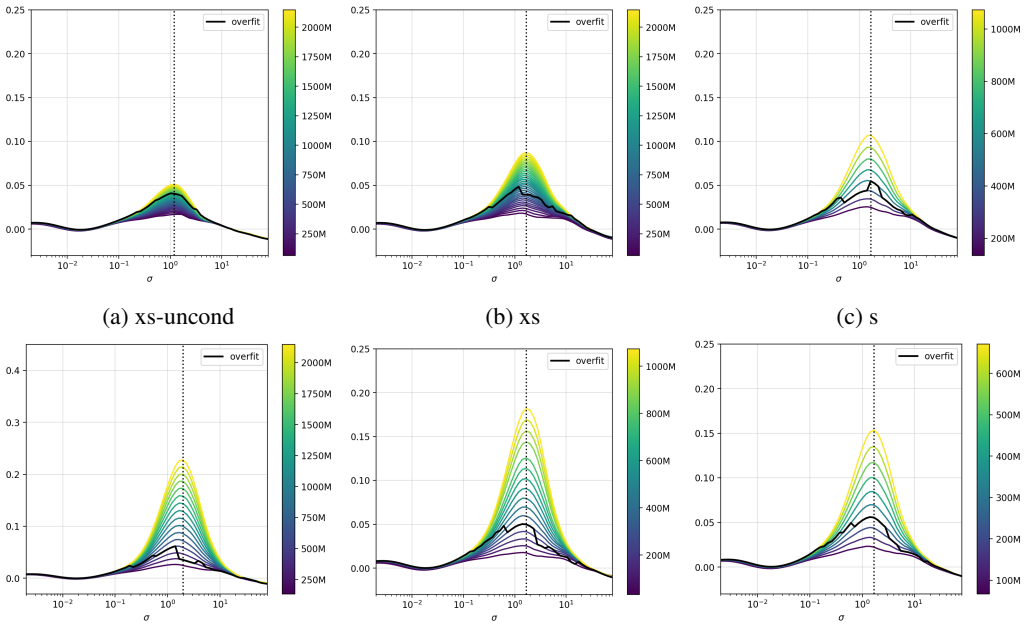
This section presents the relative generalization gaps across all our metrics with the EDM2 (Karras et al., 2024b) model on ImageNet-64/512 and the EDM (Karras et al., 2022) model on CIFAR-10/100.



(a) Error parameter δ (b) Training size N (c) Guidance weight w

Figure 12: Relative generalization gap (Eq. (1)) with M_{L2} in our 2D toy model with random data on a circular manifold and **(a)** different settings of the model error parameter δ, **(b)** different number of training points, and **(c)** different settings for the guidance weight w. $w = 0$ corresponds to no guidance.

