How Diffusion Models Memorize

Juyeop Kim **Songkuk Kim*** **Jong-Seok Lee***
School of Integrated Technology & BK21 Graduate Program in Intelligent Semiconductor Technology, Yonsei University, Republic of Korea
{juyeopkim, songkuk, jong-seok.lee}@yonsei.ac.kr

Abstract

Despite their success in image generation, diffusion models can unintentionally memorize and reproduce training data, raising serious privacy and copyright concerns. Although prior work has studied memorization from different perspectives, a unified account of how this reproduction occurs during generation remains incomplete. In this paper, we develop a mathematical account of this process in text-to-image latent diffusion models with classifier-free guidance. We trace reproduction across denoising steps through predictions of the final denoised latent made with and without the prompt. Assuming these predictions are Bayes-optimal and uniformly bounded, we show that sufficiently low loss on a training prompt–image pair enables the prediction made with the prompt to reproduce the paired training latent with high probability, even at the first denoising step. At this first step, we bound the difference between the prediction made without the prompt and the empirical mean of the training latents, showing how guidance amplifies the reproduced training latent's deviation from that mean. We then quantify how the latent passed between denoising steps carries prompt influence to subsequent predictions made without the prompt. When both predictions are sufficiently close to the paired training latent at the final step, their guided combination also reproduces it in the final generated latent. We also analyze how their difference evolves, showing that closer agreement alone does not establish reproduction. Finally, our account provides a unified explanation of existing detection and mitigation strategies.

1 Introduction

Text-to-image (T2I) diffusion models have emerged as a dominant paradigm for image generation [Starodubcev et al., 2024], allowing users to generate high-quality images by specifying their desired content through text prompts [Ramesh et al., 2022, Rombach et al., 2022, Nichol et al., 2022]. While the models have surpassed earlier approaches in image quality and prompt–image alignment [Nichol et al., 2022, Saharia et al., 2022], they can also exhibit unintended *memorization* of training prompt–image pairs, generating near-identical copies of training images in response to their paired prompts even across different random seeds [Webster, 2023, Somepalli et al., 2023a, Carlini et al., 2023, Wen et al., 2024, Hintersdorf et al., 2024]. This behavior raises serious privacy and copyright concerns, as it can expose sensitive information or reproduce copyrighted content [Carlini et al., 2023, Jiang et al., 2023].

To address this issue, prior work has sought to understand memorization from different perspectives [Somepalli et al., 2023b, Gu et al., 2025, Baptista et al., 2025, Ross et al., 2025, Jeon et al., 2025]. Yet a unified account of how a model reproduces a training image from its paired prompt remains incomplete. Without such an account, we lack a common framework for explaining the effectiveness of existing detection and mitigation strategies, which have been motivated by intuition and empirical observations of model behavior [Chen et al., 2024, Wen et al., 2024,

*Corresponding authors.

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

Ren et al., 2024, Hintersdorf et al., 2024, Jain et al., 2025]. Their empirical success alone does not establish whether the same model behaviors remain reliable indicators of memorization or whether the same mitigation strategies remain effective beyond the settings in which they were evaluated. Connecting these model behaviors to training-image reproduction provides a basis for assessing the reliability and limitations of existing strategies and guiding their adaptation to new models and generation settings.

In this paper, we develop a mathematical account of how memorization occurs during generation in T2I diffusion models. Specifically, assuming Bayes-optimal and uniformly bounded model predictions, we formalize how information learned from a training prompt–image pair can lead to reproduction of the training image from its paired prompt at inference time, focusing on widely used latent diffusion models [Rombach et al., 2022, Li et al., 2025] with classifier-free guidance (CFG) [Ho and Salimans, 2021, Meng et al., 2023]. Our contributions are as follows:

1. We trace memorization through the first, subsequent, and final denoising steps using *clean predictions*, which estimate the final denoised latent representation of the output image. At the first step, we derive a lower bound on the probability that the clean prediction from CFG's *conditional branch*, which receives the prompt, reproduces the paired training latent. This bound yields a threshold on the pair's expected training loss determined by the noise schedule: **lower loss strengthens the guarantee of reproduction even at the first denoising step**. For CFG's *unconditional branch*, which receives no prompt, we bound the difference between its clean prediction and the empirical mean of the training latents using the noise schedule and dataset statistics. Together, these findings establish that **CFG amplifies the reproduced training latent's deviation from the empirical mean by the guidance scale**, up to the derived error bounds.
2. At subsequent denoising steps, we derive a condition for continued reproduction of the paired training latent by the conditional branch and bound how a CFG update changes the unconditional branch's clean prediction at the next step through the latent produced by that update. This latent carries the prompt's influence from the preceding step and is supplied to both branches. Thus, **even though the unconditional branch receives no prompt, it can be influenced by the prompt through its input**, explaining how the initial amplification can affect predictions during subsequent denoising steps.
3. At the final denoising step, we bound the difference between the final generated latent and the paired training latent using the clean-prediction errors of both branches and the guidance scale. This establishes that **if both branches reproduce the paired training latent sufficiently accurately, their CFG combination also reproduces it in the final generated latent**. Together with the preceding results, this connects the pair's training loss and its influence during denoising to the conditions for final reproduction.
4. We further examine whether agreement between the branches indicates memorization. In an idealized model, we characterize when the gap between the probabilities the branches assign to the paired training latent widens or narrows as their shared input increasingly favors that latent. We connect this probability gap to the norm of the difference between the branches' clean predictions and bound how contributions from other training latents affect this norm. These results show that **closer agreement alone does not establish reproduction of the paired training latent**.
5. Finally, we use our account to explain existing detection and mitigation strategies within a common mathematical framework. We connect the model behaviors used for detection and the changes introduced for mitigation to how the two branches' clean predictions lead to reproduction of the paired training latent. This provides **a unified mathematical explanation of why these strategies work**, linking prior empirical findings to the mechanisms established by our analysis.

2 Preliminaries

Forward and reverse processes. Diffusion models consist of a forward process that gradually adds noise to data and a reverse process that generates samples by denoising [Sohl-Dickstein et al., 2015, Ho et al., 2020]. For computational efficiency, diffusion is often performed in the lower-dimensional latent space $\mathbb{R}^d$ of a pretrained autoencoder [Rombach et al., 2022]. Let $\beta_t \in (0,1)$ denote the noise schedule, and define $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$. Given a clean latent $\mathbf{x} \sim q$, the forward process adds noise over T steps, with the distribution at timestep t given by

$$q(\mathbf{x}_t \mid \mathbf{x}) = \mathcal{N}\big(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\,\mathbf{x}, (1-\bar{\alpha}_t)\mathbf{I}\big)\,. \tag{1}$$

Equivalently, the forward latent $\mathbf{x}_t$ can be sampled through reparameterization as

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{x} + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{2}$$

The schedule decreases $\bar{\alpha}_t$ from $\bar{\alpha}_0 = 1$ toward zero. The reverse process generates $\mathbf{x}_0$ by denoising $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Skipping steps can accelerate sampling [Song et al., 2021].

Training and inference. The denoising network is trained with the per-sample loss [Ho et al., 2020]

$$\mathcal{L}_c(\mathbf{x}, t) := \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_c\big(\sqrt{\bar{\alpha}_t}\,\mathbf{x} + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}, t\big) \right\|_2^2, \tag{3}$$

where $\boldsymbol{\epsilon}_c(\mathbf{z}, t)$ denotes the denoising network's noise prediction for input $\mathbf{z}$ at timestep t, conditioned on the text prompt c. Training minimizes the expectation of this loss over training pairs, timesteps, and forward noise. For a fixed pair $(\mathbf{x}, c)$ and timestep t, the expected loss is $\mathbb{E}_{\boldsymbol{\epsilon}}[\mathcal{L}_c(\mathbf{x}, t)]$. During training, the prompt c is randomly omitted (denoted by $\emptyset$), enabling the network to also learn the unconditional noise prediction $\boldsymbol{\epsilon}_\emptyset$ through the analogous loss $\mathcal{L}_\emptyset(\mathbf{x}, t)$ [Ho and Salimans, 2021].

For either the conditional branch ($b = c$) or the unconditional branch ($b = \emptyset$), the *clean prediction* at timestep t estimates the final denoised latent and is defined by

$$\hat{\mathbf{x}}_{t,b}(\mathbf{z}) := \frac{\mathbf{z} - \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}_b(\mathbf{z}, t)}{\sqrt{\bar{\alpha}_t}}. \tag{4}$$

Below, we write $\hat{\mathbf{x}}_{t,b} := \hat{\mathbf{x}}_{t,b}(\mathbf{x}_t)$ for notational simplicity. At inference, classifier-free guidance (CFG) [Ho and Salimans, 2021] combines the two noise predictions as

$$\boldsymbol{\epsilon}_g(\mathbf{x}_t, t, c) = g\,\boldsymbol{\epsilon}_c(\mathbf{x}_t, t) + (1-g)\,\boldsymbol{\epsilon}_\emptyset(\mathbf{x}_t, t), \tag{5}$$

where $g > 1$ is the guidance scale. The corresponding *guided clean prediction* is

$$\hat{\mathbf{x}}_{t,g} = \frac{\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}_g(\mathbf{x}_t, t, c)}{\sqrt{\bar{\alpha}_t}} = g\,\hat{\mathbf{x}}_{t,c} + (1-g)\,\hat{\mathbf{x}}_{t,\emptyset}. \tag{6}$$

The following sampling update shows how the guided clean prediction affects the latent at the next denoising step [Song et al., 2021][1]

$$\mathbf{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\,\hat{\mathbf{x}}_{t,g} + \sqrt{1-\bar{\alpha}_{t-1}}\,\boldsymbol{\epsilon}_g(\mathbf{x}_t, t, c). \tag{7}$$

3 How diffusion models memorize

We say that a diffusion model has memorized a training latent $\mathbf{x}^\star$ (the *target*) when its final generated latent $\mathbf{x}_0$ reproduces $\mathbf{x}^\star$ from its paired prompt c. Since $\bar{\alpha}_0 = 1$, the sampling update in Eq. 7 gives $\mathbf{x}_0 = \hat{\mathbf{x}}_{1,g}$. The final sample therefore depends on the CFG combination of the clean predictions from the conditional and unconditional branches (Eq. 6), motivating us to examine how both branches contribute to target reproduction.[2] At subsequent denoising steps, the branches' shared input $\mathbf{x}_t$ is shaped by earlier CFG updates, complicating the analysis of their clean predictions. We therefore begin with the simplest case: the first denoising step, $t = T$, where the input $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is pure Gaussian noise. We then extend the analysis to subsequent denoising steps and final reproduction, completing our account of how memorization occurs during generation. Let $\mathbf{X} \sim q$ denote a clean latent with mean $\boldsymbol{\mu} := \mathbb{E}_{\mathbf{X}\sim q}[\mathbf{X}]$, and let $\mathbf{X}_t$ denote its forward latent in Eq. 2.

Bayes-optimal noise prediction. Following analyses based on optimal denoisers [Kong et al., 2023, Daras et al., 2023, Kong et al., 2024], we assume that both branches make Bayes-optimal noise predictions. These predictions minimize expected squared error by taking the conditional mean of the injected noise:

$$\boldsymbol{\epsilon}_b(\mathbf{z}, t) = \mathbb{E}_{\boldsymbol{\epsilon}}[\boldsymbol{\epsilon} \mid \mathbf{X}_t = \mathbf{z}, t, b], \qquad b \in \{c, \emptyset\}. \tag{8}$$

Under this assumption, the clean predictions in Eq. 4 equal the corresponding posterior means:

$$\hat{\mathbf{x}}_{t,b}(\mathbf{z}) = \mathbb{E}_{\mathbf{X}}[\mathbf{X} \mid \mathbf{X}_t = \mathbf{z}, t, b]. \tag{9}$$

[1] While the displayed DDIM update is deterministic, the analysis extends to stochastic updates affine in the clean prediction, including DDPM [Ho et al., 2020] (Fig. 2), since the added noise terms cancel when both updates use the same sampled noise.

[2] Reproduction of the target by either branch alone does not guarantee memorization, since CFG combines both clean predictions.

During sampling, each branch's clean prediction is this posterior mean evaluated at the current realized latent $\mathbf{z} = \mathbf{x}_t$.

Uniformly bounded clean predictions. We further assume that the clean predictions are uniformly bounded over the timesteps and inputs considered:

$$\|\hat{\mathbf{x}}_{t,b}(\mathbf{z})\|_2 \le M, \qquad b \in \{c, \emptyset\}, \tag{10}$$

for some finite M. By Eq. 9 and Jensen's inequality, this holds if $\|\mathbf{X}\|_2 \le M$ almost surely. For a finite empirical distribution over fixed training latents, M can be chosen as their maximum norm.

