PROBING LOCAL CONDITIONAL FIT FOR DIFFUSION MODEL MEMBERSHIP INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Membership inference determines whether a candidate example was used to train a text-to-image diffusion model and is an important tool for auditing training-data use. We identify a directional membership signal in *conditional prediction shifts*, the changes in predicted noise under alternative text conditions at a fixed noisy input: whether movement toward the alternative prediction initially lowers or raises the denoising loss from its value under the original caption. Members exhibit less room for such local improvement. To quantify this response, we propose *LoFIT* (Local Conditional Fit), which measures the signed initial response of the denoising loss along normalized prediction shifts and aggregates it across multiple text conditions and diffusion states. LoFIT requires no target-model gradients and can be computed using only forward queries to the denoiser. Evaluations on four datasets show strong membership discrimination, particularly for fine-tuned models. LoFIT distinguishes members early in fine-tuning, while the naive denoising-loss baseline remains near chance. Controlled comparisons show that the membership signal is concentrated in the signed first-order response, with little discrimination from prediction-shift magnitude alone.

1 INTRODUCTION

Text-to-image diffusion models are trained on large collections of paired images and captions (Rombach et al., 2022) and can memorize and reproduce training images (Carlini et al., 2023; Somepalli et al., 2023b; Hong et al., 2024). Because these collections may contain private, copyrighted, or otherwise unauthorized material, determining whether a candidate example was used in training is important for auditing data provenance and training-data use. Membership inference attacks (MIAs) determine whether a queried example participated in training (Shokri et al., 2017) and can provide such an audit for text-to-image diffusion models (Zhai et al., 2024; Qi et al., 2026). We study a gray-box setting in which an auditor can query the noise predictor at chosen noisy inputs, timesteps, and text conditions, but cannot inspect model weights or obtain target-model gradients. This access enables direct evaluation of conditional denoising fit without differentiating through the target model.

Several diffusion-model MIAs use denoising or reconstruction errors as their primary membership signals (Duan et al., 2023; Kong et al., 2024; Tang et al., 2023). Denoising loss is a natural starting point because it is the training objective and is connected to likelihood through the variational bound (Ho et al., 2020). Generalization allows diffusion models to denoise unseen images accurately, making absolute denoising error alone an incomplete indicator of membership (Zhai et al., 2024). CLiD (Zhai et al., 2024) extends this view to the relationship between an image and its text condition. Motivated by conditional overfitting, it compares losses under the original caption and modified conditions. Its results suggest that membership information can reside in the learned image-text relationship, beyond denoising accuracy under a single condition.

Condition-dependent evidence also appears in a model's response to changes in the text condition. SD-MIA studies pre-training membership under black-box access by perturbing the original textual instruction and measuring the relevance of generated images to the queried image (Qi et al., 2026). It uses behavior across text conditions as membership evidence, although its observations and access setting differ from those of CLiD.

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

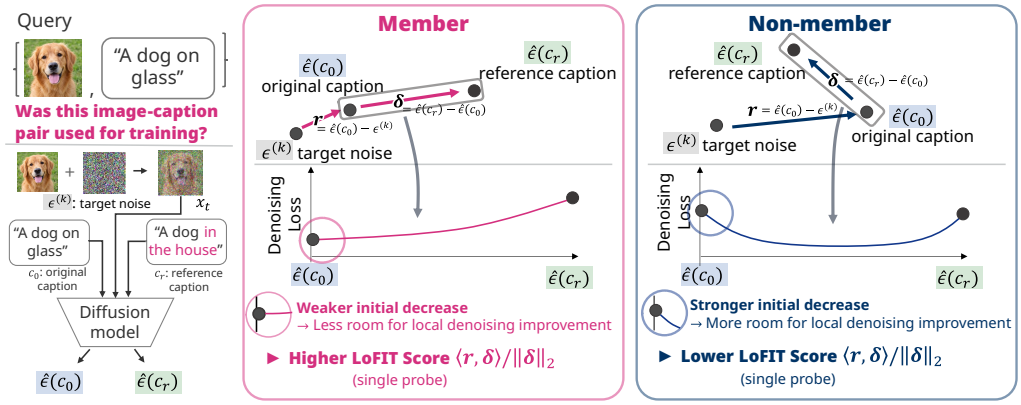


Figure 1: Illustration of LoFIT for a queried image-caption pair. At a fixed noisy input, r denotes the original-caption prediction error relative to the target noise, and δ the prediction shift toward a reference-caption prediction. LoFIT measures the initial denoising-loss response along this shift through the normalized alignment $\langle r, \delta \rangle / \|\delta\|_2$. Members exhibit a weaker initial loss decrease than non-members, indicating less room for local denoising improvement and yielding a higher score. The curves are conceptual; measured class distributions appear in Figure 6.

CLiD and SD-MIA provide complementary motivations. CLiD shows that membership information can lie in the image-text relationship, while SD-MIA shows that changing the text condition can expose it. These findings motivate us to ask how a change in text condition affects the denoising fit for the same image. With access to the noise predictor, we examine this question by comparing predictions across conditions while holding the noisy input and target noise fixed.

An alternative condition moves the prediction by a condition-induced displacement, while the original prediction is separated from the target noise by the denoising residual. Denoising loss initially decreases when the displacement has a component opposite to the residual, and increases when it has a component along it. Thus the direction of the displacement relative to the residual determines the sign of the initial loss response; its length alone does not. We call this directional relationship *local conditional fit*. Here, local refers to the initial loss response at the original prediction, even when the two captions differ substantially.

This geometry alone does not imply a membership difference. Because training optimizes denoising loss on observed image-caption pairs, we hypothesize that alternative conditions offer less initial denoising improvement for members than for non-members (Figure 1). We quantify this response with *LoFIT* (Local Conditional Fit): it measures the signed alignment between each condition-induced displacement and the denoising residual per unit displacement, then averages these responses across text conditions and diffusion states. A higher score indicates less initial reduction in loss toward alternative predictions. LoFIT requires only forward queries to the denoiser, without target-model gradients.

We evaluate LoFIT on four datasets spanning fine-tuned and pretrained diffusion models. It achieves AUROC 0.935, 0.954, and 0.896 on the three fine-tuned SD v1.4 models, outperforming the evaluated baselines, including CLiD (Zhai et al., 2024) and PFAMI (Fu et al., 2024) (Section 4.2). Across training checkpoints, its discrimination strengthens as fixed image-caption pairs are used for training, while absolute denoising loss remains a weak discriminator (Section 5.1). On identical predictions, controlled decompositions locate membership separation in the first-order alignment response (Section 5.3). Matched-budget comparisons show LoFIT's advantage at equal query cost (Section 4.2), and direction ablations show that retaining the sign improves discrimination (Appendix B.3).

Our contributions are:

- We identify *local conditional fit*, the directional relationship between a condition-induced displacement and the original denoising residual, as a membership signal, and examine how its discrimination changes with training exposure.
- We introduce *LoFIT*, a score that quantifies this signal from forward predictions alone, requiring no access to target-model gradients.

- We evaluate LoFIT on four evaluation sets and use controlled comparisons to identify the contribution of its signed first-order response.

2 Background and Related Work

2.1 Related work

Denoising-based membership inference. SecMI uses deterministic reconstruction errors at intermediate steps (Duan et al., 2023). PIA uses the noise predicted at $t = 0$ to perturb the image, then measures how accurately the denoiser recovers that noise at a later timestep (Kong et al., 2024). SimA uses the norm of the predicted score as its membership statistic (Rao et al., 2025). Quantile regression learns image-specific reconstruction-loss thresholds from non-member data and identifies members whose loss falls below the predicted threshold (Tang et al., 2023). Our method also uses denoising predictions, but probes how their error changes along normalized prediction shifts on a shared noisy input.

Conditional comparisons. Text conditioning influences training-image replication (Somepalli et al., 2023a), and the magnitude of text-conditional predictions can identify memorized prompts (Wen et al., 2024). CLiD compares denoising losses under the original caption and alternative conditions to probe the learned image-caption relationship (Zhai et al., 2024). MoFit addresses missing training captions by optimizing a surrogate image and conditioning embedding, then measuring conditional versus unconditional loss discrepancy (Jeon et al., 2026). Its gradient-based optimization requires stronger model access than our forward-only interface. We use the supplied caption and probe local conditional fit without such optimization. Appendix B relates our score to conditional loss gaps.

Generated-image signals. Black-box attacks infer membership from generated images without accessing model weights or intermediate denoising predictions (Li et al., 2025). Noise as a Probe injects image semantics into the initial noise (Lian et al., 2026). SD-MIA perturbs textual instructions and measures how the generated images remain related to the target image (Qi et al., 2026).

2.2 From conditional membership signals to local conditional fit

We examine CLiD and SD-MIA more closely because they connect membership evidence to two aspects of conditioning: how well the original caption fits the image and how model behavior changes when the condition changes.

CLiD: conditional likelihood discrepancy. CLiD (Zhai et al., 2024) asks how much the supplied caption increases the likelihood of an image:

$$I_\theta(x_0, c_0) = \log p_\theta(x_0|c_0) - \log p_\theta(x_0). \tag{1}$$

This is the pointwise mutual information under the model's joint distribution; its expectation gives mutual information. Under CLiD's conditional-overfitting assumption, this discrepancy provides membership evidence beyond the image's marginal likelihood.

CLiD approximates this quantity through the diffusion variational bound, using denoising-loss differences. It also compares against reduced captions c to mitigate estimation variability:

$$\mathcal{D}_{\mathrm{CLiD}}(x_0, c_0; c) = \mathbb{E}_{t,\epsilon}\left[\|\epsilon_\theta(x_t, t, c) - \epsilon\|_2^2 - \|\epsilon_\theta(x_t, t, c_0) - \epsilon\|_2^2\right]. \tag{2}$$

Here x_t is the noisy image at timestep t, and $\epsilon \sim \mathcal{N}(0, I)$ is the injected noise; both predictions use the same x_t, t, and ϵ. The empty caption supplies the unconditional comparison, while other reduced conditions retain partial caption information. A larger gap means that the original caption improves denoising more than the reduced condition. CLiD combines these gaps with the original-caption loss, using either a thresholded scalar score or a learned classifier. Its central contribution is to exploit the learned image-caption association as membership evidence, beyond absolute denoising accuracy.

