6. Appendix

6.1. Preliminary

Diffusion-based image generation models [23, 28, 32] aim to generate an image $x \in \mathbb{R}^{H\times W\times 3}$ conditioned on a text prompt c, by learning the joint distribution $p(x, c)$ or conditional distribution $p(x|c)$ via a denoising diffusion process in either pixel or latent space. Formally, the forward diffusion process is defined as:

$$q(z_t|z_{t-1}) = \mathcal{N}\Big(z_t; \sqrt{1-\beta_t}z_{t-1}, \beta_t\mathbf{I}\Big), \tag{8}$$

where $\{\beta_t\}_{t=1}^T$ denotes the noise schedule, and z_0 corresponds to the image x or its latent representation $\mathcal{E}(x)$ encoded by a variational autoencoder. The reverse process is parameterized by a neural network $\epsilon_\theta(z_t, c, t)$ predicting the noise conditioned on the text embedding c:

$$p_\theta(z_{t-1}|z_t, c) = \mathcal{N}\left(z_{t-1}; \frac{1}{\sqrt{1-\beta_t}}(z_t - \beta_t\epsilon_\theta(z_t, c, t)), \tilde{\beta}_t\mathbf{I}\right). \tag{9}$$

The model is trained on large-scale text–image pairs by minimizing the denoising score matching loss:

$$\mathcal{L}(\theta) = \mathbb{E}_{z_0,c,t,\epsilon}\left[\|\epsilon - \epsilon_\theta(z_t, c, t)\|_2^2\right], \tag{10}$$

where (z_0, c) are sampled from the training distribution $\mathcal{D}_{\text{train}}$ of text-image pairs. Through this optimization, the model learns cross-modal correspondences between textual semantics and visual representations. Such tightly coupled associations can lead to memorization of specific training pairs, which can subsequently be exploited for MIAs.

6.2. Baseline Methods

Membership inference attacks on image-generation models have recently attracted increasing attention, aiming to determine whether a given sample was involved in model training. Depending on the adversary's level of access to the target model, existing MIAs can be broadly categorized into white-box, gray-box, and black-box paradigms (Table 2). Representative approaches under these settings include the following:

- Loss [19]: A loss-based attack that exploits the observation that diffusion models yield lower denoising or reconstruction losses for members than for non-members.
- SecMIA [3]: A query-based approach that infers membership by evaluating how well the model's step-wise denoising matches the forward process posterior estimation compared to non-members.
- PIA [14]: A proximal-initialization-based approach that initializes the diffusion process with optimized noise and traces the denoising trajectory of candidate samples.
- CLiD [31]: A conditional-likelihood-based method that measures the discrepancy between conditional and marginal likelihoods in text-to-image diffusion models, capturing overfitting of conditional distributions as a strong membership signal.
- DRC [7]: A degrade–restore–compare framework that deliberately corrupts salient regions of an input image, restores them using the target diffusion model, and compares the reconstruction quality to infer membership.
- REDIFFUSE [16]: A black-box MIA that uses the diffusion model's image-to-image or variation API. It repeatedly inputs a candidate sample for slight modification, averages the outputs, and compares the reconstruction consistency to the original image to infer membership.
- Reconstruction [20]: A black-box reconstruction-based attack leveraging image-to-image or variation APIs to repeatedly regenerate candidate samples and measure reconstruction consistency.

Table 2. Details of MIA methods against diffusion models, where AT, PD, and CM denote the attack type, pre-training data-based evaluation, and use of cross-modal cues for detection, respectively.

Method	AT	CM	PD
Loss (Matsumoto et al., 2023) [19]	White	×	×
SecMI (Duan et al., 2023) [3]	Gray	×	×
PIA (Kong et al., 2024) [14]	Gray	×	✓
CLiD (Zhai et al., 2024) [31]	Gray	×	✓
DRC (Fu et al., 2025) [7]	Gray	×	✓
REDIFFUSE (Li et al., 2025) [16]	Black	×	×
Reconstruction (Pang and Wang, 2025) [20]	Black	×	×
SD-MIA	Black	✓	✓

6.3. Experimental Environments

All experiments were conducted on a high-performance computing server running Ubuntu 22.04.5 LTS. The hardware configuration includes 128 Intel(R) Xeon(R) Gold 6342 CPUs @ 2.80GHz and 2 NVIDIA A800 80GB PCIe GPUs, providing substantial computational capacity for training and inference. The implementation of all methods was carried out using Python 3.12.9 and PyTorch 2.6.0, which served as the primary framework for model development and experimentation.