Throughout this section and its proofs, we fix the data distribution q, the training pair $(\mathbf{x}^\star, c)$, the latent dimension d, the guidance scale g, and the bound M, unless stated otherwise.

3.1 Initial reproduction and amplification

We first relate the pair's expected training loss to reproduction of the target by the conditional clean prediction at the first denoising step. Since the loss in Eq. 3 is evaluated on forward latents, we first obtain a bound for the target's forward latent and then extend it to Gaussian initialization.

Lemma 1 (Initial conditional reproduction; proof in App. A.1)**.** *Let $\mathbf{z} = \sqrt{\bar{\alpha}_T}\mathbf{x}^\star + \sqrt{1-\bar{\alpha}_T}\boldsymbol{\epsilon}$ be a forward latent of the target at $t = T$, with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. If $\mathbb{E}_{\boldsymbol{\epsilon}}[\mathcal{L}_c(\mathbf{x}^\star, T)] \le \varepsilon_T$, then, for every tolerance $r > 0$,*

$$\Pr\big(\|\hat{\mathbf{x}}_{T,c}(\mathbf{z}) - \mathbf{x}^\star\|_2 \le r\big) \ge 1 - \frac{1-\bar{\alpha}_T}{\bar{\alpha}_T}\frac{\varepsilon_T}{r^2}.$$

The bound also holds at Gaussian initialization $\mathbf{z} = \mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ up to $o(1)$ as $\bar{\alpha}_T \to 0$.

Thus, for sufficiently small $\bar{\alpha}_T$, a small $(1-\bar{\alpha}_T)\varepsilon_T/\bar{\alpha}_T$ allows target reproduction even at the first denoising step. The unconditional branch, however, receives Gaussian noise without the paired prompt to identify the target. Its clean prediction averages over clean latents (Eq. 9) and need not reproduce the target. At Gaussian initialization, this posterior mean satisfies the following bound.

Lemma 2 (Initial unconditional prediction; proof in App. A.2)**.** *Assume Eq. 8 and $\|\mathbf{X}\|_2 \le M$ q-almost surely. For each realized initial latent $\mathbf{x}_T$, the unconditional clean prediction satisfies*

$$\big\|\hat{\mathbf{x}}_{T,\emptyset} - \boldsymbol{\mu}\big\|_2 \le M\left(e^{2\kappa_T} - 1\right), \qquad \kappa_T := \frac{\sqrt{\bar{\alpha}_T}M\|\mathbf{x}_T\|_2 + \frac{1}{2}\bar{\alpha}_T M^2}{1-\bar{\alpha}_T}.$$

For fixed $\mathbf{x}_T$, this bound goes to zero as $\bar{\alpha}_T \to 0$.

Thus, for each fixed $\mathbf{x}_T$, the unconditional clean prediction approaches the data mean $\boldsymbol{\mu}$ as $\bar{\alpha}_T \to 0$. Combining this bound with conditional target reproduction yields the following guarantee for the guided clean prediction.

Theorem 3 (Initial CFG amplification; proof in App. A.3)**.** *Assume the conditions of Lem. 2 and $\mathbb{E}_{\boldsymbol{\epsilon}}[\mathcal{L}_c(\mathbf{x}^\star, T)] \le \varepsilon_T$. For $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $g > 1$, and any tolerance $r > 0$,*

$$\Pr\big(\|\hat{\mathbf{x}}_{T,g} - [\boldsymbol{\mu} + g(\mathbf{x}^\star - \boldsymbol{\mu})]\|_2 > r\big) \le g^2\frac{1-\bar{\alpha}_T}{\bar{\alpha}_T}\frac{\varepsilon_T}{r^2} + o(1),$$

as $\bar{\alpha}_T \to 0$.

When the probability bound is small, the guided clean prediction is close to $\boldsymbol{\mu} + g(\mathbf{x}^\star - \boldsymbol{\mu})$ with high probability. Thus, CFG amplifies the target's deviation from the data mean by the guidance scale g, up to the derived error bounds. This guided clean prediction enters the sampling update in Eq. 7, shaping the latent passed to both branches at the next denoising step.

3.2 Propagation of prompt influence

We next examine how conditional reproduction and CFG amplification at the first denoising step can affect subsequent denoising steps. The conditional branch continues to receive the paired prompt c, but its clean prediction also depends on the current latent $\mathbf{x}_t$, which has been shaped by earlier CFG updates. Applying the algebraic relation underlying Lem. 1 at this latent gives the following condition for continued target reproduction.

Proposition 4 (Subsequent conditional reproduction; proof in App. A.4)**.** *Fix $t \in \{1, \dots, T\}$, a realized sampling latent $\mathbf{x}_t$, and a tolerance $r \geq 0$. Then $\|\hat{\mathbf{x}}_{t,c} - \mathbf{x}^\star\|_2 \leq r$ if and only if*

$$\left\| \frac{\mathbf{x}_t - \sqrt{\bar{\alpha}_t}\mathbf{x}^\star}{\sqrt{1-\bar{\alpha}_t}} - \boldsymbol{\epsilon}_c(\mathbf{x}_t, t) \right\|_2 \leq r\sqrt{\frac{\bar{\alpha}_t}{1-\bar{\alpha}_t}}.$$

The unconditional branch, in contrast, receives no prompt. Yet both branches receive the same latent $\mathbf{x}_t$, which has been shaped by earlier CFG updates. Thus, the prompt can still influence the unconditional branch's clean prediction indirectly through the shared input $\mathbf{x}_t$. To quantify this indirect influence, we examine how the conditional clean prediction used in a CFG update can affect the unconditional clean prediction at the next step.

For $t \in \{2, \dots, T\}$, eliminating $\boldsymbol{\epsilon}_g$ from Eq. 7 using Eq. 6 gives the CFG update below:

$$\mathbf{x}_{t-1} = \gamma_t \mathbf{x}_t + \eta_t \hat{\mathbf{x}}_{t,g}, \tag{11}$$

where $\gamma_t := \sqrt{(1-\bar{\alpha}_{t-1})/(1-\bar{\alpha}_t)}$ and $\eta_t := \sqrt{\bar{\alpha}_{t-1}} - \gamma_t\sqrt{\bar{\alpha}_t} > 0$. We compare it with an update using only the unconditional clean prediction at the same realized latent $\mathbf{x}_t$:

$$\mathbf{x}^\emptyset_{t-1} := \gamma_t \mathbf{x}_t + \eta_t \hat{\mathbf{x}}_{t,\emptyset}. \tag{12}$$

The sampling coefficients γ_t and η_t determine the contributions of the current latent and the clean prediction, respectively. Both updates share the same input $\mathbf{x}_t$ and therefore retain the effects of earlier CFG updates. Their difference isolates the contribution of the conditional clean prediction at the current step. Defining the difference between the branches' clean predictions as $\boldsymbol{\Delta}_t := \hat{\mathbf{x}}_{t,c} - \hat{\mathbf{x}}_{t,\emptyset}$, Eq. 6 gives

$$\mathbf{x}_{t-1} - \mathbf{x}^\emptyset_{t-1} = g\eta_t \boldsymbol{\Delta}_t. \tag{13}$$

When the conditional clean prediction exactly reproduces the target, the update difference is $g\eta_t(\mathbf{x}^\star - \hat{\mathbf{x}}_{t,\emptyset})$, showing how the reproduced target enters the next input to both branches.

We now compare the unconditional clean predictions at timestep $t-1$ for the two inputs $\mathbf{x}_{t-1}$ and $\mathbf{x}^\emptyset_{t-1}$ produced by the CFG and unconditional updates, respectively. To quantify the difference between these predictions, we consider the line segment joining their inputs:

$$\mathbf{z}_t(s) := \mathbf{x}^\emptyset_{t-1} + sg\eta_t \boldsymbol{\Delta}_t, \qquad s \in [0,1]. \tag{14}$$

Its endpoints are $\mathbf{z}_t(0) = \mathbf{x}^\emptyset_{t-1}$ and $\mathbf{z}_t(1) = \mathbf{x}_{t-1}$. Then, the Jacobian $\mathbf{J}_{t-1,\emptyset}(\mathbf{z}) := D_{\mathbf{z}}\hat{\mathbf{x}}_{t-1,\emptyset}(\mathbf{z})$ describes how the unconditional clean prediction responds to changes in its input. Integrating this Jacobian along the segment gives the following exact identity for the difference between the two clean predictions and a bound on its norm.

Theorem 5 (Propagation of prompt influence; proof in App. A.5)**.** *For $t \in \{2, \dots, T\}$, suppose $\hat{\mathbf{x}}_{t-1,\emptyset}(\cdot)$ is continuously differentiable on a neighborhood of the segment in Eq. 14. Then*

$$\hat{\mathbf{x}}_{t-1,\emptyset}(\mathbf{x}_{t-1}) - \hat{\mathbf{x}}_{t-1,\emptyset}(\mathbf{x}^\emptyset_{t-1}) = g\eta_t \int_0^1 \mathbf{J}_{t-1,\emptyset}(\mathbf{z}_t(s))\boldsymbol{\Delta}_t \, ds,$$

and consequently

$$\left\| \hat{\mathbf{x}}_{t-1,\emptyset}(\mathbf{x}_{t-1}) - \hat{\mathbf{x}}_{t-1,\emptyset}(\mathbf{x}^\emptyset_{t-1}) \right\|_2 \leq g\eta_t \sup_{s\in[0,1]} \left\| \mathbf{J}_{t-1,\emptyset}(\mathbf{z}_t(s)) \right\|_{\mathrm{op}} \|\boldsymbol{\Delta}_t\|_2.$$

This result quantifies how the prompt's contribution to a CFG update can affect the next unconditional clean prediction through the latent supplied as input, even though the unconditional branch receives no prompt. At timestep $t-1$, this unconditional clean prediction enters CFG again, so the indirect prompt influence can also affect the guided clean prediction.

3.3 Final reproduction

We now relate the clean predictions at the final denoising step to reproduction of the target in the final generated latent. Using $\mathbf{x}_0 = \hat{\mathbf{x}}_{1,g}$ and Eq. 6 at the realized latent $\mathbf{x}_1$, we obtain

$$\mathbf{x}_0 - \mathbf{x}^\star = \left(\hat{\mathbf{x}}_{1,c} - \mathbf{x}^\star\right) + (g-1)\boldsymbol{\Delta}_1, \tag{15}$$

where $\boldsymbol{\Delta}_1 := \hat{\mathbf{x}}_{1,c} - \hat{\mathbf{x}}_{1,\emptyset}$. Thus, final reproduction depends on the conditional clean prediction's error and the difference between the branches' clean predictions. Under Eq. 8, both clean predictions equal the corresponding posterior means in Eq. 9 evaluated at $\mathbf{x}_1$. To quantify the target's contribution to these means, we define the *target posterior probability* at timestep $t \in \{1, \dots, T\}$ and input $\mathbf{z} \in \mathbb{R}^d$ by

$$\pi_{t,b}(\mathbf{z}) := \Pr(\mathbf{X} = \mathbf{x}^\star \mid \mathbf{X}_t = \mathbf{z}, t, b)\,, \qquad b \in \{c, \emptyset\}. \tag{16}$$

At $t = 1$ and $\mathbf{z} = \mathbf{x}_1$, the law of total expectation gives

$$\begin{aligned} \hat{\mathbf{x}}_{1,b} = \pi_{1,b}(\mathbf{x}_1)\mathbf{x}^\star \\ + \big[1 - \pi_{1,b}(\mathbf{x}_1)\big]\mathbb{E}_{\mathbf{X}}[\mathbf{X} \mid \mathbf{X}_1 = \mathbf{x}_1, 1, b, \mathbf{X} \neq \mathbf{x}^\star]\,. \end{aligned} \tag{17}$$

The second term in Eq. 17 represents the contribution of other clean latents to the clean prediction and is interpreted as zero when $\pi_{1,b}(\mathbf{x}_1) = 1$, since these latents then have zero posterior probability. The following theorem uses this decomposition and CFG to bound the final reproduction error, $\|\mathbf{x}_0 - \mathbf{x}^\star\|_2$.

Theorem 6 (Final reproduction; proof in App. A.6)**.** *Assume Eq. 8 and $\|\mathbf{X}\|_2 \leq M$ almost surely under the conditional and unconditional data distributions. For each realized latent $\mathbf{x}_1$ at the final denoising step and $g > 1$, the final generated latent satisfies*

$$\|\mathbf{x}_0 - \mathbf{x}^\star\|_2 \leq \big(M + \|\mathbf{x}^\star\|_2\big)\Big[g\big(1 - \pi_{1,c}(\mathbf{x}_1)\big) + (g-1)\big(1 - \pi_{1,\emptyset}(\mathbf{x}_1)\big)\Big].$$

Thus, sufficiently high target posterior probabilities in both branches ensure that the final generated latent is close to the target.