SD-MIA: stability under textual perturbations. SD-MIA (Qi et al., 2026) studies whether the image-caption association persists under textual perturbations. It argues that locally overfitted text-to-image mappings can make generations for members less sensitive to such changes. For a caption embedding e_0 and a small perturbation h, its motivating sensitivity is

$$\left|p_\theta(x_0|e_0+h) - p_\theta(x_0|e_0)\right| \approx \left|\langle \nabla_e p_\theta(x_0|e_0), h\rangle\right| \approx 0. \tag{3}$$

This probability and its gradient are unavailable under SD-MIA's generation-only access. Instead, it constructs coherent caption variants through token, style, and semantic edits, and contrasts the CLIP (Radford et al., 2021) relevance of their generated images to the target with that under the original caption. Its signal therefore concerns how well generated content remains associated with the target when conditioning changes, without observing denoising losses or target-model gradients.

Local conditional fit hypothesis. Together, conditional overfitting and stability under textual perturbations motivate our hypothesis: member captions are closer to a denoising optimum relative to alternative conditions. Training directly minimizes denoising loss on observed image-caption pairs, suggesting less potential for local denoising improvement for members. Section 3 measures this response along alternative-caption prediction shifts while holding the noisy input and target noise fixed.

3 Measuring Local Conditional Fit

We assess whether member captions are closer to a denoising optimum by measuring the local loss response toward alternative-caption predictions. Denoising loss is the training objective and is connected to likelihood through the diffusion variational bound (Ho et al., 2020).

For a queried pair (x_0, c_0), let $\mathcal{C} = \operatorname{supp} p(c)$ be the support of the condition prior and $\mathcal{U}(c_0) \subseteq \mathcal{C}$ a set of comparison captions containing c_0. At each timestep t and noise draw $\epsilon \sim \mathcal{N}(0, I)$, all captions use the same noisy input x_t and target noise ϵ. The denoising loss and optimality criterion are

$$\begin{aligned} \mathcal{L}(x_0, c) &= \mathbb{E}_{t,\epsilon}\left[\|\epsilon_\theta(x_t, t, c) - \epsilon\|_2^2\right], \\ \mathcal{L}(x_0, c_0) &\le \mathcal{L}(x_0, c) \quad \text{for all } c \in \mathcal{U}(c_0). \end{aligned} \tag{4}$$

This criterion concerns optimality over the comparison set. Locality in LoFIT refers to infinitesimal movement in prediction space. We measure the response at each noisy state, then aggregate across states and references.

3.1 A first-order criterion for conditional fit

Squared error is convex in prediction space. For a given noisy image x_t, timestep t, and noise ϵ, the following first-order condition is sufficient for c_0 to minimize the denoising loss over $\mathcal{U}(c_0)$ (Boyd & Vandenberghe, 2004, Section 4.2.3):

$$\begin{aligned} &\left\langle \epsilon_\theta(x_t, t, c_0) - \epsilon,\ \epsilon_\theta(x_t, t, c) - \epsilon_\theta(x_t, t, c_0)\right\rangle \ge 0, \\ &\qquad \text{for all } c \in \mathcal{U}(c_0), \\ \implies &\quad c_0 \in \arg\min_{c \in \mathcal{U}(c_0)} \|\epsilon_\theta(x_t, t, c) - \epsilon\|_2^2. \end{aligned} \tag{5}$$

The inner product aligns prediction error with the conditional prediction shift. Nonnegative alignment for every comparison caption ensures that none yields a lower loss. If this holds at every noisy state, averaging gives Equation 4. This criterion motivates our alignment score.

3.2 Measuring alignment

Checking the condition over all captions is impractical. LoFIT instead evaluates the inner product using a finite set of reference captions from $\mathcal{U}(c_0)$. For a reference caption c, the original prediction error r and the conditional prediction shift δ are

$$\begin{aligned} r &= \epsilon_\theta(x_t, t, c_0) - \epsilon, \\ \delta &= \epsilon_\theta(x_t, t, c) - \epsilon_\theta(x_t, t, c_0). \end{aligned} \tag{6}$$

For $\delta \neq 0$, we use the *normalized prediction shift* $\delta/\|\delta\|_2$ to compare directions independently of shift magnitude. The *alignment response* is

$$v = \frac{\langle r, \delta \rangle}{\|\delta\|_2}. \tag{7}$$

Normalization preserves the sign of the inner product. The normalized response can also be interpreted as half the directional derivative of the denoising loss along the normalized prediction shift $\delta/\|\delta\|_2$. For a step of length s toward the reference prediction,

$$\frac{\mathrm{d}}{\mathrm{d}s} \left\| r + s\frac{\delta}{\|\delta\|_2} \right\|_2^2 \Bigg|_{s=0} = 2\frac{\langle r, \delta \rangle}{\|\delta\|_2} = 2v. \tag{8}$$

Thus $v < 0$ indicates an initial loss decrease toward the reference prediction, while $v \geq 0$ indicates no first-order decrease. At the reference prediction, $s = \|\delta\|_2$; the finite-step loss change $2sv + s^2$ can therefore be positive even when $v < 0$. CLiD's discrepancy term uses this finite loss gap, whereas LoFIT isolates the initial response (Section 5.3). LoFIT's response is local in prediction space even when the two captions differ substantially.

Score composition gives this local response a conditional interpretation. Compositional diffusion and classifier-free guidance combine predictions under different conditions to control generation (Liu et al., 2022; Ho & Salimans, 2022). At a fixed noise level, a convex combination of exact conditional scores is itself the score of a normalized geometric mixture of their densities. Noise predictions estimate these scores up to a shared scale, so interpolating predictions forms composed score estimates. LoFIT tests whether a small contribution from an alternative condition initially reduces the original denoising error, measuring local fit within this composed family.

3.3 Aggregating alignment responses

To compute LoFIT, let $\mathcal{R}$ index reference captions c_j, including caption edits, randomly sampled captions, and the empty prompt (Appendix A.1). We evaluate T timesteps with K Gaussian noise draws each, using $i = (t, k)$ to index the TK states. For each draw $\epsilon^{(k)} \sim \mathcal{N}(0, I)$, we form

$$x_{t,k} = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1 - \bar{\alpha}_t}\, \epsilon^{(k)}. \tag{9}$$

Here $\bar{\alpha}_t$ is the diffusion noise schedule, and x_0 is the queried image or its latent representation. Each comparison uses the same noisy image $x_{t,k}$ and target noise $\epsilon^{(k)}$ for both captions. At state i, r_i denotes the original prediction error, while δ_{ij} and v_{ij} denote the prediction shift and alignment for reference j. Let $\mathcal{A}$ be the set of valid state-reference pairs (i, j) after omitting zero prediction shifts; Appendix A.2 specifies the numerical tolerance. The LoFIT score is

$$S(x_0, c_0) = \begin{cases} |\mathcal{A}|^{-1} \displaystyle\sum_{(i,j)\in\mathcal{A}} v_{ij}, & |\mathcal{A}| > 0, \\ 0, & |\mathcal{A}| = 0. \end{cases} \tag{10}$$

When every pair is valid, $|\mathcal{A}| = TK|\mathcal{R}|$. Higher S indicates less initial denoising improvement and provides membership evidence. The query cost is $TK(1 + |\mathcal{R}|)$, reusing the original prediction at each state. Appendices B.1 and B.3 analyze loss minima and the contribution of the response sign.

4 Experiments

We evaluate LoFIT on the full evaluation sets and compare it with published attacks. Controlled readout comparisons use the same reference predictions and diffusion states.

4.1 Settings and comparisons

Reference construction and score computation. References comprise randomly sampled captions, the empty prompt, and language-model edits. LoFIT computes each score directly from forward predictions using a fixed reference-construction procedure, without fitting an attack classifier. The configurations below specify the reference sets and query allocations.

Table 1: Full-set membership inference on fine-tuned SD v1.4, SD v2.1, and SD-XL. Bold marks the best value per backbone and metric; dashes denote unreported results among the listed methods.

	Flickr30k		COCO		Pokemon	
Method	AUROC	TPR@1%FPR	AUROC	TPR@1%FPR	AUROC	TPR@1%FPR
SD v1.4						
Naive loss	0.524	0.013	0.527	0.018	0.576	0.029
PIA	0.507	0.010	0.503	0.012	0.552	0.027
SecMI	0.551	0.017	0.562	0.012	0.538	0.022
SD-MIA	0.519	0.013	0.517	0.013	0.502	0.019
PFAMI	0.659	0.052	0.715	0.073	0.685	0.031
CLiD_{th}	0.718	0.063	0.745	0.072	0.695	0.058
CLiD_{vec}	0.718	0.064	0.751	0.073	0.642	0.043
LoFIT (ours)	**0.935**	**0.454**	**0.954**	**0.478**	**0.896**	**0.272**
SD v2.1						
Naive loss	0.519	0.012	0.516	0.014	0.557	0.019
PIA	0.505	0.011	0.500	0.008	0.546	0.029
SecMI	0.537	0.013	0.530	0.010	0.503	0.017
PFAMI	0.631	0.038	0.653	0.060	0.631	0.029
CLiD_{th}	0.705	0.088	0.681	0.070	0.686	0.026
CLiD_{vec}	0.783	0.131	0.739	0.079	0.684	0.022
LoFIT (ours)	**0.894**	**0.296**	**0.851**	**0.166**	**0.812**	**0.130**
SD-XL						
Naive loss	0.514	0.011	0.516	0.010	0.530	0.014
PIA	0.498	0.009	0.501	0.008	0.517	0.012
SecMI	0.515	0.011	0.506	0.012	0.477	0.019
PFAMI	0.660	0.069	0.657	0.052	0.664	0.048
CLiD_{th}	0.866	0.188	0.819	0.114	0.804	0.127
CLiD_{vec}	0.880	0.217	0.836	0.121	0.784	0.084
LoFIT (ours)	**0.978**	**0.719**	**0.969**	**0.631**	**0.883**	**0.351**

Models and data. We evaluate fine-tuned SD v1.4, SD v2.1-base (ϵ-prediction), and SD-XL base 1.0 on Flickr30k (Young et al., 2014), COCO (Lin et al., 2014), and Pokemon (Lambda, 2023), with $10{,}000/10{,}000$, $2{,}500/2{,}500$, and $416/417$ members/non-members, respectively. LAION-MI uses pretrained SD v1.5 with $6{,}400$ examples per class, matched in CLIP (Radford et al., 2021) image-text similarity using 20 bins (Dubiński et al., 2024). Appendix A gives the training configurations, and Appendix A.4 details LAION-MI matching and preprocessing.