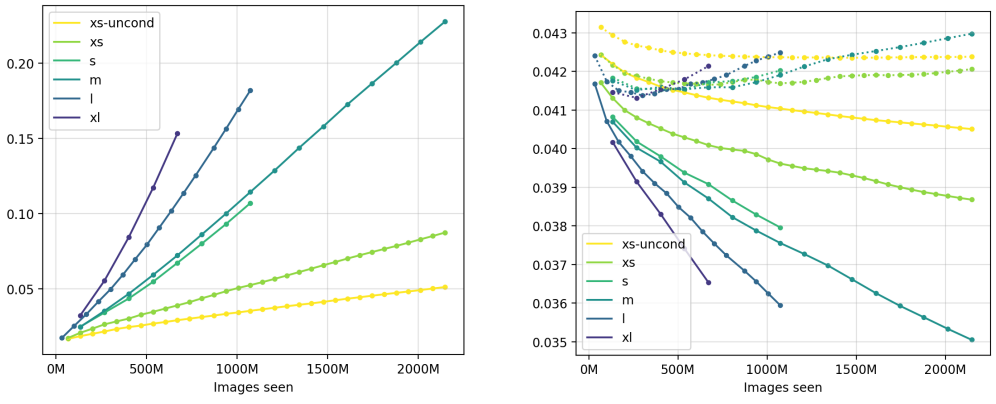
F.1 IMAGENET-64



(a) xs-uncond (b) xs (c) s

(d) m (e) l (f) xl

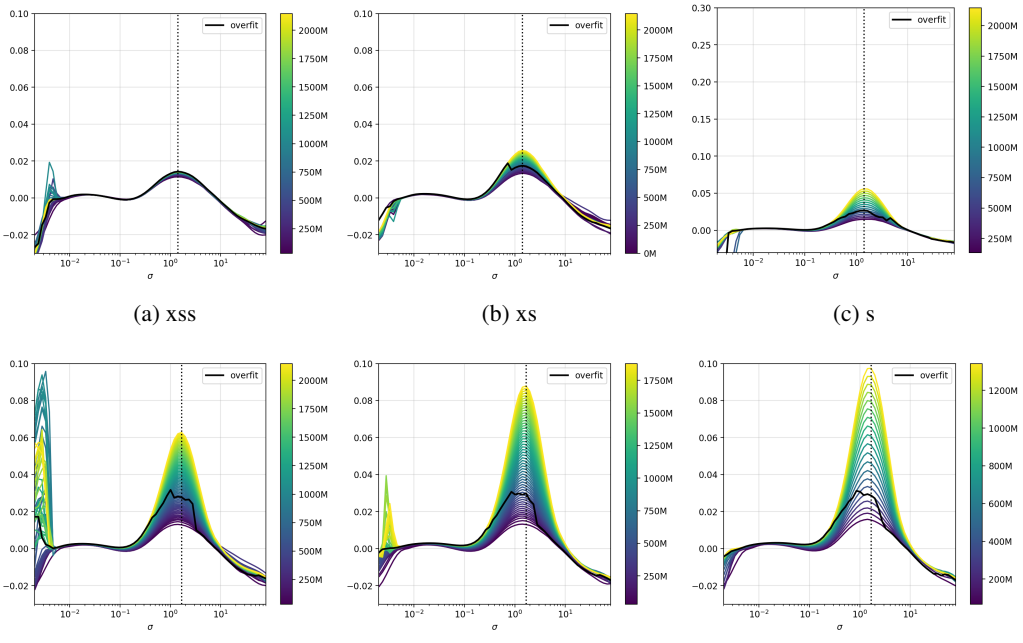
Figure 13: Relative generalization gap (Eq. (1)) with EDM2 on ImageNet-64 for various model sizes, plotted vs σ. Colorbar shows images seen during training. Dotted black lines indicate σ values used in Fig. 14.



(a) xs-uncond (b) xs

Figure 14: Relative generalization gap (Eq. (1)) and training and validation results, with EDM2 on ImageNet-64 for various model sizes. For each model size, σ is fixed at the peak of the generalization gap in Fig. 13.