3.4 Memorization dynamics

Since sufficiently accurate target reproduction by both branches ensures final reproduction, one might expect earlier or closer agreement between their clean predictions to indicate stronger memorization. We investigate this intuition by studying $\boldsymbol{\Delta}_t$ during denoising, which under Eq. 8 is the difference between the conditional and unconditional posterior means. We therefore call $\boldsymbol{\Delta}_t$ the *posterior lag*, motivated by the possibility that the conditional clean prediction is already close to the target while the unconditional clean prediction remains farther away. Accordingly, the *posterior lag magnitude*, $\|\boldsymbol{\Delta}_t\|_2$, measures the separation between the branches' clean predictions. This analysis clarifies how to interpret posterior lag trajectories and provides a mathematical basis for developing detection and mitigation strategies.

The paired prompt can give the target a higher prior probability than it has without the prompt. We express this advantage through the *target prior probabilities*:

$$p_b := \Pr(\mathbf{X} = \mathbf{x}^\star \mid b), \qquad b \in \{c, \emptyset\}, \qquad 0 < p_\emptyset < p_c < 1. \tag{18}$$

To isolate the effect of this prior advantage on the target posterior probabilities, we consider an idealized model in which both branches use the same evidence from the current latent. Specifically, after excluding the target, we assume that the average forward likelihood of the current latent is the same for both branches:

$$\mathbb{E}_{\mathbf{X}}[q(\mathbf{x}_t \mid \mathbf{X}) \mid c, \mathbf{X} \neq \mathbf{x}^\star] = \mathbb{E}_{\mathbf{X}}[q(\mathbf{x}_t \mid \mathbf{X}) \mid \emptyset, \mathbf{X} \neq \mathbf{x}^\star]\,. \tag{19}$$

Each expectation averages the Gaussian forward density in Eq. 1 over the corresponding data distribution, excluding the target. Since the target likelihood $q(\mathbf{x}_t \mid \mathbf{x}^\star)$ is also common to both branches, we define their shared *target evidence* by

$$\lambda_t := \log \frac{q(\mathbf{x}_t \mid \mathbf{x}^\star)}{\mathbb{E}_{\mathbf{X}}[q(\mathbf{x}_t \mid \mathbf{X}) \mid \emptyset, \mathbf{X} \neq \mathbf{x}^\star]}. \tag{20}$$

Larger λ_t means that the current latent more strongly favors the target over other clean latents. When $\lambda_t = 0$, the target's forward likelihood equals their average forward likelihood. Bayes' rule gives

$$\pi_{t,b}(\mathbf{x}_t) = \frac{p_b e^{\lambda_t}}{p_b e^{\lambda_t} + 1 - p_b} = \sigma(\lambda_t + \operatorname{logit}(p_b))\,, \qquad b \in \{c, \emptyset\}, \tag{21}$$

where $\sigma(u) := (1 + \exp(-u))^{-1}$ and $\operatorname{logit}(p) := \log(p/(1-p))$ for $p \in (0, 1)$. For the same evidence, the larger conditional prior gives the target a higher posterior probability. The unconditional

branch therefore needs stronger evidence to reach the same target posterior probability. To study how the difference between these probabilities changes with evidence, we define the *posterior probability gap*:

$$G(\lambda) := \sigma(\lambda + \operatorname{logit}(p_c)) - \sigma(\lambda + \operatorname{logit}(p_\emptyset)), \tag{22}$$

so $G(\lambda_t) = \pi_{t,c}(\mathbf{x}_t) - \pi_{t,\emptyset}(\mathbf{x}_t) > 0$. The following lemma characterizes G as evidence increases from its first-step value λ_T, with $\lambda_T = 0$ representing the ideal uninformative case.

Lemma 7 (Posterior probability gap; proof in App. A.7)**.** *Assume $0 < p_\emptyset < p_c < 1$. The posterior probability gap $G(\lambda)$ is unimodal in the target evidence λ, with unique maximizer*

$$\lambda^\star := -\frac{\operatorname{logit}(p_c) + \operatorname{logit}(p_\emptyset)}{2}.$$

On $[\lambda_T, \infty)$, G strictly decreases if $\lambda_T \geq \lambda^\star$. Otherwise, it strictly increases on $[\lambda_T, \lambda^\star]$ and strictly decreases on $[\lambda^\star, \infty)$. When $\lambda_T = 0$, these cases correspond to $p_c + p_\emptyset \geq 1$ and $p_c + p_\emptyset < 1$, respectively.

If target evidence increases during denoising, both branches assign increasing probability to the target. Below $\lambda^\star$, the conditional probability increases faster, widening the gap. Above $\lambda^\star$, the unconditional probability increases faster and catches up, narrowing the gap. The following theorem connects this probability gap to posterior lag by bounding the difference between $\|\boldsymbol{\Delta}_t\|_2$ and $G(\lambda_t)\|\mathbf{x}^\star\|_2$ in terms of contributions from other clean latents.

Theorem 8 (Posterior lag dynamics; proof in App. A.8)**.** *Assume Eq. 8 and $0 < p_\emptyset < p_c < 1$. For each timestep $t \in \{1, \ldots, T\}$ and realized latent $\mathbf{x}_t$ satisfying Eq. 19, the posterior lag and its magnitude satisfy*

$$\begin{aligned} \left| \|\boldsymbol{\Delta}_t\|_2 - G(\lambda_t)\|\mathbf{x}^\star\|_2 \right| &\leq \|\boldsymbol{\Delta}_t - G(\lambda_t)\mathbf{x}^\star\|_2 \\ &\leq \sum_{b \in \{c, \emptyset\}} \left[1 - \pi_{t,b}(\mathbf{x}_t)\right] \left\| \mathbb{E}_{\mathbf{X}}[\mathbf{X} \mid \mathbf{X}_t = \mathbf{x}_t, t, b, \mathbf{X} \neq \mathbf{x}^\star] \right\|_2 . \end{aligned}$$

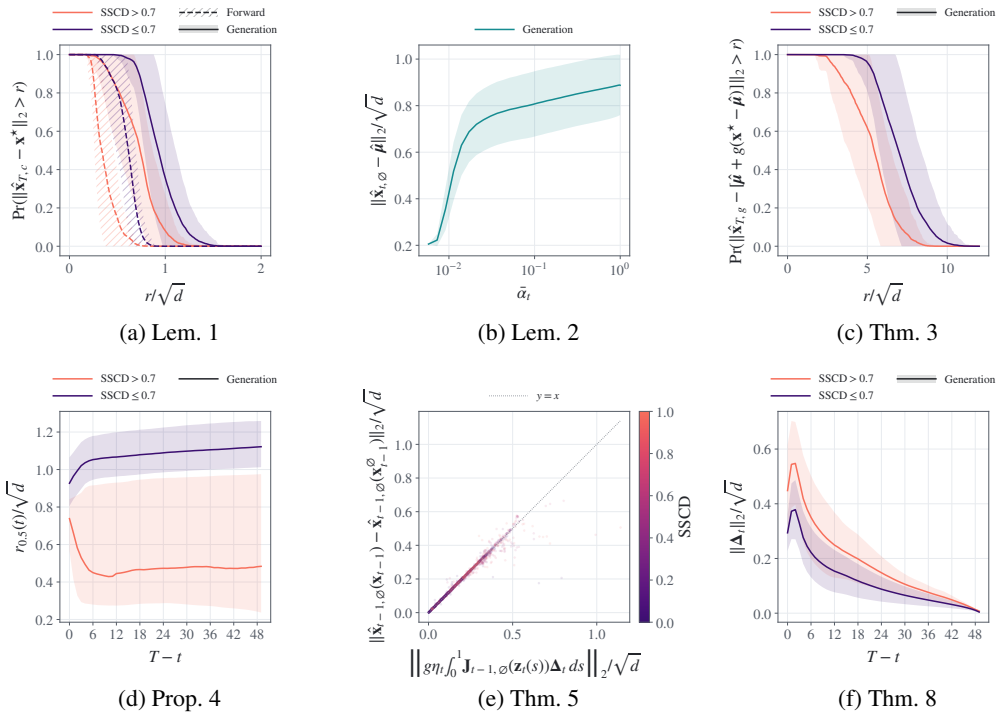
Therefore, small weighted contributions from other clean latents keep the posterior lag magnitude close to the ideal curve $G(\lambda_t)\|\mathbf{x}^\star\|_2$. Since $G(\lambda) \to 0$ as $\lambda \to \pm\infty$, the magnitude can be small when both branches assign either little or nearly all probability to the target. Thus, close agreement alone does not establish target reproduction: the predictions may agree away from the target when target evidence is very low. To understand where they agree, we must account for the target's posterior probability in each branch, which determines its weight in the prediction, and the remaining contributions from other clean latents.

4 Experiments

4.1 Settings

We conduct experiments with Stable Diffusion v1.4 (SDv1) [Rombach et al., 2022], Stable Diffusion v2.1 (SDv2) [Lopez, 2022], and RealisticVision (RealVis) [SG161222, 2023]. We use DDIM [Song et al., 2021] for all three models and additionally evaluate SDv1 with DDPM [Ho et al., 2020]. All configurations use $T = 50$ inference steps and guidance scale $g = 7.5$. We present SDv1 (DDIM) results in the main text and report results for the other configurations in App. C. We use all available paired training images and prompts from Webster [2023], yielding 446 pairs for SDv1 (see App. B for details). For each prompt, we generate 20 RGB images at a resolution of 512×512 pixels from Gaussian initializations $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The same 20 seeds are reused across prompts. All experiments are run on $4\times$ NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs, requiring less than 1.2TB of storage and completing within 12 hours.

Quantifying memorization. We assess reproduction of the paired training image using SSCD [Pizzi et al., 2022], which has shown strong performance in detecting image replication [Somepalli et al., 2023a]. We compute the cosine similarity between the SSCD descriptors of each generated image and its paired training image. Following Wen et al. [2024], we use a threshold of 0.7. For each model–sampler configuration, we classify a prompt as memorized if at least one of its 20 generated images has SSCD > 0.7, and as normal if all scores are ≤ 0.7. For SDv1



(a) Lem. 1 (b) Lem. 2 (c) Thm. 3

(d) Prop. 4 (e) Thm. 5 (f) Thm. 8

Figure 1: **Denoising measurements on SDv1 (DDIM).** Coral and purple in (a,c,d,f) denote memorized and normal prompt groups, respectively (Sec. 4.1). In (a), dashed and solid curves correspond to forward latents and Gaussian initializations, respectively. In (a–c,f), curves show means across prompts, and shaded bands indicate ± 1 standard deviation. Seeds are first averaged within each prompt in (b,f). In (d), curves and shaded bands show the median and interquartile range, respectively, with each prompt receiving equal total weight. In (e), point colors indicate each generated image's final SSCD, and the dotted diagonal denotes equality.

(DDIM), these groups contain 135 and 311 prompts, respectively. All generated samples remain in their prompt's group, and prompt-group summaries give each prompt equal weight.

Measurements. We encode each paired training image with the model's pretrained autoencoder to obtain the target latent $\mathbf{x}^\star$ and record both branches' clean predictions at every denoising step. Their difference gives the posterior lag $\boldsymbol{\Delta}_t$. Displayed Euclidean norms are divided by $\sqrt{d}$, where d is the latent dimension. To compare initial conditional reproduction under forward noise and Gaussian initialization, we additionally evaluate 20 independently sampled forward latents per target at the noise level of the first denoising step. For measurements involving the data mean $\boldsymbol{\mu}$, we use a model-based estimate $\hat{\boldsymbol{\mu}}$, motivated by Eq. 9 and Lem. 2. For each model, we obtain this estimate by averaging 4,096 unconditional clean predictions evaluated on standard Gaussian inputs at the largest timestep used in training. The same estimate is used for the unconditional-prediction and initial CFG measurements. To measure propagation of prompt influence, we compare the CFG and unconditional updates from the same current latent, as in Eqs. 11 and 12. We hold any sampled noise fixed across the two updates and evaluate the next unconditional clean prediction at both resulting latents. We approximate the Jacobian line integral in Thm. 5 using Gauss–Legendre quadrature of Jacobian–vector products.

4.2 Results

Initial conditional reproduction. Fig. 1a reports the probability that the initial conditional clean prediction lies farther than a tolerance r from the target. For both forward latents and Gaussian initializations, memorized prompts have lower failure probabilities than normal prompts at the same tolerance, indicating closer initial conditional predictions to their targets (coral curves below the corresponding purple curves). Gaussian initializations require larger tolerances than forward latents to

achieve comparable failure probabilities (solid curves shifted to the right of the corresponding dashed curves). This separation is consistent with the finite-noise distribution mismatch in Lem. 1: forward latents retain the target-dependent component $\sqrt{\bar{\alpha}_T}\mathbf{x}^\star$ and have noise variance $1 - \bar{\alpha}_T$, whereas Gaussian initializations have zero mean and unit variance. The two input distributions approach each other as $\bar{\alpha}_T \to 0$.