Default evaluation configuration. All three fine-tuned backbones use the same diffusion-state grids and noise-draw counts. SD v2.1 and SD-XL use the reduced reference sets in Appendix A.2. On the three fine-tuned SD v1.4 models, we use all eligible reference families: random captions, the empty prompt, and language-model rewrites. LAION-MI uses six randomly sampled captions and the empty prompt ($|\mathcal{R}| = 7$); adding rewrites weakens discrimination there (Appendix C.1). Flickr30k and LAION-MI use $T = 12$ timesteps $\{60, 120, \ldots, 720\}$ and $K = 4$ noise draws; COCO and Pokemon use $T = 4$ timesteps (150, 300, 450, 600) and $K = 4$. On Flickr30k/COCO/Pokemon, SD v1.4 requires $2{,}064/752/1{,}088$ denoiser evaluations per image, and SD v2.1 and SD-XL require $528/176/272$; LAION-MI requires 384. These counts include the original caption.

Metrics. We report AUROC and true-positive rates (TPR) at 1% false-positive rate (FPR) for fine-tuned models and 5% FPR for LAION-MI.

Baselines and controlled comparisons. We compare with naive loss, PIA (Kong et al., 2024), SecMI (Duan et al., 2023), SD-MIA (Qi et al., 2026), CLiD (Zhai et al., 2024), and PFAMI (Fu et al., 2024). CLiD uses character-based caption thirds and the empty prompt at three fixed timesteps. PFAMI measures relative loss fluctuations under image crops, with calibration by the corresponding pretrained backbone on fine-tuned targets and no calibration on LAION-MI. Appendix A.3 details these settings and deviations from the original protocols. To isolate the contribution of LoFIT's scoring rule, we also evaluate CLiD variants using our expanded reference captions, including LLM

rewrites and random captions, or our timestep configuration. Table 5 compares readouts on identical predictions, while Section 4.2 examines query budgets.

4.2 RESULTS

Membership inference performance. LoFIT outperforms the evaluated baselines in both AUROC and low-FPR TPR on all three datasets with fine-tuned SD v1.4, SD v2.1, and SD-XL (Table 1). On pretrained LAION-MI, it achieves the highest reported AUROC while attaining TPR comparable to SD-MIA at 5% FPR (Table 2).

Table 2: Membership inference on pretrained SD v1.5 (LAION-MI). Bold marks the best values; underlining marks the second-best TPR.

Method	AUROC	TPR@5%FPR
Naive loss	0.527	0.072
PIA	0.521	0.072
SecMI	0.512	0.063
SD-MIA	0.571	**0.112**
PFAMI	0.516	0.066
CLiD_{th}	0.557	0.075
CLiD_{vec}	0.548	0.076
LoFIT (ours)	**0.602**	<u>0.101</u>

Query efficiency. LoFIT outperforms CLiD_{th} at matched query budgets, including variants using our captions or timestep grid (Figure 2). At 15 denoiser evaluations, AUROC is 0.834 versus 0.767 for the strongest CLiD variant on Flickr30k, and 0.580 versus 0.559 on LAION-MI. Appendix A.3 specifies calibration.

On identical reference predictions and diffusion states, LoFIT outperforms CLiD_{th} with both fixed $\alpha = 0.5$ and oracle α (Table 3). With all reference families, its AUROC advantage over oracle CLiD_{th} is 0.081–0.130 on the fine-tuned models at the same query cost. Appendix A.3 specifies scaling and oracle selection.

With the empty reference alone, LoFIT also outperforms both CLiD_{th} variants across all four datasets. On Flickr30k and COCO, empty-only LoFIT exceeds even oracle CLiD_{th} using all reference families.

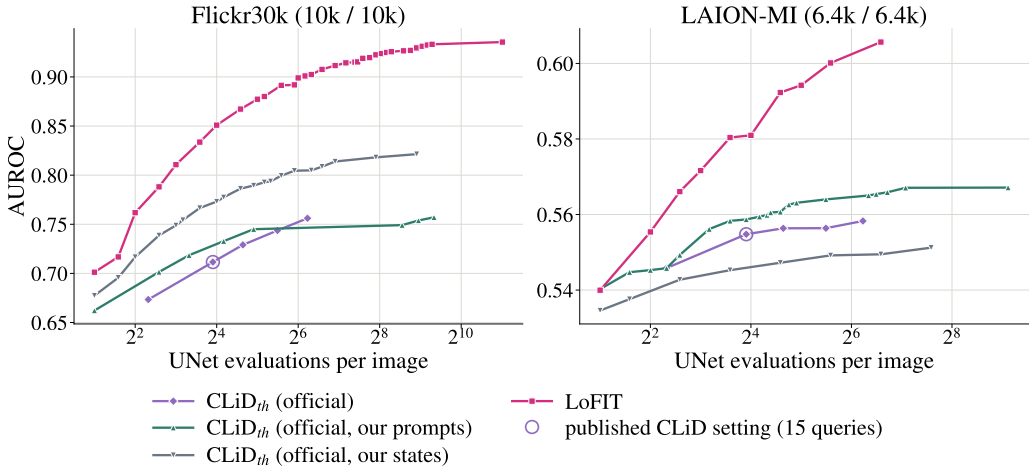


Figure 2: Query efficiency on Flickr30k and LAION-MI. CLiD curves use the released statistic with two-fold cross-fitted calibration (Appendix A.3); "our prompts" and "our states" denote our reference captions and timestep grid. Open circles mark the original 15-query configuration.

Image membership with generated captions. With BLIP-2 OPT-6.7b (Li et al., 2023) captions, prompt-based LoFIT achieves higher AUROC than official CLiD_{th} with oracle α. Using random captions and the empty reference, LoFIT reaches 0.641, 0.669, and 0.621 AUROC on fine-tuned SD v1.4 for Flickr30k, COCO, and Pokemon, respectively, compared with 0.603, 0.650, and 0.576 for CLiD_{th}. On LAION-MI, the corresponding AUROC values are 0.546 and 0.535. Gaussian conditioning probes provide further improvements on the fine-tuned models. Appendix C.3 reports the setup and full comparisons.

Table 3: LoFIT and CLiD_{th} on identical reference predictions and diffusion states. Bold marks the best value in each reference group.

	Flickr30k		COCO		Pokemon		LAION-MI	
Method	AUROC	TPR@1% FPR	AUROC	TPR@1% FPR	AUROC	TPR@1% FPR	AUROC	TPR@5% FPR
All reference families (LAION-MI: 36 references; Table 2 uses 7)								
LoFIT	**0.935**	**0.454**	**0.954**	**0.478**	**0.896**	**0.272**	**0.573**	**0.085**
CLiD_{th} ($\alpha = 0.5$)	0.731	0.106	0.752	0.130	0.764	0.058	0.563	0.084
CLiD_{th} (oracle α)	0.813	0.092	0.824	0.083	0.815	0.079	0.569	0.083
Empty reference only								
LoFIT	**0.900**	**0.268**	**0.914**	**0.287**	**0.781**	**0.070**	**0.606**	**0.094**
CLiD_{th} ($\alpha = 0.5$)	0.730	0.112	0.736	0.094	0.720	0.053	0.549	0.089
CLiD_{th} (oracle α)	0.806	0.146	0.799	0.097	0.756	0.053	0.553	0.081

5 Understanding the Membership Signal

The following analyses use the SD v1.4 fine-tunes and pretrained SD v1.5.

5.1 Early membership detection during fine-tuning

LoFIT detects membership earlier and approaches saturation sooner than official CLiD_{th} (Figure 3). At 20k steps, LoFIT achieves AUROC 0.917 and TPR 0.353 at 1% FPR. By 40k, these reach 0.994 and 0.875, while CLiD_{th} continues improving through 80k steps. Absolute denoising loss remains a weak discriminator across these checkpoints. We evaluate the same 10,000 member and 10,000 non-member Flickr30k pairs throughout. LoFIT keeps its references and 12×4 diffusion states fixed; CLiD_{th} uses the released caption reductions and three-state configuration.

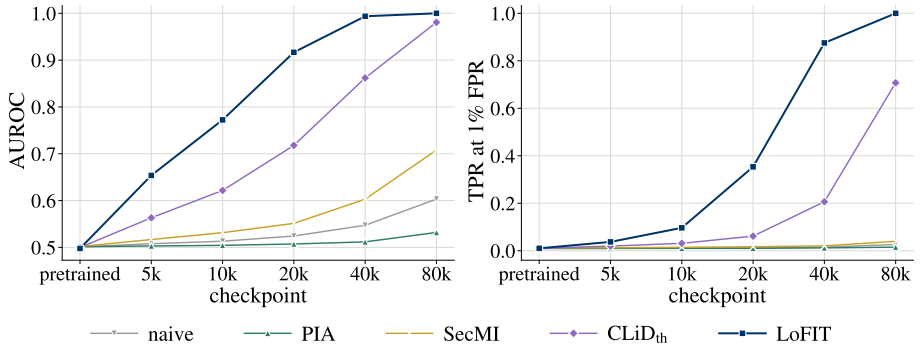


Figure 3: Membership inference across Flickr30k fine-tuning checkpoints, including official CLiD_{th}. Membership denotes participation in this fine-tuning run.