6.4. Evidence of Representation Collapse

We provide gray-box evidence of representation collapse and analyze how SD-MIA probes it. First, we directly perturb embeddings of members and non-members and measure the reconstruction success rate (RSR). Figure 8 shows: under small perturbations ($|\delta| < 2.5$), RSR of members remains close to the unperturbed case, while that of non-members drops sharply, indicating that members lie in a locally collapsed region. Second, under black-box access, SD-MIA perturbs textual prompts to indirectly induce embedding variations. Results (on SD v1-3, v1-4, and v1-5)

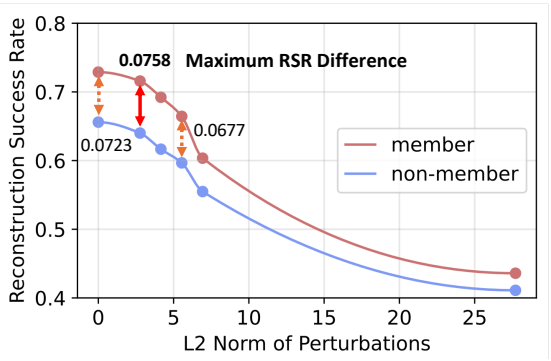

Figure 8. Gray-box evidence of representation collapse.

show that the induced perturbations have an average norm of 0.3858 and a cosine similarity of 0.0066, falling within the validated collapse region above, confirming the effectiveness of the proposed SD-MIA method.

6.5. Efficiency-Utility Trade-off

We evaluate the computational efficiency of SD-MIA alongside its membership inference performance to quantify the trade-off between runtime and utility. Figure 9 presents the membership inference attack performance of SD-MIA and the Reconstruction under varying numbers of repeated generations. The results demonstrate that, with various computational budgets, SD-MIA consistently outperforms the baseline methods, achieving better AUC. This highlights the practicality and efficiency of SD-MIA for large-scale auditing of diffusion models, where maintaining a balance between computational cost and inference accuracy is crucial.

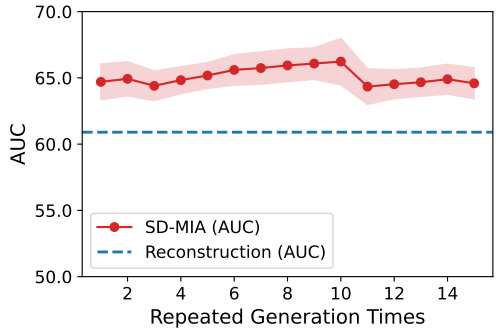

Figure 9. Efficiency-utility trade-off of SD-MIA compared to the Reconstruction baseline, evaluated across different numbers of repeated generations.

6.6. Performance on Image Autoregressive Models

Since SD-MIA is a black-box method that relies only on generated images and avoids diffusion-specific features, it readily transfers to image autoregressive models (IARs), whereas many gray-box methods cannot. We use Infinity[1] as the target IAR model in our experiments. Table 3 shows that SD-MIA effectively identifies training data of IAR models and outperforms black-box baselines.

Table 3. Performance on a representative IAR model (Infinity).

Method	AUC	TPR@1%	TPR@5%	TPR@10%
Reconstruction	$55.55_{\pm 2.10}$	$3.20_{\pm 0.81}$	$7.80_{\pm 2.60}$	$14.00_{\pm 4.04}$
SD-MIA	$\mathbf{58.68}_{\pm 1.94}$	$\mathbf{6.20}_{\pm 3.42}$	$\mathbf{15.20}_{\pm 2.41}$	$\mathbf{22.80}_{\pm 1.54}$