Initial unconditional prediction. Fig. 1b plots the distance between the unconditional clean prediction and the estimated data mean $\hat{\boldsymbol{\mu}}$ against $\bar{\alpha}_t$ along the generated trajectories. The distance is smallest at the first denoising steps, where $\bar{\alpha}_t$ is smallest, and increases as denoising proceeds (teal curve rising from left to right). In Lem. 2, decreasing $\bar{\alpha}_T$ at fixed $\mathbf{x}_T$ decreases the bound on the distance to the data mean, with the bound going to zero as $\bar{\alpha}_T \to 0$. The observed proximity to $\hat{\boldsymbol{\mu}}$ at small $\bar{\alpha}_t$ agrees with the initial unconditional behavior characterized by the lemma.

Initial CFG amplification. Fig. 1c measures the probability that the initial guided clean prediction deviates by more than r from $\hat{\boldsymbol{\mu}} + g(\mathbf{x}^\star - \hat{\boldsymbol{\mu}})$. Memorized prompts have lower failure probabilities at the same tolerance, indicating closer agreement with this extrapolated target (coral curve shifted to the left of the purple curve). This agrees with Thm. 3: when the conditional clean prediction is close to the target and the unconditional clean prediction is close to the data mean, CFG amplifies the target's deviation from that mean by g. The stronger agreement for memorized prompts connects their initial conditional reproduction to the predicted CFG amplification.

Subsequent conditional reproduction. Fig. 1d tracks the median reproduction tolerance $r_{0.5}(t)$ under equal prompt weighting. A smaller tolerance means that at least half of the samples under this weighting have conditional clean predictions within a smaller distance of their paired targets. For memorized prompts, the tolerance decreases sharply during the first few updates and remains substantially below that of normal prompts throughout denoising (coral curve). Normal prompts instead exhibit an increasing median tolerance (purple curve). By Prop. 4, this lower median error corresponds to a smaller median discrepancy in the noise-prediction condition for target reproduction at the same timestep. This separation shows that memorized prompts retain more accurate conditional target reproduction throughout denoising.

Propagation of prompt influence. Fig. 1e compares the norm of the numerically integrated Jacobian response with the norm of the matched change in the next unconditional clean prediction. Most points concentrate near the dotted equality line, showing agreement between the two quantities in Thm. 5. Larger responses tend to be associated with higher final SSCD, indicating stronger prompt influence on the next unconditional clean prediction along trajectories that more closely reproduce the target (coral points extending farther from the origin). This matched comparison supports the propagation mechanism: the conditional contribution changes the sampling latent, which in turn changes the unconditional clean prediction at the next step.

Memorization dynamics. Fig. 1f shows the mean posterior lag magnitude $\|\boldsymbol{\Delta}_t\|_2$ for memorized and normal prompts throughout denoising. Both mean trajectories exhibit a single early peak followed by a gradual decrease, with memorized prompts maintaining a larger separation between the branches over most of the trajectory (coral curve above the purple curve). The larger lag for memorized prompts is consistent with their more accurate conditional target reproduction (Figs. 1a and 1d), while the unconditional clean prediction initially remains near the data mean (Fig. 1b). The paired prompt supplies information about the target directly to the conditional branch, whereas its influence reaches the unconditional branch through the guided sampling latent (Thm. 5). These observations are consistent with conditional target reproduction preceding comparable unconditional target reproduction, producing a larger separation between the two clean predictions. The single-peaked mean trajectories are qualitatively consistent with the unimodal posterior probability gap in Lem. 7 and its contribution to the posterior lag in Thm. 8. Under the shared-evidence model, when target evidence increases through the maximizer of this gap and the weighted contributions from other clean latents are small, the conditional and unconditional target posterior probabilities first move farther apart and then closer together. The smaller posterior lag for normal prompts, despite their larger conditional reproduction errors, also shows that closer agreement alone does not establish target reproduction.

5 Explaining prior work

Several prior studies propose detection or mitigation strategies for memorization in diffusion models. Our account provides a unified mathematical explanation by relating these methods to conditional target reproduction, CFG amplification, and propagation of prompt influence. A broader discussion of related work appears in App. D.

Wen et al. [2024] empirically show that text prompts strongly guide the denoising process under memorization and use the magnitude of the conditional–unconditional noise-prediction difference for detection and mitigation. However, they motivate this association through intuition and empirical observations, without deriving how the prompt's contribution propagates through the sampling latent to the unconditional branch. By Eq. 4,

$$\boldsymbol{\Delta}_t = -\sqrt{\frac{1-\bar{\alpha}_t}{\bar{\alpha}_t}}\left(\boldsymbol{\epsilon}_c(\mathbf{x}_t, t) - \boldsymbol{\epsilon}_\emptyset(\mathbf{x}_t, t)\right). \tag{23}$$

Thus, at a fixed timestep, this noise-difference magnitude and the posterior lag magnitude in our analysis give the same ranking. The posterior lag also determines the difference $g\eta_t\boldsymbol{\Delta}_t$ between the CFG and unconditional updates (Eq. 13). Thm. 5 quantifies how this update difference affects the next unconditional clean prediction. Their inference-time mitigation therefore acts on the signal that drives prompt influence through the sampling update.

Hintersdorf et al. [2024] empirically find that the first-step residuals $\boldsymbol{\epsilon}_c(\mathbf{x}_T, T) - \mathbf{x}_T$ are more consistent across seeds after normalization for memorized prompts, and use this property to identify and deactivate cross-attention neurons that trigger replication. However, they motivate this consistency through intuition about similar denoising trajectories, without deriving why the first-step residuals share target-dependent structure across seeds. Our analysis connects the pair's expected training loss to initial conditional reproduction (Lem. 1). Rearranging Eq. 4 gives

$$\begin{aligned}\boldsymbol{\epsilon}_c(\mathbf{x}_T, T) - \mathbf{x}_T = &\left(\frac{1}{\sqrt{1-\bar{\alpha}_T}} - 1\right)\mathbf{x}_T \\ &- \sqrt{\frac{\bar{\alpha}_T}{1-\bar{\alpha}_T}}\,\hat{\mathbf{x}}_{T,c}.\end{aligned} \tag{24}$$

When $\hat{\mathbf{x}}_{T,c} \approx \mathbf{x}^\star$ across seeds, the second term supplies an approximately common target-dependent component. This explains how consistent normalized residuals can arise when that component dominates the remaining variation. When the unconditional clean prediction is also close to the data mean, CFG approximately amplifies the reproduced target's deviation from that mean (Thm. 3). Their neuron-level intervention can thus be interpreted as disrupting the conditional target reproduction that supplies the target component to guided generation.

Jain et al. [2025] propose mitigation by identifying a transition timestep in denoising: generation is initially unconditional, and CFG is enabled afterward. However, their attraction-basin account does not explicitly derive how the conditional contribution to a CFG update affects the next unconditional clean prediction. When the conditional branch reproduces the target at the first denoising step, CFG can amplify its deviation from the data mean (Thm. 3). At the same current latent, using an unconditional update instead removes the update difference $g\eta_t\boldsymbol{\Delta}_t$ (Eq. 13) and its contribution to the next unconditional clean prediction (Thm. 5).

6 Conclusion

In this paper, we provided a theoretical explanation of how memorization occurs in text-to-image diffusion models and supported it empirically. We showed that accurate initial conditional target reproduction and an unconditional clean prediction near the data mean allow CFG to approximately amplify the target's deviation from that mean. We derived an exact identity describing how the prompt's contribution to a CFG update affects the next unconditional clean prediction. Under Bayes-optimal noise prediction and bounded clean latents, we bounded final reproduction error using both branches' target posterior probabilities. Our posterior lag analysis relates branch separation to target evidence and contributions from other clean latents. Together, these results explain prior empirical findings and provide a theoretical basis for existing detection and mitigation strategies. Limitations and future work are discussed in App. E.

Acknowledgments and Disclosure of Funding

This work was supported by the National Research Foundation of Korea (NRF) grant (No. RS-2024-00453301, RS-2025-00517159, RS-2026-25490405) and the Institute of Information communications Technology Planning Evaluation (IITP) grant (IITP-2025-RS-2024-00428780).

References

Nikita Starodubcev, Mikhail Khoroshikh, Artem Babenko, and Dmitry Baranchuk. Invertible consistency distillation for text-guided image editing in around 7 steps. In *NeurIPS*, 2024.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint*, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Raphael Gontijo-Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022.

Ryan Webster. A reproducible extraction of training images from diffusion models. *arXiv preprint*, 2023.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? Investigating data replication in diffusion models. In *CVPR*, 2023a.

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. In *USENIX Security*, 2023.

Yuxin Wen, Yuchen Liu, Chen Chen, and Lingjuan Lyu. Detecting, explaining, and mitigating memorization in diffusion models. In *ICLR*, 2024.

Dominik Hintersdorf, Lukas Struppek, Kristian Kersting, Adam Dziedzic, and Franziska Boenisch. Finding NeMo: Localizing neurons responsible for memorization in diffusion models. In *NeurIPS*, 2024.

Harry H. Jiang, Lauren Brown, Jessica Cheng, Mehtab Khan, Abhishek Gupta, Deja Workman, Alex Hanna, Johnathan Flowers, and Timnit Gebru. AI art and its impact on artists. In *AIES*, 2023.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and mitigating copying in diffusion models. In *NeurIPS*, 2023b.

Xiangming Gu, Chao Du, Tianyu Pang, Chongxuan Li, Min Lin, and Ye Wang. On memorization in diffusion models. *TMLR*, 2025.

Ricardo Baptista, Agnimitra Dasgupta, Nikola B. Kovachki, Assad Oberai, and Andrew M. Stuart. Memorization and regularization in generative diffusion models. *arXiv preprint*, 2025.

Brendan Leigh Ross, Hamidreza Kamkari, Tongzi Wu, Rasa Hosseinzadeh, Zhaoyan Liu, George Stein, Jesse C. Cresswell, and Gabriel Loaiza-Ganem. A geometric framework for understanding memorization in generative models. In *ICLR*, 2025.

Dongjae Jeon, Dueun Kim, and Albert No. Understanding and mitigating memorization in generative models via sharpness of probability landscapes. In *ICML*, 2025.

Chen Chen, Daochang Liu, and Chang Xu. Towards memorization-free diffusion models. In *CVPR*, 2024.

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. In *ECCV*, 2024.

Anubhav Jain, Yuya Kobayashi, Takashi Shibuya, Yuhta Takida, Nasir Memon, Julian Togelius, and Yuki Mitsufuji. Classifier-free guidance inside the attraction basin may cause memorization. In *CVPR*, 2025.

Xinghui Li, Qichao Sun, Pengze Zhang, Fulong Ye, Zhichao Liao, Wanquan Feng, Songtao Zhao, and Qian He. AnyDressing: Customizable multi-garment virtual dressing via latent diffusion models. In *CVPR*, 2025.

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

Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *CVPR*, 2023.

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

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

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021.

Xianghao Kong, Rob Brekelmans, and Greg Ver Steeg. Information-theoretic diffusion. In *ICLR*, 2023.

Giannis Daras, Yuval Dagan, Alexandros G. Dimakis, and Constantinos Daskalakis. Consistent diffusion models: Mitigating sampling drift by learning to be consistent. In *NeurIPS*, 2023.

Xianghao Kong, Ollie Liu, Han Li, Dani Yogatama, and Greg Ver Steeg. Interpretable diffusion via information decomposition. In *ICLR*, 2024.

Joshua Lopez. Stable Diffusion v2.1 and DreamStudio updates 7-dec 22, 2022. URL `https://stability.ai/news-updates/stablediffusion2-1-release7-dec-2022`.

SG161222. Realistic Vision V1.4, 2023. URL `https://huggingface.co/SG161222/Realistic_Vision_V1.4`.

Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze. A self-supervised descriptor for image copy detection. In *CVPR*, 2022.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022.

Zahra Kadkhodaie, Florentin Guth, Eero P. Simoncelli, and Stéphane Mallat. Generalization in diffusion models arises from geometry-adaptive harmonic representations. In *ICLR*, 2024.

Gerrit J. J. van den Burg and Christopher K. I. Williams. On memorization in probabilistic deep generative models. In *NeurIPS*, 2021.