5.2 Caption and Gaussian conditioning probes

Table 4 compares caption and Gaussian probes at equal single-reference cost ($2TK$ denoiser evaluations). Both use the same images, diffusion-state grids, and alignment score; original-caption losses match across all datasets (Appendix C.2). AUROC is averaged across references within each group.

Table 4: Single-reference AUROC for combined caption edits and Gaussian probes.

Reference directions	Flickr30k	COCO	Pokemon	LAION-MI
Combined edits	0.880	0.909	0.810	0.569
Gaussian, $h = 160$	0.854	0.903	0.805	0.534

Gaussian probes approach combined edits on COCO and Pokemon, but remain weak on pretrained LAION-MI (0.534, versus 0.606 for the empty caption). Pooling caption and Gaussian responses further improves discrimination on the fine-tuned models (Appendix C.2). Gaussian probes require direct modification of conditioning embeddings, extending access beyond LoFIT's text-condition gray-box interface.

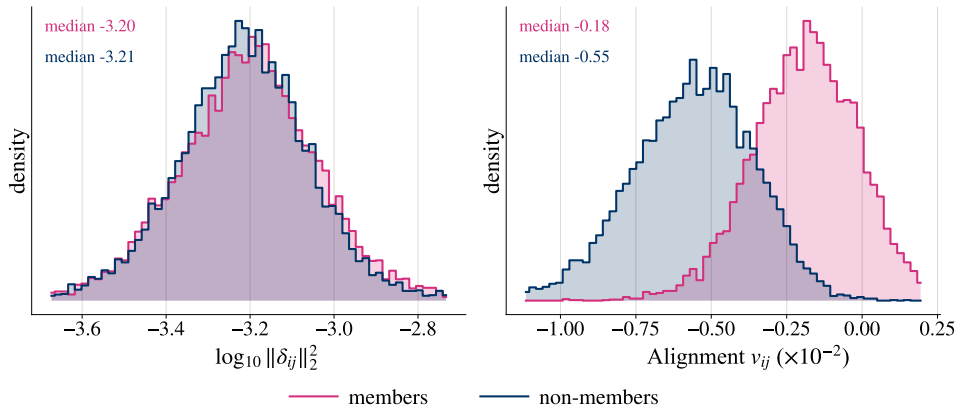


Figure 4: Conditional prediction statistics on Flickr30k. Left: distributions of squared shift magnitude $\|\delta_{ij}\|_2^2$ on a logarithmic scale. Right: normalized residual interaction $\langle r_i, \delta_{ij}\rangle / \|\delta_{ij}\|_2$, the LoFIT response. Shift magnitudes largely overlap, while the first-order response separates the classes. Magenta denotes members and blue denotes non-members. Full profiles appear in Appendix B.2.

5.3 Alignment isolates the directional membership signal

The discrepancy term in CLiD measures conditional denoising-loss gaps (Zhai et al., 2024), while LoFIT isolates their first-order response. For a fixed noisy image and target noise, the gap decomposes as

$$\Delta_{ij} = \|r_i + \delta_{ij}\|_2^2 - \|r_i\|_2^2 = 2\langle r_i, \delta_{ij}\rangle + \|\delta_{ij}\|_2^2, \tag{11}$$

$$\frac{\Delta_{ij}}{\|\delta_{ij}\|_2} = \underbrace{2v_{ij}}_{\text{first-order response}} + \underbrace{\|\delta_{ij}\|_2}_{\text{shift magnitude}}. \tag{12}$$

The normalized gap adds shift magnitude to LoFIT's first-order response $2v_{ij}$. Shift magnitudes largely overlap between the classes, whereas the response is more negative for non-members (Figure 4; full profiles in Appendix B.2).

On identical predictions under the default configuration, LoFIT outperforms both raw and normalized gaps in AUROC and low-FPR TPR across all four datasets (Table 5). Isolating the first-order response therefore improves discrimination beyond normalization alone.

Table 5: Readout comparison on identical predictions and diffusion states. Bold marks the best value per metric.

	Flickr30k		COCO		Pokemon		LAION-MI	
Readout	AUROC	TPR@1% FPR	AUROC	TPR@1% FPR	AUROC	TPR@1% FPR	AUROC	TPR@5% FPR
Loss gap Δ_{ij}	0.813	0.092	0.818	0.078	0.804	0.065	0.549	0.076
Normalized gap	0.906	0.334	0.930	0.380	0.882	0.127	0.567	0.085
Shift magnitude $\|\delta_{ij}\|_2$	0.503	0.012	0.487	0.014	0.457	0.024	0.480	0.053
LoFIT	**0.935**	**0.454**	**0.954**	**0.478**	**0.896**	**0.272**	**0.602**	**0.101**

6 Conclusion

Members exhibit less potential for local denoising improvement along alternative-caption prediction shifts. LoFIT measures this signature through normalized first-order responses using only forward predictions. Across fine-tuned and pretrained diffusion models, these responses provide membership information beyond absolute denoising loss.

AI USE STATEMENT

We used generative AI tools to assist with experimental design and manuscript drafting and editing. These tools also assisted with LaTeX formatting and the preparation of tables and figures. We used `gpt-4.1-mini` to generate reference-caption variants, as described in Appendix A.1. The authors take responsibility for the final content of this work, including all AI-assisted text and figures.

ETHICS STATEMENT

This work studies privacy leakage through membership inference in text-to-image diffusion models. LoFIT supports audits of training-data use and assessments of privacy exposure. The same capability can reveal information about the inclusion of image-caption pairs in training data. We intend this method for authorized privacy audits and research on privacy protection, with appropriate safeguards for the queried data.

REPRODUCIBILITY STATEMENT

Section 3 defines the LoFIT score and its computation. Section 4 and Appendix A document target-model training, data splits, preprocessing, diffusion states, reference construction, and baseline calibration. Controlled comparisons specify the reference sets, states, and readouts used in each configuration. Appendices B and C provide additional analyses of the membership signal and reference construction.

REFERENCES

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. ISBN 978-0-521-83378-3.

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 *32nd USENIX Security Symposium*, 2023.

Jinhao Duan, Fei Kong, Shiqi Wang, Xiaoshuang Shi, and Kaidi Xu. Are diffusion models vulnerable to membership inference attacks? In *International Conference on Machine Learning*, 2023.

Jan Dubiński, Antoni Kowalczuk, Stanisław Pawlak, Przemysław Rokita, Tomasz Trzciński, and Paweł Morawiecki. Towards more realistic membership inference attacks on large diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024.

Wenjie Fu, Huandong Wang, Liyuan Zhang, Chen Gao, Yong Li, and Tao Jiang. A probabilistic fluctuation based membership inference attack for diffusion models. *arXiv preprint arXiv:2308.12143*, 2024.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Chunsan Hong, Tae-Hyun Oh, and Minhyuk Sung. Membench: Memorized image trigger prompt dataset for diffusion models. *arXiv preprint arXiv:2407.17095*, 2024.

Joonsung Jeon, Woo Jae Kim, Suhyeon Ha, Sooel Son, and Sung eui Yoon. No caption, no problem: Caption-free membership inference via model-fitted embeddings. In *International Conference on Learning Representations*, 2026.

Fei Kong, Jinhao Duan, RuiPeng Ma, Hengtao Shen, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. An efficient membership inference attack for the diffusion model by proximal initialization. In *International Conference on Learning Representations*, 2024.

Lambda. Pokemon-blip-captions. `https://huggingface.co/datasets/lambdalabs/pokemon-blip-captions/`, 2023.

Jingwei Li, Jing Dong, Tianxing He, and Jingzhao Zhang. Towards black-box membership inference attack for diffusion models. In *International Conference on Machine Learning*, 2025.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, 2023.

Puwei Lian, Yujun Cai, Songze Li, and Bingkun Bao. Noise as a probe: Membership inference attacks on diffusion models leveraging initial noise. *arXiv preprint arXiv:2601.21628*, 2026.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, 2014.

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *Proceedings of the European Conference on Computer Vision*, 2022.

Tao Qi, Huili Wang, Yuanhong Huang, Wendan Wang, Lianchao Zhao, Jinrui Wang, Zichen Qin, Shangguang Wang, and Yongfeng Huang. Black-box membership inference attacks on the pre-training data of image-generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Mingxing Rao, Bowen Qu, and Daniel Moyer. Score-based membership inference on diffusion models. *arXiv preprint arXiv:2509.25003*, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy*, 2017.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and mitigating copying in diffusion models. In *Advances in Neural Information Processing Systems*, 2023a.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023b.

Shuai Tang, Zhiwei Steven Wu, Sergul Aydore, Michael Kearns, and Aaron Roth. Membership inference attacks on diffusion models via quantile regression. *arXiv preprint arXiv:2312.05140*, 2023.

Yuxin Wen, Yuchen Liu, Chen Chen, and Lingjuan Lyu. Detecting, explaining, and mitigating memorization in diffusion models. In *International Conference on Learning Representations*, 2024.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. doi: 10.1162/tacl_a_00166.

Shengfang Zhai, Huanran Chen, Yinpeng Dong, Jiajun Li, Qingni Shen, Yansong Gao, Hang Su, and Yang Liu. Membership inference on text-to-image diffusion models via conditional likelihood discrepancy. In *Advances in Neural Information Processing Systems*, 2024.

A EXPERIMENTAL PROTOCOLS

This appendix documents target-model training, data construction, reference captions, and baseline evaluation.

Training configurations. We fine-tune the full UNet using the diffusers text-to-image training script, with the VAE and text encoders frozen. The tokenizer and diffusion noise schedule remain fixed. All nine fine-tuned targets use AdamW with learning rate 10^{-5}, $\beta = (0.9, 0.999)$, weight decay 0.01, and optimizer epsilon 10^{-8}. The learning rate is constant with no warmup. Training minimizes epsilon-prediction MSE with timesteps sampled uniformly from a $1{,}000$-step DDPM schedule, without SNR weighting, caption dropout, or noise offset. We use FP16 mixed precision, gradient checkpointing, and gradient clipping at norm 1.0. EMA weights are maintained during training; all attacks evaluate the raw UNet weights.