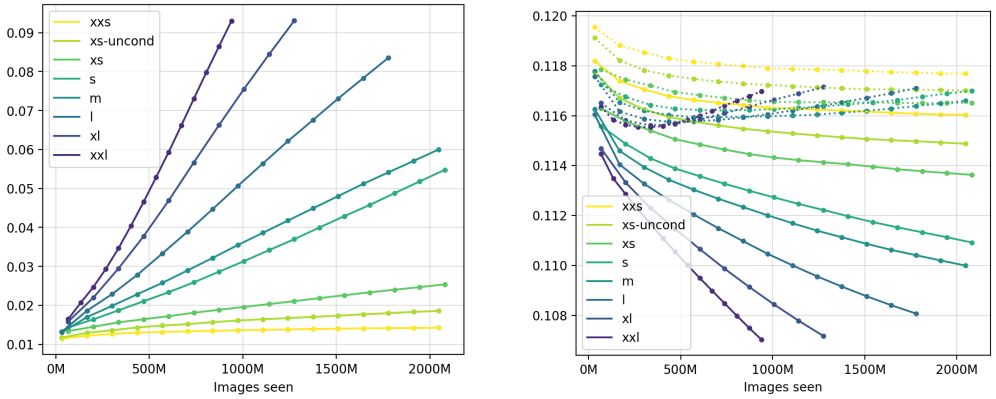
F.2 IMAGENET-512



(a) xss (b) xs (c) s

(d) m (e) l (f) xl

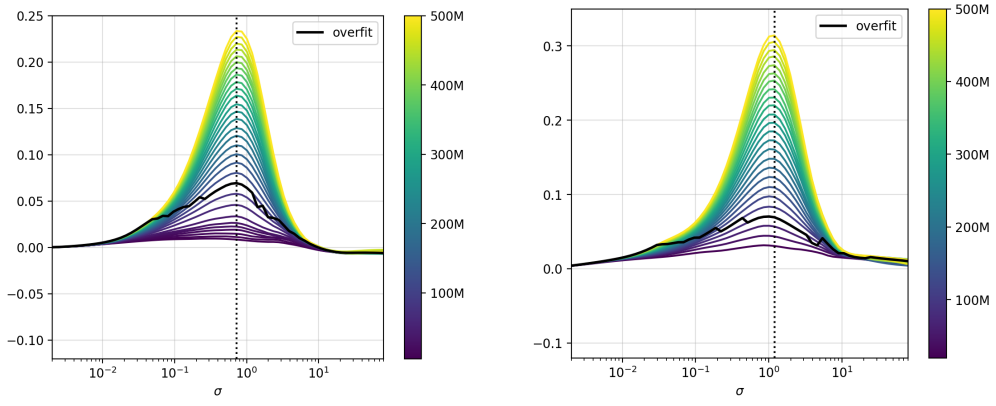
Figure 15: Relative generalization gap (Eq. (1)) with EDM2 on ImageNet-512 for various model sizes, plotted against the noise level σ. Colorbar shows images seen during training. Dotted black lines indicate σ values used in Fig. 16.



(a) xs-uncond (b) xs

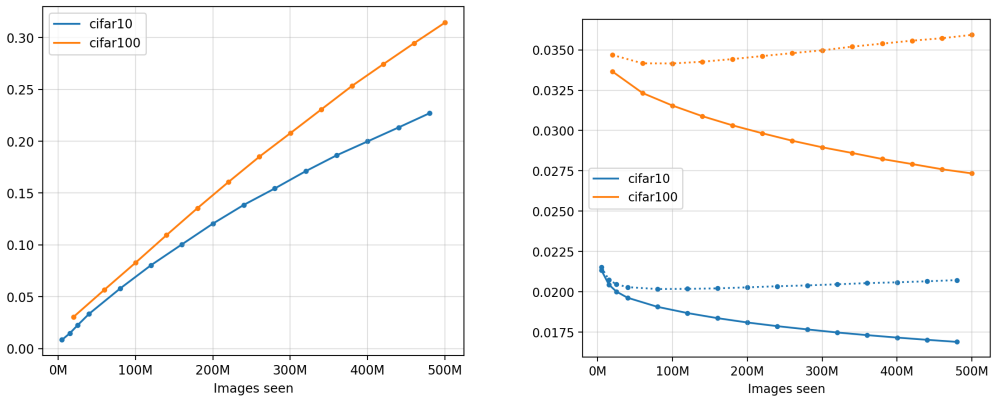
Figure 16: Relative generalization gap (Eq. (1)) and training and validation results, with EDM2 on ImageNet-64 for various model sizes. For each model size, σ is fixed at the peak of the generalization gap in Fig. 15.