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

A Proofs

A.1 Proof of Lem. 1

Lemma 1 (Initial conditional reproduction)**.** *Let* $\mathbf{z} = \sqrt{\bar{\alpha}_T}\mathbf{x}^\star + \sqrt{1-\bar{\alpha}_T}\boldsymbol{\epsilon}$ *be a forward latent of the target at* $t = T$*, with* $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$*. If* $\mathbb{E}_{\boldsymbol{\epsilon}}[\mathcal{L}_c(\mathbf{x}^\star, T)] \le \varepsilon_T$*, then, for every tolerance* $r > 0$*,*

$$\Pr\big(\|\hat{\mathbf{x}}_{T,c}(\mathbf{z}) - \mathbf{x}^\star\|_2 \le r\big) \ge 1 - \frac{1-\bar{\alpha}_T}{\bar{\alpha}_T}\frac{\varepsilon_T}{r^2}.$$

The bound also holds at Gaussian initialization $\mathbf{z} = \mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ *up to* $o(1)$ *as* $\bar{\alpha}_T \to 0$*.*

Proof. For any timestep $t \in \{1, \dots, T\}$, substituting the forward latent $\mathbf{z} = \sqrt{\bar{\alpha}_t}\mathbf{x}^\star + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$ into Eq. 4 gives

$$\hat{\mathbf{x}}_{t,c}(\mathbf{z}) - \mathbf{x}^\star = \sqrt{\frac{1-\bar{\alpha}_t}{\bar{\alpha}_t}}\left[\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_c(\mathbf{z}, t)\right]. \tag{25}$$

Taking squared norms and using Eq. 3 yields the pointwise identity

$$\|\hat{\mathbf{x}}_{t,c}(\mathbf{z}) - \mathbf{x}^\star\|_2^2 = \frac{1-\bar{\alpha}_t}{\bar{\alpha}_t}\mathcal{L}_c(\mathbf{x}^\star, t). \tag{26}$$

This identity is algebraic and does not require Bayes-optimal noise prediction. At $t = T$, taking expectation over the forward noise gives

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\epsilon}}\Big[\|\hat{\mathbf{x}}_{T,c}(\mathbf{z}) - \mathbf{x}^\star\|_2^2\Big] &= \frac{1-\bar{\alpha}_T}{\bar{\alpha}_T}\mathbb{E}_{\boldsymbol{\epsilon}}[\mathcal{L}_c(\mathbf{x}^\star, T)] \\ &\le \frac{1-\bar{\alpha}_T}{\bar{\alpha}_T}\varepsilon_T.\end{aligned} \tag{27}$$

Markov's inequality therefore implies that, for every $r > 0$,

$$\Pr\big(\|\hat{\mathbf{x}}_{T,c}(\mathbf{z}) - \mathbf{x}^\star\|_2 > r\big) \le \frac{1-\bar{\alpha}_T}{\bar{\alpha}_T}\frac{\varepsilon_T}{r^2}. \tag{28}$$

To obtain the initialization bound, we compare the target's forward distribution $\mathcal{N}(\sqrt{\bar{\alpha}_T}\mathbf{x}^\star, (1-\bar{\alpha}_T)\mathbf{I})$ with the Gaussian initialization distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. The Gaussian relative entropy is

$$\begin{aligned}&D_{\mathrm{KL}}\big(\mathcal{N}\big(\sqrt{\bar{\alpha}_T}\mathbf{x}^\star, (1-\bar{\alpha}_T)\mathbf{I}\big) \,\big\|\, \mathcal{N}(\mathbf{0}, \mathbf{I})\big) \\ &\quad= \frac{1}{2}\left[d(1-\bar{\alpha}_T) + \bar{\alpha}_T\|\mathbf{x}^\star\|_2^2 - d - d\log(1-\bar{\alpha}_T)\right] \\ &\quad= \frac{1}{2}\left[\bar{\alpha}_T\|\mathbf{x}^\star\|_2^2 + d\left[-\bar{\alpha}_T - \log(1-\bar{\alpha}_T)\right]\right].\end{aligned} \tag{29}$$

For the total variation distance, Pinsker's inequality gives

$$\begin{aligned}&\mathrm{TV}\big(\mathcal{N}\big(\sqrt{\bar{\alpha}_T}\mathbf{x}^\star, (1-\bar{\alpha}_T)\mathbf{I}\big), \mathcal{N}(\mathbf{0}, \mathbf{I})\big) \\ &\quad\le \sqrt{\frac{1}{2}D_{\mathrm{KL}}\big(\mathcal{N}\big(\sqrt{\bar{\alpha}_T}\mathbf{x}^\star, (1-\bar{\alpha}_T)\mathbf{I}\big) \,\big\|\, \mathcal{N}(\mathbf{0}, \mathbf{I})\big)} \\ &\quad= \frac{1}{2}\sqrt{\bar{\alpha}_T\|\mathbf{x}^\star\|_2^2 + d\left[-\bar{\alpha}_T - \log(1-\bar{\alpha}_T)\right]}.\end{aligned} \tag{30}$$

Total variation bounds the absolute difference between the probabilities of the same event under the two distributions. Applying this to the reproduction-error event and using Eq. 28 yields

$$\begin{aligned}&\Pr\big(\|\hat{\mathbf{x}}_{T,c} - \mathbf{x}^\star\|_2 > r\big) \\ &\quad\le \Pr\big(\|\hat{\mathbf{x}}_{T,c}(\mathbf{z}) - \mathbf{x}^\star\|_2 > r\big) \\ &\qquad+ \frac{1}{2}\sqrt{\bar{\alpha}_T\|\mathbf{x}^\star\|_2^2 + d\left[-\bar{\alpha}_T - \log(1-\bar{\alpha}_T)\right]} \\ &\quad\le \frac{1-\bar{\alpha}_T}{\bar{\alpha}_T}\frac{\varepsilon_T}{r^2} + \frac{1}{2}\sqrt{\bar{\alpha}_T\|\mathbf{x}^\star\|_2^2 + d\left[-\bar{\alpha}_T - \log(1-\bar{\alpha}_T)\right]} \\ &\quad= \frac{1-\bar{\alpha}_T}{\bar{\alpha}_T}\frac{\varepsilon_T}{r^2} + o(1),\end{aligned} \tag{31}$$

as $\bar{\alpha}_T \to 0$. Taking complements in Eqs. 28 and 31 establishes the stated lower bounds on reproduction probability. □

A.2 Proof of Lem. 2

Lemma 2 (Initial unconditional prediction)**.** *Assume Eq. 8 and $\|\mathbf{X}\|_2 \le M$ q-almost surely. For each realized initial latent $\mathbf{x}_T$, the unconditional clean prediction satisfies*

$$\left\|\hat{\mathbf{x}}_{T,\emptyset} - \boldsymbol{\mu}\right\|_2 \le M\left(e^{2\kappa_T} - 1\right), \qquad \kappa_T := \frac{\sqrt{\bar{\alpha}_T} M \|\mathbf{x}_T\|_2 + \frac{1}{2}\bar{\alpha}_T M^2}{1 - \bar{\alpha}_T}.$$

For fixed $\mathbf{x}_T$, this bound goes to zero as $\bar{\alpha}_T \to 0$.

Proof. Fix the realized initial latent $\mathbf{x}_T$. Expanding the forward likelihood in Eq. 1 and omitting factors independent of $\mathbf{x}$ gives

$$\begin{aligned} q(\mathbf{x}_T \mid \mathbf{x}) &\propto \exp\left(-\frac{\|\mathbf{x}_T - \sqrt{\bar{\alpha}_T}\mathbf{x}\|_2^2}{2(1-\bar{\alpha}_T)}\right) \\ &\propto \exp\left(\frac{\sqrt{\bar{\alpha}_T}\mathbf{x}_T^\top \mathbf{x} - \frac{1}{2}\bar{\alpha}_T\|\mathbf{x}\|_2^2}{1-\bar{\alpha}_T}\right) =: w_T(\mathbf{x}). \end{aligned} \tag{32}$$

Under Bayes-optimal noise prediction, Eq. 9 and Bayes' rule yield

$$\begin{aligned} Z_T &:= \mathbb{E}_{\mathbf{X}\sim q}[w_T(\mathbf{X})], \\ \hat{\mathbf{x}}_{T,\emptyset} &= \mathbb{E}_{\mathbf{X}}[\mathbf{X} \mid \mathbf{X}_T = \mathbf{x}_T, T, \emptyset] = \frac{\mathbb{E}_{\mathbf{X}\sim q}[\mathbf{X} w_T(\mathbf{X})]}{Z_T}. \end{aligned} \tag{33}$$

By Cauchy–Schwarz and $\|\mathbf{X}\|_2 \le M$ q-almost surely,

$$\begin{aligned} |\log w_T(\mathbf{X})| &\le \frac{\sqrt{\bar{\alpha}_T}\|\mathbf{x}_T\|_2\|\mathbf{X}\|_2 + \frac{1}{2}\bar{\alpha}_T\|\mathbf{X}\|_2^2}{1-\bar{\alpha}_T} \\ &\le \frac{\sqrt{\bar{\alpha}_T} M\|\mathbf{x}_T\|_2 + \frac{1}{2}\bar{\alpha}_T M^2}{1-\bar{\alpha}_T} = \kappa_T. \end{aligned} \tag{34}$$

Consequently,

$$e^{-\kappa_T} \le w_T(\mathbf{X}) \le e^{\kappa_T}, \qquad e^{-\kappa_T} \le Z_T \le e^{\kappa_T}, \tag{35}$$

and hence, q-almost surely,

$$\begin{aligned} e^{-2\kappa_T} \le \frac{w_T(\mathbf{X})}{Z_T} \le e^{2\kappa_T}, \\ \left|\frac{w_T(\mathbf{X})}{Z_T} - 1\right| \le \max\left\{1 - e^{-2\kappa_T}, e^{2\kappa_T} - 1\right\} = e^{2\kappa_T} - 1. \end{aligned} \tag{36}$$

Subtracting $\boldsymbol{\mu} = \mathbb{E}_{\mathbf{X}\sim q}[\mathbf{X}]$ from Eq. 33 gives

$$\hat{\mathbf{x}}_{T,\emptyset} - \boldsymbol{\mu} = \mathbb{E}_{\mathbf{X}\sim q}\left[\mathbf{X}\left(\frac{w_T(\mathbf{X})}{Z_T} - 1\right)\right]. \tag{37}$$

Therefore,

$$\begin{aligned} \left\|\hat{\mathbf{x}}_{T,\emptyset} - \boldsymbol{\mu}\right\|_2 &\le \mathbb{E}_{\mathbf{X}\sim q}\left[\|\mathbf{X}\|_2 \left|\frac{w_T(\mathbf{X})}{Z_T} - 1\right|\right] \\ &\le M\left(e^{2\kappa_T} - 1\right). \end{aligned} \tag{38}$$

At fixed $\mathbf{x}_T$, we have $\kappa_T \to 0$ as $\bar{\alpha}_T \to 0$, so the bound tends to zero. □

A.3 Proof of Thm. 3

Theorem 3 (Initial CFG amplification)**.** *Assume the conditions of Lem. 2 and $\mathbb{E}_{\boldsymbol{\epsilon}}[\mathcal{L}_c(\mathbf{x}^\star, T)] \le \varepsilon_T$. For $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $g > 1$, and any tolerance $r > 0$,*

$$\Pr\left(\|\hat{\mathbf{x}}_{T,g} - [\boldsymbol{\mu} + g(\mathbf{x}^\star - \boldsymbol{\mu})]\|_2 > r\right) \le g^2 \frac{1-\bar{\alpha}_T}{\bar{\alpha}_T}\frac{\varepsilon_T}{r^2} + o(1),$$

as $\bar{\alpha}_T \to 0$.