SD-XL starts from `stabilityai/stable-diffusion-xl-base-1.0` and follows the shared training recipe at 1024×1024 resolution. Table 6 lists the evaluated checkpoints and effective batch sizes. Steps count optimizer updates, and batch sizes include all devices and gradient accumulation. SD v2.1 Flickr30k uses training seed 0; the SD v1.4 runs and SD v2.1 COCO/Pokemon do not set a training seed. The data split seed is 0 for every backbone.

Table 6: Target-model training runs and evaluated checkpoints.

Model	Dataset	Effective batch	Evaluated update
SD v1.4	Flickr30k	16	20,000
	COCO	256	1,000
	Pokemon	256	200
SD v2.1	Flickr30k	16	20,000
	COCO	256	1,000
	Pokemon	256	200
SD-XL	Flickr30k	16	20,000
	COCO	256	1,000
	Pokemon	256	200

Image splits and captions. We shuffle each dataset once with seed 0, assign the first n images to members, and take the next n as non-members, with $n = 10{,}000$ for Flickr30k and $n = 2{,}500$ for COCO. Pokemon uses the first 416 images as members and the remaining 417 as non-members. Each source row contains one image; all captions associated with a row remain on the same side of the split. We retain only the first Flickr30k caption before shuffling. Every image has one fixed caption throughout training, and the same string serves as its original caption during evaluation. The caption columns are `caption`, `captions`, and `text` for Flickr30k, COCO, and Pokemon, respectively.

Preprocessing. We resize the shorter image side to 1024 pixels for SD-XL and 512 for SD v1.4, SD v1.5, and SD v2.1, then apply a square center crop at that resolution. Training adds horizontal flipping with probability 0.5 and normalizes pixels to $[-1, 1]$; attack evaluation omits flipping. LoFIT, naive loss, PIA, and SecMI use the VAE posterior mean. The official CLiD implementation instead samples from the VAE posterior and normalizes each channel with mean and standard deviation 0.5.

Flickr30k shadow model. After the same shuffle, zero-based rows $20{,}000$–$29{,}999$ provide $10{,}000$ shadow members and rows $30{,}000$–$31{,}782$ provide $1{,}783$ shadow non-members. Both groups are disjoint from the target-model training and evaluation images. The SD v1.4 shadow model uses the target's training recipe, effective batch size 16, and $20{,}000$ updates. Its member and non-member features train the CLiD_{vec} classifier.

A.1 REFERENCE CONSTRUCTION

Our reference families extend CLiD's caption reduction and embedding perturbations (Zhai et al., 2024). The language-model edits are inspired by SD-MIA's textual perturbations (Qi et al., 2026) and use `gpt-4.1-mini`. Table 7 summarizes their construction. The same procedure is applied to members and non-members. The expanded pools combine language-model edits, random captions, and the empty prompt. Case-only edits, word shuffling, and negation are excluded.

Table 7: Reference-caption families and their construction.

Prompt family	Construction
Standard language-model rewrites	
Token-level edits	Small wording changes, including one- or two-word synonym substitutions that preserve meaning.
Paraphrases	Rephrase the caption while preserving its meaning, including light and heavy rewrites with increasing changes in wording.
Style changes	Change how the caption is presented.
Semantic edits	Change depicted entities or actions, including variants that alter one, two, or three or more content words.
Combined edits	Change both wording and semantic content.
Additional references	
Random captions	Randomly sample M captions of other images from the dataset's pooled caption corpus; exclude identical strings.
Empty prompt	Use the empty string $\varnothing$.

Caption-rewriting prompts. The templates below share a system prompt. Here, {`cap`} is the original caption, {`n`} is the requested number of variants per family, and `H_both` denotes combined edits.

System prompt

```
You rewrite image captions for a controlled experiment. Answer with
STRICT JSON only, no prose. Keep every variant a natural image caption
of similar length to the original.
```

User prompt: paraphrases, token-level, style, and semantic edits

```
Original caption: {cap}
Produce this JSON object exactly: {"para": [{n} paraphrases: SAME
meaning, different wording, each clearly distinct from the others],
"token": [{n} variants: change only 1-2 words, meaning essentially
preserved], "style": [{n} variants: same content, add a
style/medium/lighting modifier], "sem": [{n} variants: CHANGE one
salient object or attribute so the described scene differs]}
```

User prompt: combined edits

```
Original caption: {cap}
Produce this JSON object exactly: {"H_both": [5 variants: HARD
NEGATIVES that describe a clearly DIFFERENT scene. Each variant must
(a) change THREE or more content words (objects, attributes, actions,
places) so the scene is different, AND (b) lightly paraphrase the rest
so that about half of the remaining words differ from the original
wording. Keep the same sentence style and similar length; the five
variants must be distinct from each other]}
```

A.2 Reference pools and evaluation settings

Table 8 lists the number of references per family. Each reference slot contains one caption generated or sampled for each queried image using the same construction procedure. The main SD v1.4 evaluations use all eligible references; SD v2.1 and SD-XL use the reduced sets in Table 9, and LAION-MI uses random captions and the empty prompt. For each reference, residuals, prediction shifts, and loss gaps are computed from the same predictions and target noise. Separately evaluated caption families can use different noise draws. In that case, averaging across families also averages over more noise realizations.

Table 8: References per image in the expanded pools, before zero-shift omissions.

Reference family	Flickr30k	COCO	Pokemon	LAION-MI
Standard language-model rewrites				
Token-level edits	6	8	14	6
Paraphrases	8	12	18	8
Style changes	2	4	4	2
Semantic edits	8	10	19	8
Combined edits	5	5	5	5
Additional references				
Random captions	12	6	6	6
Empty prompt	1	1	1	1
Total	42	46	67	36

Zero-shift handling. Let D be the number of elements in a noise prediction. We omit a state-reference pair when $\|\delta_{ij}\|_2^2/D \leq 10^{-12}$ and average over all remaining pairs, as in Equation 10. An image with no valid pairs receives score zero. The reference counts in Table 8 are specified before this numerical check.

Configurations for additional analyses. The scoring rules in Tables 5 and 13 are compared using the default reference sets and diffusion states in Section 4.1. Each score averages its per-reference, per-state responses.

Table 3 compares LoFIT and $\mathrm{CLiD}_{\mathrm{th}}$ using either the complete pools in Table 8 or the empty reference alone. Both scores use identical predictions over 48 states on Flickr30k and LAION-MI and 16 on COCO and Pokemon.

To track membership discrimination during fine-tuning, we evaluate successive model checkpoints using randomly sampled captions and the empty prompt as references for each dataset. Analyses of caption edits and Gaussian perturbations use the caption sets and query budgets specified for each experiment. All analyses use the same alignment score.

SD v2.1 and SD-XL reference sets. Table 9 specifies the reduced reference sets used for the SD v2.1 and SD-XL results in Table 1. Query counts use $TK(1 + |\mathcal{R}|)$, reusing the original prediction across references at each state.

Table 9: SD v2.1 and SD-XL reference counts and evaluation budgets.

Configuration	Flickr30k	COCO	Pokemon
Combined edits	5	5	5
Semantic edits	2	2	5
Heavy paraphrases	2	2	5
Empty prompt	1	1	1
Total references	10	10	16
Timesteps T	12	4	4
Noise draws K	4	4	4
Denoiser evaluations	528	176	272

A.3 BASELINE IMPLEMENTATIONS AND CALIBRATION

This section specifies the baseline scoring procedures and the data used for calibration and classifier training.

Naive loss, PIA, and SecMI. Each attack uses its default fixed-timestep setting and one noise realization per image, with seed 0. Predictions use the supplied caption without classifier-free guidance. Each method returns an image-level L2-distance score without classifier training or post-hoc calibration. Images and VAE latents follow the preprocessing above.

CLiD. We use the released character-clipping variant of CLiD (Zhai et al., 2024). Each image is evaluated under its original caption, three character-based caption thirds, and the empty prompt. One Gaussian noise draw per timestep is shared across conditions, giving 15 denoiser queries per image. Evaluation uses sampled VAE latents and FP32 model weights. The features are the original-caption denoising MSE and four conditional loss gaps, each averaged over the three states.

For CLiD_{th}, we apply `RobustScaler` to the original loss and negative mean gap before combining them with weights $1-\alpha$ and α. Calibration follows the experiment-specific protocols below.

For the supplied-caption results on SD v1.4 and LAION-MI (Tables 1 and 2), we select oracle $\alpha \in \{0, 0.1, \ldots, 1\}$ to maximize AUROC using evaluation-pool membership labels. These entries give an oracle AUROC upper bound over the tested grid, rather than an independently calibrated attack. TPR is evaluated at the same selected α. The selected values are 0.9, 0.9, 0.7, and 0.6 for Flickr30k, COCO, Pokemon, and LAION-MI, respectively. SD v2.1 uses fixed $\alpha = 0.5$ with scaling fitted on the unlabeled evaluation pool, and SD-XL uses cross-fitting (Table 10).

CLiD_{vec} uses the five features as inputs to the released XGBoost classifier. SD v1.4 Flickr30k uses shadow-model training; the remaining SD v1.4, SD v2.1, SD-XL, and LAION-MI settings use stratified two-fold cross-fitting on target-model features. We split member and non-member images separately, train on one fold, evaluate the other, and exchange the folds. We repeat this procedure with seeds 0–4 and average the fold metrics across repetitions. For LAION-MI, we compute AUROC from pooled out-of-fold predictions. Cross-fitting uses labeled target examples in the training fold, with each scored image excluded from that fold. The classifier uses `n_estimators=20`, `max_depth=3`, `gamma=0.7`, `subsample=0.7`, `colsample_bytree=0.7`, `reg_alpha=1`, and `reg_lambda=1`.

Table 10: CLiD calibration and classifier configurations for Tables 1 and 2.