F.3 CIFAR-10/100



(a) CIFAR-10 (b) CIFAR-100

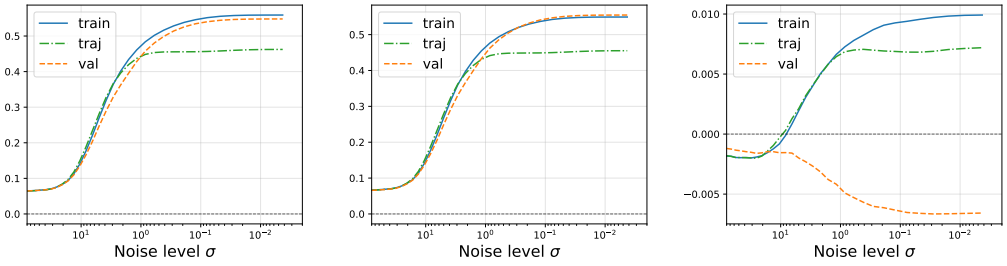
Figure 17: Relative generalization gap (Eq. (1)) with EDM on CIFAR-10/100. Colorbar shows images seen during training. Dotted black lines indicate σ values used in Fig. 18



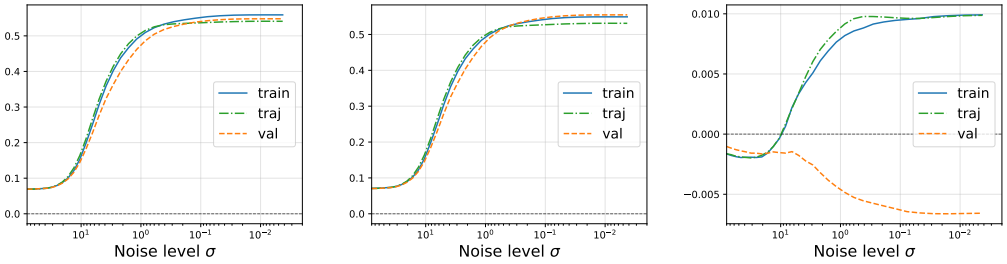
(a) Rel. gen-gap (b) Train/val results

Figure 18: Relative generalization gap (Eq. (1)) and training and validation results with EDM on CIFAR-10/100 for model sizes. σ is fixed at the peak of the generalization gap, see Fig. 17.

G ADDITIONAL RESULTS FOR THE SIMILARITY METRICS

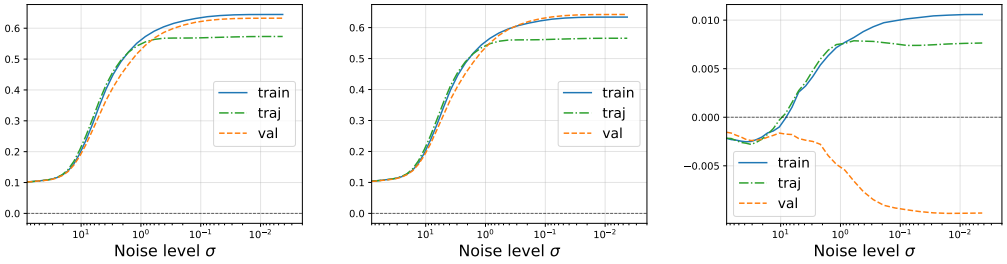


(a) $S_{20}(Q, \mathcal{D}_{\text{train}}; \sigma)$ (b) $S_{20}(Q, \mathcal{D}_{\text{train}}; \sigma)$ (c) $B_{20}(Q; \sigma)$

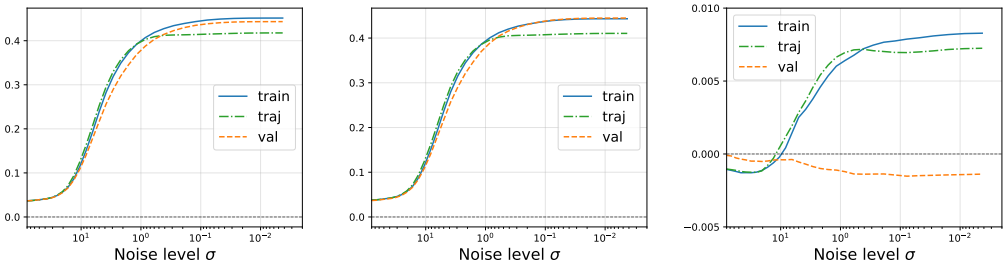


(d) $S_{20}(Q, \mathcal{D}_{\text{train}}; \sigma)$ (e) $S_{20}(Q, \mathcal{D}_{\text{val}}; \sigma)$ (f) $B_{20}(Q; \sigma)$

Figure 19: The gap between the absolute similarity to the references between predictions from intermediate states of trajectories (green line) and predictions from noised data points (blue and orange lines) closes with increasing model size. All plots show fully trained models. **First column:** EDM2-XS. **Second column:** EDM2-XL.