Proof. For each realized initial latent $\mathbf{x}_T$, the CFG decomposition in Eq. 6 gives

$$\begin{aligned}&\hat{\mathbf{x}}_{T,g} - [\boldsymbol{\mu} + g(\mathbf{x}^\star - \boldsymbol{\mu})] \\ &\quad = g\left(\hat{\mathbf{x}}_{T,c} - \mathbf{x}^\star\right) + (1-g)\left(\hat{\mathbf{x}}_{T,\emptyset} - \boldsymbol{\mu}\right).\end{aligned} \tag{39}$$

Since $g > 1$, the triangle inequality and Lem. 2 imply

$$\begin{aligned}&\left\|\hat{\mathbf{x}}_{T,g} - [\boldsymbol{\mu} + g(\mathbf{x}^\star - \boldsymbol{\mu})]\right\|_2 \\ &\quad \le g\left\|\hat{\mathbf{x}}_{T,c} - \mathbf{x}^\star\right\|_2 + (g-1)\left\|\hat{\mathbf{x}}_{T,\emptyset} - \boldsymbol{\mu}\right\|_2 \\ &\quad \le g\left\|\hat{\mathbf{x}}_{T,c} - \mathbf{x}^\star\right\|_2 + (g-1)M\left(e^{2\kappa_T} - 1\right).\end{aligned} \tag{40}$$

By Eqs. 9 and 10, $\|\hat{\mathbf{x}}_{T,b}(\mathbf{z})\|_2 \le M$ for $b \in \{c, \emptyset\}$. Thus, the conditional squared error satisfies

$$0 \le \left\|\hat{\mathbf{x}}_{T,c}(\mathbf{z}) - \mathbf{x}^\star\right\|_2^2 \le \left(M + \|\mathbf{x}^\star\|_2\right)^2 \tag{41}$$

at every input $\mathbf{z}$. For a nonnegative bounded function, the absolute difference between its expectations under two distributions is at most its upper bound times their total variation distance. Applying Eqs. 27 and 30 therefore gives

$$\begin{aligned}&\mathbb{E}_{\mathbf{x}_T}\left[\left\|\hat{\mathbf{x}}_{T,c} - \mathbf{x}^\star\right\|_2^2\right] \\ &\quad \le \frac{1-\bar{\alpha}_T}{\bar{\alpha}_T}\varepsilon_T \\ &\qquad + \frac{\left(M + \|\mathbf{x}^\star\|_2\right)^2}{2}\sqrt{\bar{\alpha}_T\|\mathbf{x}^\star\|_2^2 + d\left[-\bar{\alpha}_T - \log(1-\bar{\alpha}_T)\right]} \\ &\quad = \frac{1-\bar{\alpha}_T}{\bar{\alpha}_T}\varepsilon_T + o(1).\end{aligned} \tag{42}$$

For the unconditional branch, $\|\boldsymbol{\mu}\|_2 \le \mathbb{E}_{\mathbf{X}\sim q}[\|\mathbf{X}\|_2] \le M$ and Lem. 2 yield

$$\left\|\hat{\mathbf{x}}_{T,\emptyset} - \boldsymbol{\mu}\right\|_2 \le \min\left\{2M,\, M\left(e^{2\kappa_T} - 1\right)\right\} \longrightarrow 0 \tag{43}$$

as $\bar{\alpha}_T \to 0$ at each fixed $\mathbf{x}_T$. Since the initialization distribution remains $\mathcal{N}(\mathbf{0}, \mathbf{I})$, dominated convergence gives

$$\begin{aligned}\mathbb{E}_{\mathbf{x}_T}\left[\left\|\hat{\mathbf{x}}_{T,\emptyset} - \boldsymbol{\mu}\right\|_2\right] &\longrightarrow 0, \\ \mathbb{E}_{\mathbf{x}_T}\left[\left\|\hat{\mathbf{x}}_{T,\emptyset} - \boldsymbol{\mu}\right\|_2^2\right] &\longrightarrow 0.\end{aligned} \tag{44}$$

Squaring the first inequality in Eq. 40, taking expectation, and using these bounds yields

$$\begin{aligned}&\mathbb{E}_{\mathbf{x}_T}\left[\left\|\hat{\mathbf{x}}_{T,g} - [\boldsymbol{\mu} + g(\mathbf{x}^\star - \boldsymbol{\mu})]\right\|_2^2\right] \\ &\quad \le g^2\mathbb{E}_{\mathbf{x}_T}\left[\left\|\hat{\mathbf{x}}_{T,c} - \mathbf{x}^\star\right\|_2^2\right] \\ &\qquad + 2g(g-1)\left(M + \|\mathbf{x}^\star\|_2\right)\mathbb{E}_{\mathbf{x}_T}\left[\left\|\hat{\mathbf{x}}_{T,\emptyset} - \boldsymbol{\mu}\right\|_2\right] \\ &\qquad + (g-1)^2\mathbb{E}_{\mathbf{x}_T}\left[\left\|\hat{\mathbf{x}}_{T,\emptyset} - \boldsymbol{\mu}\right\|_2^2\right] \\ &\quad \le g^2\frac{1-\bar{\alpha}_T}{\bar{\alpha}_T}\varepsilon_T + o(1).\end{aligned} \tag{45}$$

Finally, Markov's inequality gives

$$\begin{aligned}&\Pr_{\mathbf{x}_T}\left(\left\|\hat{\mathbf{x}}_{T,g} - [\boldsymbol{\mu} + g(\mathbf{x}^\star - \boldsymbol{\mu})]\right\|_2 > r\right) \\ &\quad \le \frac{1}{r^2}\mathbb{E}_{\mathbf{x}_T}\left[\left\|\hat{\mathbf{x}}_{T,g} - [\boldsymbol{\mu} + g(\mathbf{x}^\star - \boldsymbol{\mu})]\right\|_2^2\right] \\ &\quad \le g^2\frac{1-\bar{\alpha}_T}{\bar{\alpha}_T}\frac{\varepsilon_T}{r^2} + o(1),\end{aligned} \tag{46}$$

as $\bar{\alpha}_T \to 0$ for each fixed tolerance $r > 0$. □

Finite-model correction. At the first denoising step, Eq. 39 holds without assuming Eq. 8. Under $\|\mathbf{X}\|_2 \le M$ q-almost surely, adding and subtracting the unconditional posterior mean and applying its bound from App. A.2 gives

$$\begin{aligned}
&\|\hat{\mathbf{x}}_{T,g} - [\boldsymbol{\mu} + g(\mathbf{x}^\star - \boldsymbol{\mu})]\|_2 \\
&\quad \le g\,\|\hat{\mathbf{x}}_{T,c} - \mathbf{x}^\star\|_2 + (g-1)M\left(e^{2\kappa_T} - 1\right) \\
&\qquad + (g-1)\left\|\hat{\mathbf{x}}_{T,\emptyset} - \mathbb{E}_{\mathbf{X}}[\mathbf{X} \mid \mathbf{X}_T = \mathbf{x}_T, T, \emptyset]\right\|_2 .
\end{aligned} \tag{47}$$

The additional term measures the unconditional clean prediction's departure from its posterior mean. Conditional reproduction is already quantified through the actual network loss in Lem. 1, so no separate conditional model-error term is needed.

A.4 Proof of Prop. 4

Proposition 4 (Subsequent conditional reproduction)**.** *Fix $t \in \{1, \dots, T\}$, a realized sampling latent $\mathbf{x}_t$, and a tolerance $r \ge 0$. Then $\|\hat{\mathbf{x}}_{t,c} - \mathbf{x}^\star\|_2 \le r$ if and only if*

$$\left\| \frac{\mathbf{x}_t - \sqrt{\bar{\alpha}_t}\mathbf{x}^\star}{\sqrt{1-\bar{\alpha}_t}} - \boldsymbol{\epsilon}_c(\mathbf{x}_t, t) \right\|_2 \le r\sqrt{\frac{\bar{\alpha}_t}{1-\bar{\alpha}_t}}.$$

Proof. Using Eq. 4 at $\mathbf{x}_t$ and subtracting $\mathbf{x}^\star$ gives

$$\begin{aligned}
\hat{\mathbf{x}}_{t,c} - \mathbf{x}^\star &= \frac{\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}_c(\mathbf{x}_t, t) - \sqrt{\bar{\alpha}_t}\mathbf{x}^\star}{\sqrt{\bar{\alpha}_t}} \\
&= \sqrt{\frac{1-\bar{\alpha}_t}{\bar{\alpha}_t}}\left[\frac{\mathbf{x}_t - \sqrt{\bar{\alpha}_t}\mathbf{x}^\star}{\sqrt{1-\bar{\alpha}_t}} - \boldsymbol{\epsilon}_c(\mathbf{x}_t, t)\right].
\end{aligned} \tag{48}$$

Since $0 < \bar{\alpha}_t < 1$, taking Euclidean norms and rearranging proves the equivalence. □

A.5 Proof of Thm. 5

Theorem 5 (Propagation of prompt influence)**.** *For $t \in \{2, \dots, T\}$, suppose $\hat{\mathbf{x}}_{t-1,\emptyset}(\cdot)$ is continuously differentiable on a neighborhood of the segment in Eq. 14. Then*

$$\hat{\mathbf{x}}_{t-1,\emptyset}(\mathbf{x}_{t-1}) - \hat{\mathbf{x}}_{t-1,\emptyset}(\mathbf{x}^\emptyset_{t-1}) = g\eta_t \int_0^1 \mathbf{J}_{t-1,\emptyset}(\mathbf{z}_t(s))\boldsymbol{\Delta}_t\, ds,$$

and consequently

$$\left\|\hat{\mathbf{x}}_{t-1,\emptyset}(\mathbf{x}_{t-1}) - \hat{\mathbf{x}}_{t-1,\emptyset}(\mathbf{x}^\emptyset_{t-1})\right\|_2 \le g\eta_t \sup_{s\in[0,1]} \left\|\mathbf{J}_{t-1,\emptyset}(\mathbf{z}_t(s))\right\|_{\mathrm{op}} \|\boldsymbol{\Delta}_t\|_2.$$

Proof. The segment in Eq. 14 satisfies $\mathbf{z}_t'(s) = g\eta_t\boldsymbol{\Delta}_t$, so the chain rule gives

$$\frac{d}{ds}\hat{\mathbf{x}}_{t-1,\emptyset}(\mathbf{z}_t(s)) = \mathbf{J}_{t-1,\emptyset}(\mathbf{z}_t(s))\mathbf{z}_t'(s) = g\eta_t\mathbf{J}_{t-1,\emptyset}(\mathbf{z}_t(s))\boldsymbol{\Delta}_t. \tag{49}$$

Integrating over $s \in [0,1]$ and using $\mathbf{z}_t(0) = \mathbf{x}^\emptyset_{t-1}$ and $\mathbf{z}_t(1) = \mathbf{x}_{t-1}$ proves the stated identity by the fundamental theorem of calculus. Since $g\eta_t > 0$, the triangle and operator-norm inequalities give

$$\begin{aligned}
&\left\|\hat{\mathbf{x}}_{t-1,\emptyset}(\mathbf{x}_{t-1}) - \hat{\mathbf{x}}_{t-1,\emptyset}(\mathbf{x}^\emptyset_{t-1})\right\|_2 \\
&\quad \le g\eta_t \int_0^1 \left\|\mathbf{J}_{t-1,\emptyset}(\mathbf{z}_t(s))\boldsymbol{\Delta}_t\right\|_2 ds \\
&\quad \le g\eta_t \int_0^1 \left\|\mathbf{J}_{t-1,\emptyset}(\mathbf{z}_t(s))\right\|_{\mathrm{op}} \|\boldsymbol{\Delta}_t\|_2\, ds \\
&\quad \le g\eta_t \sup_{s\in[0,1]} \left\|\mathbf{J}_{t-1,\emptyset}(\mathbf{z}_t(s))\right\|_{\mathrm{op}} \|\boldsymbol{\Delta}_t\|_2.
\end{aligned} \tag{50}$$

The supremum is finite because the Jacobian is continuous on the compact segment. □

Continuous differentiability of $\hat{\mathbf{x}}_{t-1,\emptyset}(\cdot)$. Under Eq. 8, the unconditional clean prediction equals the posterior mean in Eq. 9. For $t \ge 2$, the Gaussian forward likelihood at timestep $t-1$ has positive variance $1 - \bar{\alpha}_{t-1} > 0$. Bayes' formula expresses this posterior mean as a ratio of smooth Gaussian integrals with a strictly positive denominator, ensuring continuous differentiability.

A.6 Proof of Thm. 6

Theorem 6 (Final reproduction)**.** *Assume Eq. 8 and $\|\mathbf{X}\|_2 \le M$ almost surely under the conditional and unconditional data distributions. For each realized latent $\mathbf{x}_1$ at the final denoising step and $g > 1$, the final generated latent satisfies*

$$\|\mathbf{x}_0 - \mathbf{x}^\star\|_2 \le \big(M + \|\mathbf{x}^\star\|_2\big)\Big[g\big(1 - \pi_{1,c}(\mathbf{x}_1)\big) + (g-1)\big(1 - \pi_{1,\emptyset}(\mathbf{x}_1)\big)\Big].$$

Proof. Fix the realized latent $\mathbf{x}_1$ at the final denoising step. Under Eq. 8, each clean prediction equals the posterior mean in Eq. 9 evaluated at $\mathbf{x}_1$. For $b \in \{c, \emptyset\}$ with $\pi_{1,b}(\mathbf{x}_1) < 1$, subtracting $\mathbf{x}^\star$ from Eq. 17 and taking norms gives

$$\|\hat{\mathbf{x}}_{1,b} - \mathbf{x}^\star\|_2 = \big[1 - \pi_{1,b}(\mathbf{x}_1)\big]\, \|\mathbb{E}_{\mathbf{X}}[\mathbf{X} \mid \mathbf{X}_1 = \mathbf{x}_1, 1, b, \mathbf{X} \neq \mathbf{x}^\star] - \mathbf{x}^\star\|_2 \,. \tag{51}$$

If $\pi_{1,b}(\mathbf{x}_1) = 1$, then $\hat{\mathbf{x}}_{1,b} = \mathbf{x}^\star$, and the weighted non-target term is defined as zero.