Target setting	CLiD_{th}	CLiD_{vec}
SD v1.4 Flickr30k	Oracle $\alpha = 0.9$	Shadow-trained classifier
SD v1.4 COCO, Pokemon	Oracle $\alpha = 0.9$ and 0.7	Target two-fold cross-fitting
SD v2.1, all datasets	Fixed $\alpha = 0.5$; unlabeled target-pool scaling	Target two-fold cross-fitting
SD-XL, all datasets	Target cross-fitting for scaler and α	Target two-fold cross-fitting
SD v1.5 LAION-MI	Oracle $\alpha = 0.6$	Target two-fold cross-fitting

CLiD_{th} on LoFIT's references and states. Table 3 isolates the scoring rule by computing both methods from identical predictions. For each dataset and reference configuration, CLiD_{th} fits `RobustScaler` on the pooled evaluation examples and combines the scaled original loss and negative mean gap as above. We report fixed $\alpha = 0.5$ and an oracle α selected to maximize evaluation AUROC. Each oracle row reports AUROC and TPR at that same α. In Flickr30k/COCO/Pokemon/LAION-MI order, the oracle values are $1.0/0.9/0.8/0.7$ for all reference families and $0.9/0.9/0.7/0.7$ for the empty reference alone.

CLiD with BLIP-generated captions. In Table 17, CLiD_{th} uses the released caption reductions and diffusion states with the BLIP-2 caption as the original condition. The same oracle protocol selects $\alpha = 0.8$, 0.8, 0.2, and 0.4 for Flickr30k, COCO, Pokemon, and LAION-MI, respectively. We report TPR at the same AUROC-maximizing α.

CLiD calibration in the query-budget comparison. Figure 2 uses two-fold cross-fitting to calibrate $\mathrm{CLiD}_{\mathrm{th}}$ on each dataset. Within each training fold, we fit the scaler and select α from $\{0, 0.1, \ldots, 1\}$ by AUROC, then score the held-out fold. We repeat the stratified two-fold split twice using a random generator initialized with seed 0 and average the held-out metrics. This procedure is separate from the main-table calibration in Table 10.

SD-MIA. We use the SD-MIA generation, caption-perturbation, and CLIP-scoring pipeline, with classifier-free guidance scale 6 for generation.

PFAMI. We follow the released PFAMI statistical attack (Fu et al., 2024). For text-to-image diffusion, we VAE-encode perturbed images at the target model's resolution and condition on the original caption with guidance scale 1. Fine-tuned SD v1.4 and SD v2.1 targets use their corresponding pretrained backbones for calibration; LAION-MI uses the uncalibrated statistic. Evaluation costs 110 denoiser queries per image, or 220 with calibration.

A.4 LAION-MI MATCHING AND POPULATION CONTROLS

LAION-MI members originate from LAION-2B-en records included in Stable Diffusion training; non-members come from a later crawl (Dubiński et al., 2024). CLIP matching controls their image-text similarity discrepancy. We examine its effect on attack performance and the population differences that remain.

CLIP matching. We test whether CLIP matching removes class differences in image-text similarity. Before matching, mean similarities are 0.2698 ± 0.0389 for members and 0.2411 ± 0.0526 for non-members. Matching partitions their shared range into 20 bins and subsamples each class to the smaller count per bin. Using CLIP similarity as the membership score, AUROC falls from 0.668 to 0.503 (Table 11). This matching analysis retains $6{,}389$ examples per class; the final attack evaluation uses a separate matched population of $6{,}400$ examples per class.

Table 11: AUROC using CLIP image-text similarity as the membership score, before and after matching.

Membership score	Before matching	After matching
CLIP image-text similarity	0.668	0.503

Effect on SD-MIA. With the same $6{,}400$ members, matching reduces SD-MIA AUROC from 0.611 against all $9{,}578$ non-members to 0.571 against the $6{,}400$ retained non-members (Table 12). The retained and excluded non-member groups are themselves distinguishable, showing that population differences can inflate apparent membership discrimination.

Table 12: SD-MIA AUROC before and after CLIP matching, with population controls on LAION-MI.

Comparison	Group 1 size	Group 2 size	AUROC
Members vs. all non-members (before)	6,400	9,578	0.611
Members vs. matched non-members (after)	6,400	6,400	0.571
Members vs. excluded non-members	6,400	3,178	0.691
Matched vs. excluded non-members	6,400	3,178	0.635

B EVIDENCE FOR THE DIRECTIONAL MEMBERSHIP SIGNAL

These analyses examine the position of the original prediction relative to denoising minima and identify which components of conditional predictions carry the membership signal. The loss-landscape and prediction-shift analyses use captions with varying degrees of wording and semantic modification, together with random captions and the empty prompt (Appendix A.2).

B.1 Loss minima along prediction shifts

At fixed target checkpoints, we test whether members' original predictions lie closer to a denoising minimum along prediction shifts. Figures 5 and 6 show the loss curves and minimum locations.

This visualization uses a separate caption-edit experiment with 4 timesteps and 4 noise draws (16 states) per image, rather than the 48-state Flickr30k configuration in the main evaluation. For each image-reference pair, $\bar{X}_j$ and $\bar{S}_j$ are the state averages of $\langle r_i, \delta_{ij} \rangle / D$ and $\|\delta_{ij}\|_2^2 / D$, respectively, where D is the prediction dimension. With the step expressed in root-mean-square prediction units, the mean loss change is

$$\overline{L}_j(s) - \overline{L}_j(0) = 2s\widetilde{v}_j + s^2, \qquad \widetilde{v}_j = \frac{\bar{X}_j}{\sqrt{\bar{S}_j}}, \qquad s_j^* = -\widetilde{v}_j. \tag{13}$$

Here $s = 0$ is the original prediction, the reference lies at $s = \sqrt{\bar{S}_j}$, and s_j^* is the unconstrained minimum of the averaged curve. The curve coefficient $\widetilde{v}_j$ normalizes after state averaging; LoFIT instead normalizes each v_{ij} before averaging.

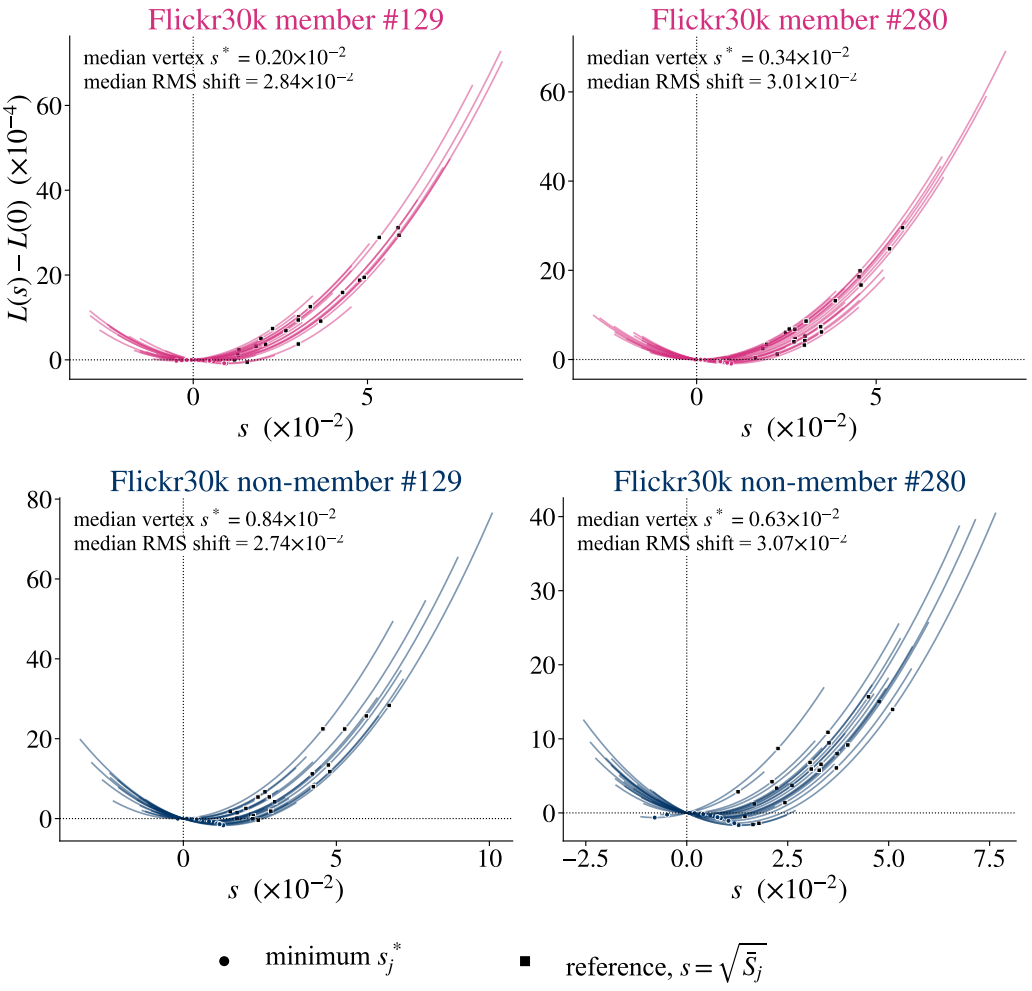


Figure 5: Loss landscapes for two Flickr30k members (top) and two non-members (bottom) under alternative reference captions. Each curve uses the state-averaged statistics for one reference. The original prediction is at $s = 0$, and the vertical axis gives the mean loss change. Dots mark minima and squares reference predictions; curves include extrapolation. The RMS shift is $\sqrt{\bar{S}_j}$ (Equation 13).

Figure 5 illustrates the contrast on four Flickr30k examples: members' minima lie near the original prediction, while non-members' minima lie farther toward the references. Figure 6 shows this

pattern across all four datasets. Median s_j^* values for members/non-members are $0.23/0.82$ on Flickr30k, $0.16/0.88$ on COCO, and $0.14/0.62$ on Pokemon, all in units of 10^{-2}. The difference is smaller on LAION-MI ($0.34/0.47$). Thus, the fine-tuned models show a consistent geometric signature: members' original predictions are closer to the minima along these reference-defined lines.