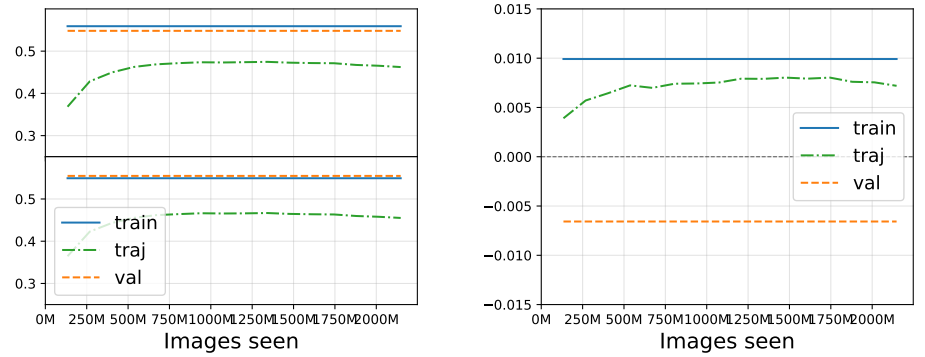


(a) $S_5(Q, \mathcal{D}_{\text{train}}; \sigma)$ (b) $S_5(Q, \mathcal{D}_{\text{val}}; \sigma)$ (c) $B_5(Q; \sigma)$

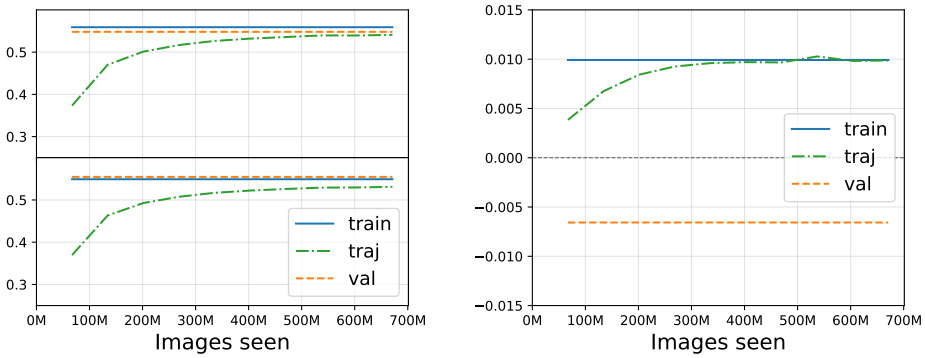


(d) $S_{45}(Q, \mathcal{D}_{\text{train}}; \sigma)$ (e) $S_{45}(Q, \mathcal{D}_{\text{val}}; \sigma)$ (f) $B_{45}(Q; \sigma)$

Figure 20: Qualitative trends in the metrics of Section 4 are robust to the choice of k. All plots show fully trained models.



(a) $S_{20}(Q, \mathcal{D}_{\text{train}}; \sigma)$ (b) $S_{20}(Q, \mathcal{D}_{\text{train}}; \sigma)$



(c) $S_{20}(Q, \mathcal{D}_{\text{train}}; \sigma)$ (d) $S_{20}(Q, \mathcal{D}_{\text{train}}; \sigma)$

Figure 21: The gap between the absolute similarity to the references between predictions from intermediate states of trajectories (green line) and predictions from noised data points (blue and orange lines) closes with increasing model size. All plots show fully trained models. **First column:** EDM2-XS. **Second column:** EDM2-XL.