The identity $\mathbf{x}_0 = \hat{\mathbf{x}}_{1,g}$ and Eq. 6 give

$$\mathbf{x}_0 - \mathbf{x}^\star = g\big(\hat{\mathbf{x}}_{1,c} - \mathbf{x}^\star\big) + (1-g)\big(\hat{\mathbf{x}}_{1,\emptyset} - \mathbf{x}^\star\big). \tag{52}$$

Since $g > 1$, the triangle inequality and Eq. 51 imply

$$\begin{aligned}
\|\mathbf{x}_0 - \mathbf{x}^\star\|_2 &\le g\|\hat{\mathbf{x}}_{1,c} - \mathbf{x}^\star\|_2 + (g-1)\|\hat{\mathbf{x}}_{1,\emptyset} - \mathbf{x}^\star\|_2 \\
&= g\big[1 - \pi_{1,c}(\mathbf{x}_1)\big]\, \|\mathbb{E}_{\mathbf{X}}[\mathbf{X} \mid \mathbf{X}_1 = \mathbf{x}_1, 1, c, \mathbf{X} \neq \mathbf{x}^\star] - \mathbf{x}^\star\|_2 \\
&\quad + (g-1)\big[1 - \pi_{1,\emptyset}(\mathbf{x}_1)\big]\, \|\mathbb{E}_{\mathbf{X}}[\mathbf{X} \mid \mathbf{X}_1 = \mathbf{x}_1, 1, \emptyset, \mathbf{X} \neq \mathbf{x}^\star] - \mathbf{x}^\star\|_2 \,.
\end{aligned} \tag{53}$$

This bound retains the full non-target contribution and does not use the assumption $\|\mathbf{X}\|_2 \le M$.

Since $0 < \bar{\alpha}_1 < 1$, the Gaussian forward likelihood is strictly positive, and Bayes' rule preserves each prior's almost-sure bound $\|\mathbf{X}\|_2 \le M$ under posterior conditioning. For $\pi_{1,b}(\mathbf{x}_1) < 1$, the triangle inequality and Jensen's inequality therefore give

$$\begin{aligned}
&\|\mathbb{E}_{\mathbf{X}}[\mathbf{X} \mid \mathbf{X}_1 = \mathbf{x}_1, 1, b, \mathbf{X} \neq \mathbf{x}^\star] - \mathbf{x}^\star\|_2 \\
&\quad \le \mathbb{E}_{\mathbf{X}}[\|\mathbf{X}\|_2 \mid \mathbf{X}_1 = \mathbf{x}_1, 1, b, \mathbf{X} \neq \mathbf{x}^\star] + \|\mathbf{x}^\star\|_2 \\
&\quad \le M + \|\mathbf{x}^\star\|_2.
\end{aligned} \tag{54}$$

Substituting Eq. 54 into Eq. 53 proves the theorem. □

A.7 Proof of Lem. 7

Lemma 7 (Posterior probability gap)**.** *Assume $0 < p_\emptyset < p_c < 1$. The posterior probability gap $G(\lambda)$ is unimodal in the target evidence λ, with unique maximizer*

$$\lambda^\star := -\frac{\operatorname{logit}(p_c) + \operatorname{logit}(p_\emptyset)}{2}.$$

On $[\lambda_T, \infty)$, G strictly decreases if $\lambda_T \ge \lambda^\star$. Otherwise, it strictly increases on $[\lambda_T, \lambda^\star]$ and strictly decreases on $[\lambda^\star, \infty)$. When $\lambda_T = 0$, these cases correspond to $p_c + p_\emptyset \ge 1$ and $p_c + p_\emptyset < 1$, respectively.

Proof. Differentiating Eq. 22 gives

$$G'(\lambda) = \sigma'(\lambda + \operatorname{logit}(p_c)) - \sigma'(\lambda + \operatorname{logit}(p_\emptyset)) \,. \tag{55}$$

The sigmoid derivative satisfies

$$\sigma'(u) = \sigma(u)\big[1 - \sigma(u)\big] = \frac{1}{4\cosh^2(u/2)}, \tag{56}$$

so σ' is even and strictly decreases as $|u|$ increases. Since $\operatorname{logit}(p_c) > \operatorname{logit}(p_\emptyset)$, comparing the squared arguments yields

$$\begin{aligned}
G'(\lambda) > 0 \quad &\Longleftrightarrow \quad |\lambda + \operatorname{logit}(p_c)| < |\lambda + \operatorname{logit}(p_\emptyset)| \\
&\Longleftrightarrow \quad 2\lambda + \operatorname{logit}(p_c) + \operatorname{logit}(p_\emptyset) < 0 \\
&\Longleftrightarrow \quad \lambda < \lambda^\star.
\end{aligned} \tag{57}$$

Similarly, $G'(\lambda) < 0$ for $\lambda > \lambda^\star$, and $G'(\lambda^\star) = 0$. Thus, G strictly increases on $(-\infty, \lambda^\star]$ and strictly decreases on $[\lambda^\star, \infty)$, proving unimodality and the unique maximizer.

By the definition of $\lambda^\star$,

$$\lambda_T \geq \lambda^\star \quad \iff \quad 2\lambda_T + \operatorname{logit}(p_c) + \operatorname{logit}(p_\emptyset) \geq 0. \tag{58}$$

If $\lambda_T \geq \lambda^\star$, then $[\lambda_T, \infty)$ lies within $[\lambda^\star, \infty)$, where G is strictly decreasing. Otherwise, $\lambda_T < \lambda^\star$, so G strictly increases on $[\lambda_T, \lambda^\star]$ and strictly decreases on $[\lambda^\star, \infty)$.

When $\lambda_T = 0$, the condition $\lambda_T \geq \lambda^\star$ becomes

$$\begin{aligned} \operatorname{logit}(p_c) + \operatorname{logit}(p_\emptyset) \geq 0 \quad &\iff \quad \log \frac{p_c p_\emptyset}{(1-p_c)(1-p_\emptyset)} \geq 0 \\ &\iff \quad p_c p_\emptyset \geq (1-p_c)(1-p_\emptyset) \\ &\iff \quad p_c + p_\emptyset \geq 1. \end{aligned} \tag{59}$$

The complementary case $\lambda_T < \lambda^\star$ therefore corresponds to $p_c + p_\emptyset < 1$. □

A.8 Proof of Thm. 8

Theorem 8 (Posterior lag dynamics)**.** *Assume Eq. 8 and $0 < p_\emptyset < p_c < 1$. For each timestep $t \in \{1, \dots, T\}$ and realized latent $\mathbf{x}_t$ satisfying Eq. 19, the posterior lag and its magnitude satisfy*

$$\begin{aligned} \big|\|\boldsymbol{\Delta}_t\|_2 - G(\lambda_t)\|\mathbf{x}^\star\|_2\big| &\leq \|\boldsymbol{\Delta}_t - G(\lambda_t)\mathbf{x}^\star\|_2 \\ &\leq \sum_{b \in \{c, \emptyset\}} \big[1 - \pi_{t,b}(\mathbf{x}_t)\big] \, \big\|\mathbb{E}_{\mathbf{X}}[\mathbf{X} \mid \mathbf{X}_t = \mathbf{x}_t, t, b, \mathbf{X} \neq \mathbf{x}^\star]\big\|_2 . \end{aligned}$$

Proof. Fix $t \in \{1, \dots, T\}$ and a realized sampling latent $\mathbf{x}_t$ satisfying Eq. 19. Since $0 < \bar{\alpha}_t < 1$, the Gaussian forward likelihood is strictly positive, and $0 < p_b < 1$ ensures $0 < \pi_{t,b}(\mathbf{x}_t) < 1$ for $b \in \{c, \emptyset\}$. Bayes' rule, Eq. 19, and the definition of λ_t give

$$\begin{aligned} \frac{\pi_{t,b}(\mathbf{x}_t)}{1 - \pi_{t,b}(\mathbf{x}_t)} &= \frac{p_b}{1-p_b} \frac{q(\mathbf{x}_t \mid \mathbf{x}^\star)}{\mathbb{E}_{\mathbf{X}}[q(\mathbf{x}_t \mid \mathbf{X}) \mid b, \mathbf{X} \neq \mathbf{x}^\star]} \\ &= \exp(\lambda_t + \operatorname{logit}(p_b)) . \end{aligned} \tag{60}$$

Taking logarithms and applying σ gives Eq. 21. Since $p_c > p_\emptyset$ and σ is strictly increasing, Eq. 22 yields $\pi_{t,c}(\mathbf{x}_t) - \pi_{t,\emptyset}(\mathbf{x}_t) = G(\lambda_t) > 0$.

Under Eq. 8, each clean prediction equals the forward-model posterior mean in Eq. 9 evaluated at $\mathbf{x}_t$. For $b \in \{c, \emptyset\}$, the law of total expectation gives

$$\begin{aligned} \hat{\mathbf{x}}_{t,b} = \pi_{t,b}(\mathbf{x}_t)\mathbf{x}^\star \\ + \big[1 - \pi_{t,b}(\mathbf{x}_t)\big]\mathbb{E}_{\mathbf{X}}[\mathbf{X} \mid \mathbf{X}_t = \mathbf{x}_t, t, b, \mathbf{X} \neq \mathbf{x}^\star] . \end{aligned} \tag{61}$$

Subtracting the unconditional decomposition from the conditional decomposition and using $\boldsymbol{\Delta}_t = \hat{\mathbf{x}}_{t,c} - \hat{\mathbf{x}}_{t,\emptyset}$ yields

$$\begin{aligned} \boldsymbol{\Delta}_t - G(\lambda_t)\mathbf{x}^\star = \big[1 - \pi_{t,c}(\mathbf{x}_t)\big]\mathbb{E}_{\mathbf{X}}[\mathbf{X} \mid \mathbf{X}_t = \mathbf{x}_t, t, c, \mathbf{X} \neq \mathbf{x}^\star] \\ - \big[1 - \pi_{t,\emptyset}(\mathbf{x}_t)\big]\mathbb{E}_{\mathbf{X}}[\mathbf{X} \mid \mathbf{X}_t = \mathbf{x}_t, t, \emptyset, \mathbf{X} \neq \mathbf{x}^\star] . \end{aligned} \tag{62}$$

Since $G(\lambda_t) > 0$, the reverse triangle inequality and the triangle inequality applied to Eq. 62 give

$$\begin{aligned} \big|\|\boldsymbol{\Delta}_t\|_2 - G(\lambda_t)\|\mathbf{x}^\star\|_2\big| &\leq \|\boldsymbol{\Delta}_t - G(\lambda_t)\mathbf{x}^\star\|_2 \\ &\leq \sum_{b \in \{c, \emptyset\}} \big[1 - \pi_{t,b}(\mathbf{x}_t)\big] \, \big\|\mathbb{E}_{\mathbf{X}}[\mathbf{X} \mid \mathbf{X}_t = \mathbf{x}_t, t, b, \mathbf{X} \neq \mathbf{x}^\star]\big\|_2 . \end{aligned} \tag{63}$$

This proves both inequalities. □

B Dataset

In this paper, we use prompts from Webster [2023], sourced from the LAION-5B dataset [Schuhmann et al., 2022], for Stable Diffusion v1.4 (SDv1) [Rombach et al., 2022], Stable Diffusion v2.1 (SDv2) [Lopez, 2022], and RealisticVision (RealVis) [SG161222, 2023]. The released metadata group the prompts into four categories:

(i) Matching verbatim (`MV`**)**: generated outputs that closely reproduce their paired training images;

(ii) Retrieval verbatim (`RV`**)**: cases where the output does not match the paired training image but reproduces another training image identified through retrieval;

(iii) Template verbatim (`TV`**)**: images that share the overall template with a training image but differ in details such as colors or textures; and

(iv) None (`N`**)**: cases where no verbatim reproduction was identified in the original evaluation.

Each prompt was provided with a URL linking to its paired training image. However, some URLs were inaccessible, preventing direct retrieval of the corresponding images. For these cases, our data preparation procedure also checks a ground-truth mirror, accepting an image only when its original index, prompt, and URL match the source entry.[3] We exclude pairs whose target images remain unavailable and use all remaining pairs, as shown in Table 1. Each prompt's paired target image remains fixed throughout the experiments.