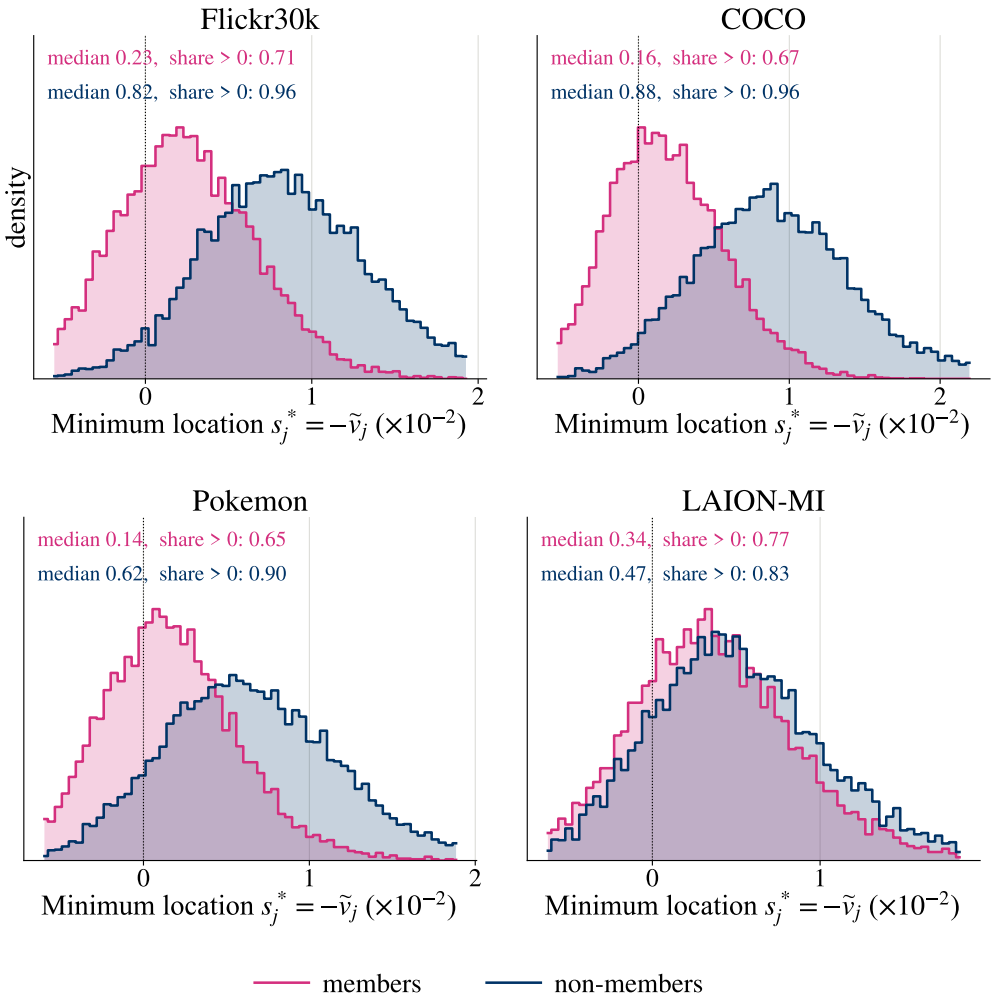


Figure 6: Distribution of minimum locations $s_j^* = -\tilde{v}_j$ from the state-averaged loss curves. Each observation is an image-reference pair. Members' minima (magenta) concentrate closer to the original prediction at $s = 0$ than non-members' minima (blue), with weaker separation on LAION-MI. Positive values indicate that the minimum lies toward the reference prediction.

B.2 Membership information in the loss-gap components

We extend the loss-gap decomposition in Section 5.3 across all four datasets to examine which component carries membership information. For each dataset, Figure 7 shows squared shift magnitudes in the upper panel and the LoFIT response $\langle r_i, \delta_{ij} \rangle / \|\delta_{ij}\|_2$ in the lower panel. Shift magnitudes largely overlap between the classes, while LoFIT responses are more negative for non-members. Separation is pronounced on the fine-tuned models and weaker on LAION-MI. Table 5 tests these components on the full evaluation sets under the default reference configuration: mean shift magnitude remains near chance, while alignment provides stronger discrimination than the loss gap.

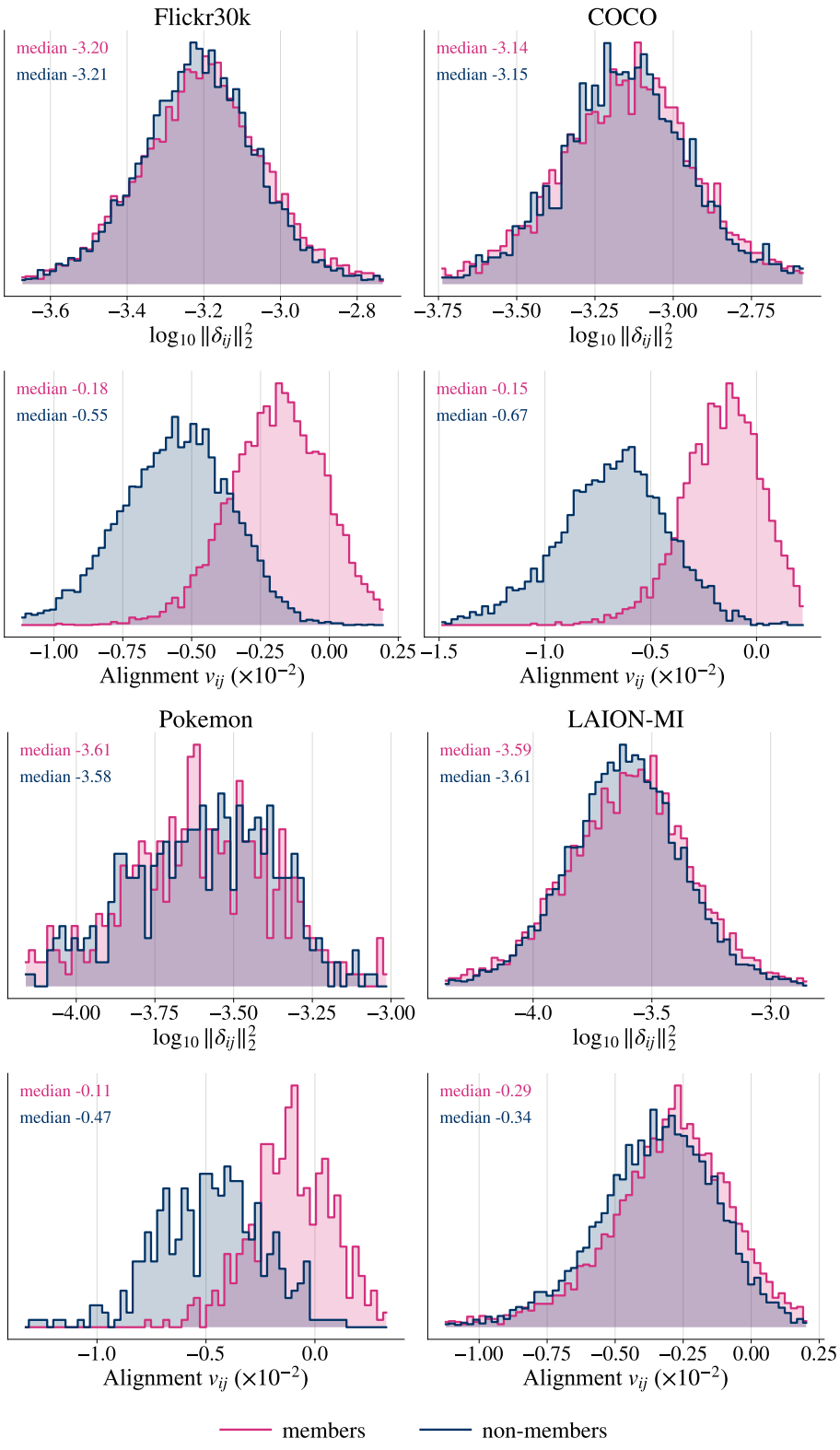


Figure 7: Conditional prediction statistics across four datasets. For each dataset, the upper panel shows $\log_{10} \|\delta_{ij}\|_2^2$ and the lower panel shows normalized alignment $\langle r_i, \delta_{ij} \rangle / \|\delta_{ij}\|_2$, the LoFIT response. Magenta and blue denote members and non-members, respectively.

B.3 Preserving the response sign improves discrimination

We test whether retaining the sign of the alignment response improves membership discrimination. A negative v_{ij} indicates that movement toward the reference prediction initially decreases denoising loss; a positive value indicates an increase. Taking the absolute value or square removes this distinction.

Table 13 compares these scoring rules on identical predictions using the settings in Appendix A.2. Here $\mathbb{E}$ denotes averaging over references and diffusion states. The signed mean improves both AUROC and low-FPR TPR over the absolute and squared responses on all three fine-tuned models, and improves AUROC on LAION-MI. Preserving the direction of the initial loss change therefore strengthens the membership signal. We also test cosine normalization, which additionally divides each response by the residual norm. It lowers AUROC on the fine-tuned models and ties LoFIT on LAION-MI.

Table 13: Membership scores with and without the response sign, evaluated on identical predictions.

	Flickr30k		COCO		Pokemon		LAION-MI	
Score	AUROC	TPR@1% FPR	AUROC	TPR@1% FPR	AUROC	TPR@1% FPR	AUROC	TPR@5% FPR
$\mathbb{E}[v]$ (ours)	**0.935**	**0.454**	**0.954**	**0.478**	**0.896**	**0.272**	**0.602**	**0.101**
$-\mathbb{E}[\lvert v\rvert]$	0.893	0.183	0.915	0.246	0.837	0.062	0.595	0.086
$-\mathbb{E}[v^2]$	0.886	0.184	0.910	0.216	0.835	0.087	0.595	0.088
$\mathbb{E}[v/\lVert r\rVert_2]$	0.917	0.377	0.944	0.410	0.891	0.255	0.602	0.101

C Choosing Informative Conditioning Directions

This section examines how reference selection and perturbation design affect LoFIT's membership discrimination. We study which references provide strong and complementary signals, compare conditioning changes with alternative perturbations, and assess how reference selection changes when the original caption is unavailable.