6.7. Experimental Settings for Benchmark Unbiasedness Visualization

To validate the unbiasedness of our benchmarks, we first visualize their embedding-level distribution. Specifically, we encode all samples using CLIP ViT-L/14 [22], then apply Principal Component Analysis (PCA) to reduce the embedding dimensionality to two and plot the resulting distributions, as shown in Figure 3. The visualization indicates that the embeddings of member and non-member samples exhibit well-aligned distributions, suggesting no observable bias. To further quantify this absence of bias, we train a logistic regression classifier to distinguish between member and non-member embeddings. The classifier uses a regularization strength of 1000 to mitigate overfitting and allows up to 1000 iterations to ensure convergence under high-dimensional, strongly regularized conditions. The resulting classification accuracies are nearly equivalent to random guessing (0.518 for FlickrMIA-25 and 0.533 for LAION-mi), which confirms that our benchmarks are effectively unbiased.

6.8. Prompts for Perturbation Generation

In this section, we present the prompts used to generate multi-view textual perturbations within the SD-MIA framework. These prompts are designed to guide the language model in producing variations at three levels: token-level, style-level, and semantic-level. While introducing these perturbations, the prompts ensure that each part of the original description is preserved, allowing for controlled modifications in both the lexical and stylistic elements of the text.

[1] https://huggingface.co/FoundationVision/Infinity

Prompt 1: token-view perturbation

Rewrite the given image caption by rephrasing the text while preserving both the original content/subject and the artistic style exactly. Do not change the main subject or any style modifiers (e.g., 'photorealistic', 'oil painting', 'cartoon style'). Only modify wording, word order, or small descriptive phrasing.
Examples:
- 'photorealistic, a cat on a chair' → 'photorealistic, a cat sitting on a chair'
- 'oil painting of mountains at sunset' → 'oil painting of mountain peaks at sunset'
- 'cartoon style, child playing' → 'cartoon style, a child at play'
- 'digital art, futuristic cityscape' → 'digital art, a futuristic city skyline'
Rules: 1) Preserve the exact subject/content and any style modifiers. Do not introduce new subjects or styles. 2) Only rephrase or slightly rearrange words; avoid adding new objects or changing factual content. 3) Output only the new caption, no quotes or extra text. 4) Ensure the output remains truthful and consistent with the original caption.

Prompt 2: style-view perturbation

Rewrite the given image caption so that the content/-subject remains exactly the same, but change the artistic style of the image. Add only 1-2 style modifiers like 'photorealistic', 'cinematic', 'oil painting', 'cartoon style', etc. before, after, or within the caption.
Examples:
- 'a cat on a chair' → 'photorealistic, a cat on a chair'
- 'UK Active logo' → 'UK Active logo, in the style of oil painting'
- 'person smiling' → 'a watercolor painting of person smiling'
- 'sunset over mountains' → 'cinematic, sunset over mountains'
- 'Salad with chestnuts' → 'Salad with chestnuts, digital art'
Common style modifiers (choose 1-2 only):
- photorealistic, cinematic, highly detailed, 4k
- oil painting, watercolor painting, acrylic painting
- pencil sketch, ink drawing, charcoal drawing
- cartoon style, anime style, manga
- digital art, 3D render, vector art
- in the style of [artist/movement]
Rules: 1) Keep the exact same content/subject. 2) Add only 1-2 style modifiers (not more). 3) Output only the new caption, no quotes or extra text. 4) Ensure that the output caption conforms to objective facts.

Prompt 3: semantic-view perturbation

Rewrite the given image caption so that the content/-subject is changed, but keep the same artistic STYLE. Keep the same style modifiers (if any) but change the main subject/content.
Examples:
- 'photorealistic, a cat on a chair' → 'photorealistic, a dog on a sofa'
- 'UK Active logo' → 'Nike logo' (both simple descriptions without style modifiers)
- 'oil painting of mountains' → 'oil painting of an ocean'
- 'sunset over mountains, digital art' → 'sunrise over cityscape, digital art'
- 'person smiling' → 'person running' (both simple, no style modifiers)
Rules: 1) Change the subject/content to something different. 2) Keep the same style modifiers if present in the original. 3) If no style modifiers in original, keep the same simple format. 4) Output only the new caption, no quotes or extra text. 5) Ensure that the output caption conforms to objective facts.