The original categories describe generation outcomes under the settings and evaluation criteria of Webster [2023], which can differ from those used in our experiments. We therefore re-score prompts using SSCD [Pizzi et al., 2022] to assess reproduction of each prompt's paired target under our model and sampler configurations (Sec. 4.1). This provides a consistent criterion for grouping prompts across all configurations. For each model–sampler configuration, we classify a prompt as memorized if at least one of its 20 generated images has SSCD > 0.7 with its paired target, and as normal if all scores are ≤ 0.7. These groups are used for the experimental comparisons, while the categories in Table 1 retain the original source annotations. All generated samples remain in their prompt's group, and prompt-group summaries give each prompt equal weight.

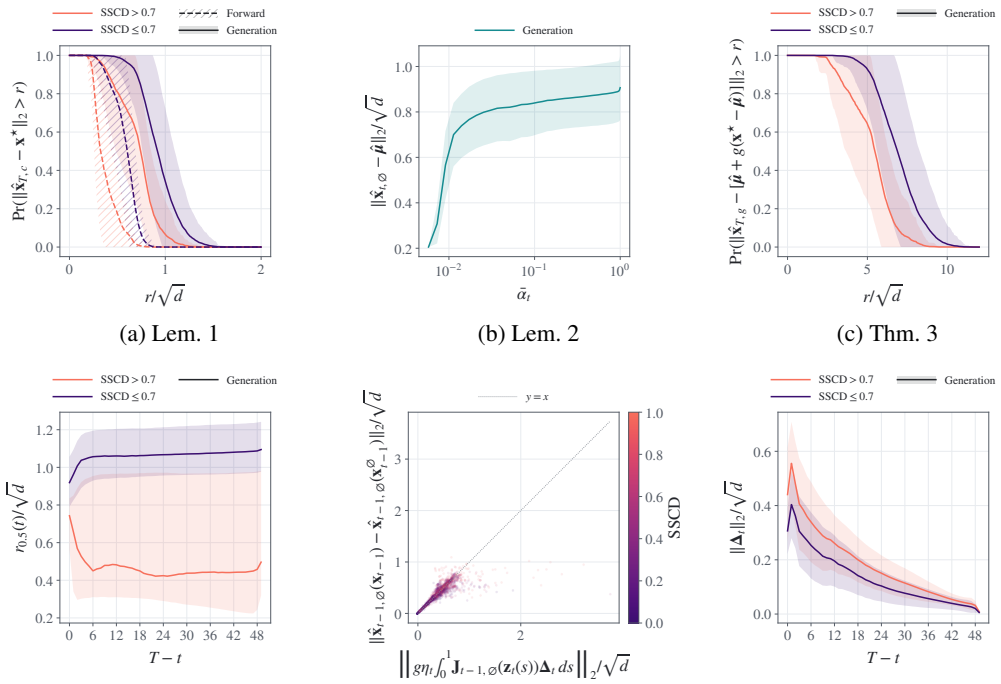
Table 1: **Dataset.** Original prompt categories and counts across different diffusion models. Fractions indicate the number of prompts with available paired target images over the total number of prompts originally provided by Webster [2023].

	`MV`	`RV`	`TV`	`N`	Total
SDv1	86/86	30/30	229/229	101/155	446/500
SDv2	2/4	0/0	208/215	180/281	390/500
RealVis	89/90	34/34	229/230	94/146	446/500

C Additional results

Figs. 2, 3 and 4 show results for SDv1 (DDPM), SDv2 (DDIM), and RealVis (DDIM), respectively. The outcomes closely mirror those reported for SDv1 (DDIM) in the main text (Fig. 1), supporting our findings across different diffusion models and samplers.

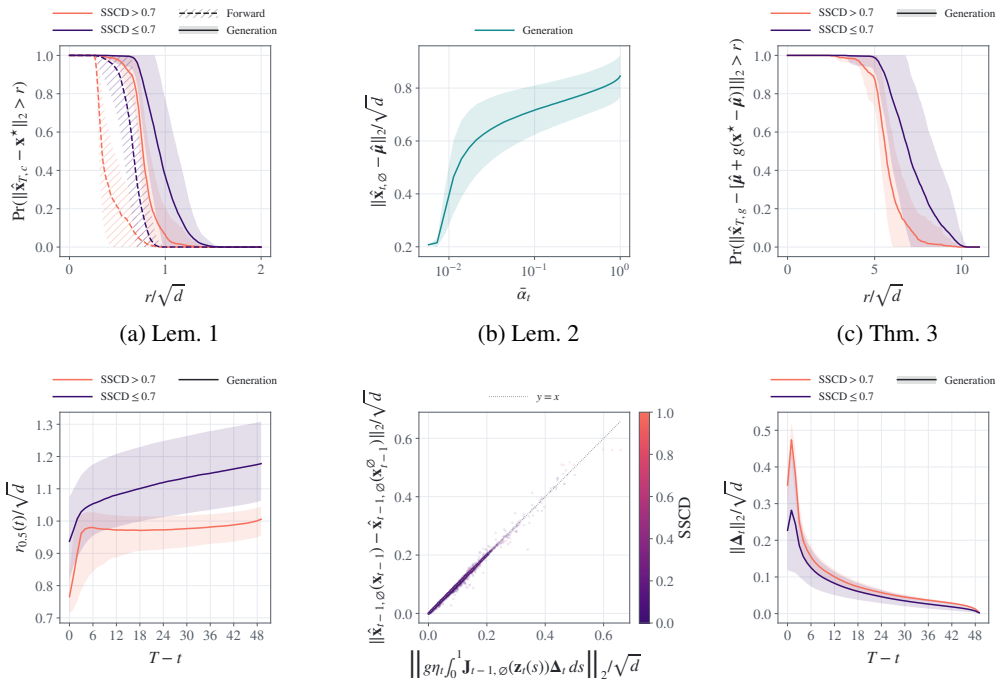
[3] `https://huggingface.co/datasets/gdhanuka/memorization_data2`



(a) Lem. 1 (b) Lem. 2 (c) Thm. 3

(d) Prop. 4 (e) Thm. 5 (f) Thm. 8

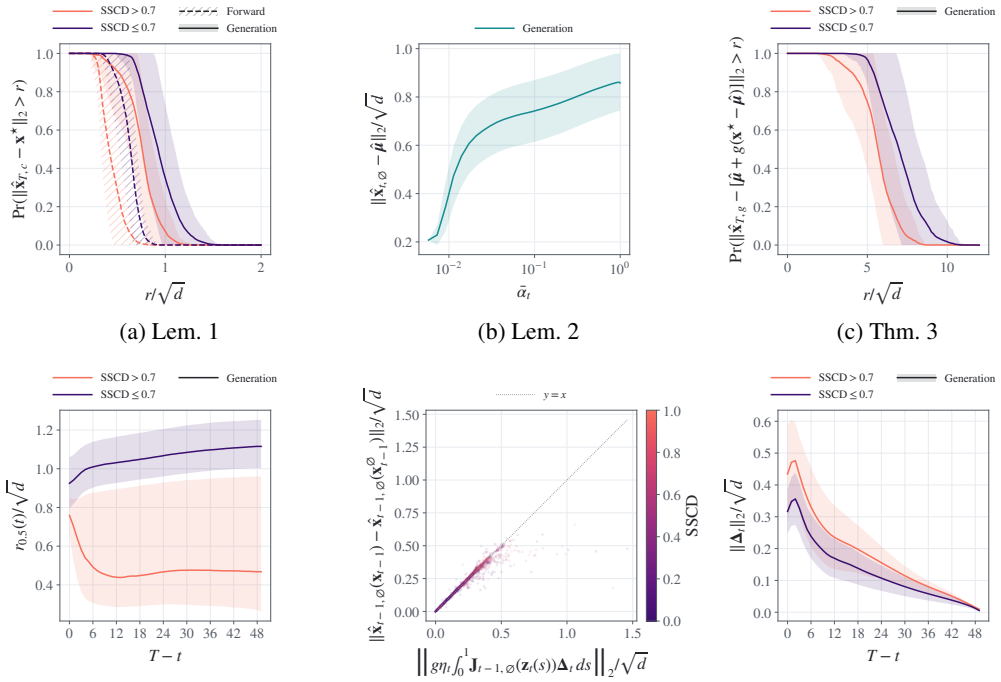
Figure 2: **Denoising measurements on SDv1 (DDPM).** Coral and purple in (a,c,d,f) denote memorized and normal prompt groups, respectively (Sec. 4.1). In (a), dashed and solid curves correspond to forward latents and Gaussian initializations, respectively. In (a–c,f), curves show means across prompts, and shaded bands indicate ± 1 standard deviation. Seeds are first averaged within each prompt in (b,f). In (d), curves and shaded bands show the median and interquartile range, respectively, with each prompt receiving equal total weight. In (e), point colors indicate each generated image's final SSCD, and the dotted diagonal denotes equality.



(a) Lem. 1 (b) Lem. 2 (c) Thm. 3

(d) Prop. 4 (e) Thm. 5 (f) Thm. 8

Figure 3: **Denoising measurements on SDv2 (DDIM).** Coral and purple in (a,c,d,f) denote memorized and normal prompt groups, respectively (Sec. 4.1). In (a), dashed and solid curves correspond to forward latents and Gaussian initializations, respectively. In (a–c,f), curves show means across prompts, and shaded bands indicate ± 1 standard deviation. Seeds are first averaged within each prompt in (b,f). In (d), curves and shaded bands show the median and interquartile range, respectively, with each prompt receiving equal total weight. In (e), point colors indicate each generated image's final SSCD, and the dotted diagonal denotes equality.



(a) Lem. 1 (b) Lem. 2 (c) Thm. 3

(d) Prop. 4 (e) Thm. 5 (f) Thm. 8

Figure 4: **Denoising measurements on RealVis (DDIM).** Coral and purple in (a,c,d,f) denote memorized and normal prompt groups, respectively (Sec. 4.1). In (a), dashed and solid curves correspond to forward latents and Gaussian initializations, respectively. In (a–c,f), curves show means across prompts, and shaded bands indicate ± 1 standard deviation. Seeds are first averaged within each prompt in (b,f). In (d), curves and shaded bands show the median and interquartile range, respectively, with each prompt receiving equal total weight. In (e), point colors indicate each generated image's final SSCD, and the dotted diagonal denotes equality.

D Related work

Memorization and generalization. Prior work has studied the relationship between memorization and generalization in diffusion models [Kadkhodaie et al., 2024]. van den Burg and Williams [2021] showed that removing a sample from training data can induce local density changes, which they use to quantify sample-level memorization. Other studies found that duplicated images are more likely to be reproduced [Somepalli et al., 2023a,b, Carlini et al., 2023]. Theoretical analyses also examine how empirical score optima, training dynamics, and regularization affect memorization [Gu et al., 2025, Baptista et al., 2025, Bonnaire et al., 2025]. Recent works propose geometric views: memorization is related to the local intrinsic dimensions of the data and learned manifolds [Ross et al., 2025]. Similarly, Jeon et al. [2025] link memorization to sharp regions of the log density.

Text guidance and memorization. Prior work has established that text guidance is closely related to memorization. For example, Somepalli et al. [2023b] highlight the role of text conditioning, and Chen et al. [2024] propose Anti-Memorization Guidance (AMG) to mitigate memorization associated with duplicated images, duplicated captions, and highly specific prompts. Wen et al. [2024] use the magnitude of the difference between conditional and unconditional noise predictions for detection and mitigation. Other studies identify concentrated cross-attention on trigger tokens [Ren et al., 2024] and cross-attention neurons that trigger replication [Hintersdorf et al., 2024]. Jain et al. [2025] provide an attraction-basin account and show how the timing of CFG affects memorization.

Our analysis complements these perspectives by explicitly deriving how the prompt's contribution to a CFG update changes the next unconditional clean prediction. We connect this mechanism to conditional target reproduction and CFG amplification at the first denoising step. We also bound the final reproduction error using both branches' target posterior probabilities at the final denoising step and show that closer agreement alone does not establish target reproduction.

E Limitations and future work

Our analysis is currently limited to diffusion models and has not yet been extended to recent flow-based generative models. Since flow models also generate samples through continuous trajectories, a similar analysis may reveal whether guidance amplifies and propagates information about the target through the same underlying mechanism. Future work includes this extension and the development and comparison of mitigation methods that weaken conditional target reproduction, reduce early guidance-induced latent changes, or prevent subsequent unconditional target reproduction.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Please see App. E.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

Justification: Please see App. A.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: Please see Sec. 4.1.

   Guidelines:

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

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

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

Answer: [No]

Justification: We do not provide additional code or data. The models and datasets used in our experiments are publicly available, and the experimental setup is described in Sec. 4.1. We will consider releasing code in a future version to further improve reproducibility.

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Please see Sec. 4.1.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: Please see Sec. 4.2 and App. C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Please see Sec. 4.1

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

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

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [N/A]

    Justification: N/A

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: N/A

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

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

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [N/A]

    Justification: N/A

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: N/A

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: N/A

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: N/A

Guidelines:

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