C.1 Reference quality and complementarity

To guide reference selection, we examine how well individual reference captions distinguish members from non-members and whether combining caption families improves discrimination.

Reference quality at equal cost. We evaluate each reference independently at equal query cost to separate reference quality from the benefit of averaging. Table 14 reports the mean and standard deviation of AUROC across reference slots within each family. We use the expanded reference pools in Appendix A.2, the full evaluation sets, and the diffusion-state configurations in Section 4.1, requiring $2TK$ denoiser evaluations per image. The empty prompt contributes a single slot.

Table 14: Single-reference LoFIT AUROC (mean $\pm$ standard deviation across slots) at equal query cost. Bold marks the highest mean per dataset.

Reference family	Flickr30k	COCO	Pokemon	LAION-MI
Token-level edits	0.642 ± 0.022	0.679 ± 0.026	0.655 ± 0.030	0.515 ± 0.008
Paraphrases	0.737 ± 0.027	0.756 ± 0.029	0.701 ± 0.041	0.517 ± 0.009
Style changes	0.650 ± 0.033	0.700 ± 0.024	0.638 ± 0.015	0.504 ± 0.010
Semantic edits	0.743 ± 0.064	0.793 ± 0.058	0.738 ± 0.047	0.529 ± 0.017
Combined edits	0.880 ± 0.002	0.909 ± 0.007	$\mathbf{0.810} \pm 0.014$	0.569 ± 0.006
Random captions	0.836 ± 0.009	0.819 ± 0.011	0.740 ± 0.017	0.574 ± 0.007
Empty prompt	$\mathbf{0.900}$	$\mathbf{0.914}$	0.781	$\mathbf{0.606}$

The empty prompt is strongest on Flickr30k, COCO, and LAION-MI; combined edits lead on Pokemon and among language-model rewrites on every dataset. Semantic edits vary more across refer-

ences; on COCO, their mean single-reference AUROC is 0.793, compared with 0.819 for random captions.

Combined edits also outperform random captions when both use four references. For each family, we randomly sample four references and evaluate AUROC, repeating this procedure 25 times. The mean AUROC is 0.919 versus 0.894 on Flickr30k and 0.940 versus 0.872 on COCO. Averaging also benefits semantic edits: on COCO, their mean AUROC increases from 0.793 with one reference to 0.897 with four.

C.2 Caption-based and Gaussian conditioning probes

We compare the membership information supplied by individual caption and Gaussian references, then examine their combination. We use the full evaluation sets and diffusion-state grids in Table 14. Each reference is scored independently by averaging alignment responses over states. We report the mean and standard deviation of AUROC across references, at a cost of $2TK$ queries per image per reference.

Gaussian probes query $e_0 + hU_j$, where $e_0 \in \mathbb{R}^{77\times768}$ is the original caption's conditioning tensor and U_j is Gaussian noise normalized to unit Frobenius norm. We sample four Gaussian directions per image at each $h \in \{40, 80, 160\}$ and reuse each direction across that image's diffusion states. Table 15 averages the four single-direction AUROCs at each scale and reports their standard deviation where available; caption-family counts appear in Table 8. The denoiser converts each conditioning perturbation into a prediction shift, which is scored using the same alignment formula as a caption reference. These probes require direct access to the conditioning embedding interface, beyond the text-condition queries assumed by LoFIT.

We match the caption and Gaussian collections by image and timestep/noise-draw index. Their original-caption denoising losses agree exactly on all four datasets.

Table 15: Single-reference AUROC for caption and Gaussian probes. Entries show means and standard deviations where available.

Reference	Flickr30k	COCO	Pokemon	LAION-MI
Caption references				
Empty prompt	0.900	0.914	0.781	0.606
Combined edits	0.880 ± 0.002	0.909 ± 0.007	0.810 ± 0.014	0.569
Random captions	0.836 ± 0.009	0.819 ± 0.011	0.740 ± 0.017	0.574
Semantic edits	0.743 ± 0.064	0.793 ± 0.058	0.738 ± 0.047	0.529
Paraphrases	0.737 ± 0.027	0.756 ± 0.029	0.701 ± 0.041	0.517
Style changes	0.650 ± 0.033	0.700 ± 0.024	0.638 ± 0.015	0.504
Token-level edits	0.642 ± 0.022	0.679 ± 0.026	0.655 ± 0.030	0.515
Gaussian conditioning probes				
$h = 40$	0.688 ± 0.002	0.731 ± 0.004	0.646 ± 0.025	0.520
$h = 80$	0.806 ± 0.001	0.863 ± 0.004	0.761 ± 0.014	0.534
$h = 160$	0.854 ± 0.001	0.903 ± 0.003	0.805 ± 0.009	0.534

On the fine-tuned models, a single Gaussian direction at $h = 160$ outperforms a random caption on average and approaches combined edits, particularly on COCO and Pokemon. This ranking differs on pretrained LAION-MI: Gaussian probes yield AUROC 0.520–0.534, compared with 0.574 for random captions and 0.606 for the empty prompt. The comparison therefore supports Gaussian perturbations as competitive probes in the fine-tuning setting, while captions remain more informative on the pretrained model.

Combining captions and Gaussian probes. Adding Gaussian probes to the caption pool improves membership discrimination on all three fine-tuned models (Table 16). We pool alignment responses from the complete caption reference sets in Table 8 and eight additional Gaussian probes, four at $h = 80$ and four at $h = 160$. These combined configurations use more queries than the caption-only configurations. AUROC increases from 0.935 to 0.946 on Flickr30k, 0.954 to 0.966 on COCO, and 0.896 to 0.910 on Pokemon. TPR at 1% FPR rises from 0.454 to 0.501 on Flickr30k

and from 0.478 to 0.565 on COCO. On LAION-MI, combining the full caption pool with Gaussian probes changes AUROC from 0.573 to 0.575.

Table 16: AUROC after adding Gaussian probes to the full caption pool.

References	Flickr30k	COCO	Pokemon	LAION-MI
Caption pool	0.935	0.954	0.896	0.573
Caption + Gaussian	0.946	0.966	0.910	0.575

C.3 Membership inference with generated captions

We evaluate image-membership inference when the original caption is unavailable. A BLIP-2 OPT-6.7b (Li et al., 2023) caption, generated by greedy decoding, replaces c_0 throughout scoring and reference construction. Membership labels indicate whether the image was used for training. All results use the full evaluation populations. LoFIT and the loss-gap controls use the timestep grids in Section 4.1 and four noise draws.

Reference construction. Caption references comprise randomly sampled BLIP-2 captions of other images in the pool, the empty prompt, and language-model rewrites of the generated caption. Gaussian references perturb its conditioning tensor as $e_{\mathrm{B}} + hU_j$, where e_{B} encodes the BLIP-2 caption and U_j is normalized Gaussian noise, as defined in Appendix C.2. We test three directions at each $h \in \{40, 80, 160\}$ and average alignment responses over directions and states; the nine-direction configuration pools all three scales. The loss-gap baseline averages conditional loss differences on the same predictions used by LoFIT. A single-scale Gaussian configuration requires 192 denoiser evaluations per image on Flickr30k and LAION-MI, and 64 on COCO and Pokemon, excluding caption generation.

Official CLiD uses the released evaluation configuration with oracle α, as specified in Appendix A.3.

Table 17: Image-membership inference with BLIP-2 captions. Bold marks the best value per metric.

	Flickr30k		COCO		Pokemon		LAION-MI	
		TPR@1%		TPR@1%		TPR@1%		TPR@5%
Score and references	AUROC	FPR	AUROC	FPR	AUROC	FPR	AUROC	FPR
Official CLiD								
$\mathrm{CLiD}_{\mathrm{th}}$, oracle α	0.603	0.031	0.650	0.044	0.576	0.041	0.535	0.084
Loss-based controls								
Naive denoising loss	0.522	0.014	0.525	0.018	0.554	0.026	0.521	0.071
Mean loss gap, empty	0.627	0.063	0.673	0.059	0.523	0.041	0.536	0.081
Mean loss gap, random + empty	0.587	0.033	0.619	0.039	0.567	0.046	0.527	0.087
LoFIT with caption references								
Empty only	0.680	0.073	0.763	0.103	0.530	0.007	**0.564**	**0.092**
Random + empty	0.641	0.029	0.669	0.027	0.621	0.034	0.546	0.076
Language-model rewrites	0.684	0.070	0.752	0.091	0.605	0.062	0.545	0.077
LoFIT with Gaussian references								
$h = 40$	0.749	0.081	0.800	0.106	0.669	0.046	0.535	0.063
$h = 80$	0.838	0.190	0.894	0.236	0.782	0.130	0.544	0.066
$h = 160$	**0.842**	**0.214**	**0.904**	**0.264**	**0.802**	**0.139**	0.532	0.068
All nine directions	0.837	0.192	0.896	0.252	0.777	0.111	0.540	0.066
Mean loss gap, same nine	0.618	0.038	0.648	0.057	0.600	0.036	0.494	0.054
LoFIT with mixed references								
Random + empty + Gaussian	0.755	0.096	0.849	0.193	0.761	0.072	0.552	0.081
All captions + Gaussian	0.749	0.117	0.830	0.186	0.717	0.087	0.554	0.078

Alignment remains informative with generated captions. With random captions and the empty reference, LoFIT exceeds official $\mathrm{CLiD}_{\mathrm{th}}$ with oracle α in AUROC across all four datasets (Table 17). On identical predictions, it also outperforms the loss-gap baseline in AUROC with the empty reference alone, random captions plus empty, and the nine Gaussian directions.

With BLIP-generated captions, Gaussian probes outperform caption references on the fine-tuned models. At $h = 160$, AUROC reaches 0.842, 0.904, and 0.802 on Flickr30k, COCO, and Pokemon, respectively. Using the supplied captions yields higher AUROC (Table 1). Within the BLIP setting, combining caption references with Gaussian probes lowers AUROC relative to using the nine Gaussian directions alone on these datasets. On LAION-MI, the empty reference yields AUROC 0.564, above the caption mixtures and Gaussian probes in Table 17. These results show that the most informative references depend on the available query caption.