Supplementary Materials for Unveiling Structural Memorization: Structural Membership Inference Attack for Text-to-Image Diffusion Models

Anonymous Authors

A DESCRIPTIONS OF IMPLEMENTATION DETAILS

A.1 Mathematical Details of DDIM and DDIM Inversion

Denoising Diffusion Implicit Models (DDIM). The denoising process of diffusion models can be refactored as non-Markov, where a skip-step sampling strategy [4] can be applied to accelerate the generation process. Here, the forward diffusion is described by:

$$x_t = \sqrt{\overline{\alpha}_t} x_0 + \sqrt{1 - \overline{\alpha}_t} z \quad z \sim N(0, \mathbf{I}) \tag{1}$$

The denoising process can be represented as:

$$x_{t-1} = \sqrt{\overline{\alpha}_{t-1}} f_\theta(x_t, t) + \sqrt{1 - \overline{\alpha}_{t-1} - \sigma_t^2} \epsilon_\theta(x_t, t) + \sigma_t^2 z \tag{2}$$

where $f_\theta(x_t, t)$ is the prediction of x_0 by the model θ given x_t:

$$f_\theta(x_t, t) = \frac{x_t - \sqrt{1 - \overline{\alpha}_t} \epsilon_\theta(x_t, t)}{\sqrt{\overline{\alpha}_t}} \tag{3}$$

Different samplers are adopted by changing the value of σ_t. Especially, when σ_t is set to 0, the sampling process becomes deterministic, which is DDIM sampling.

DDIM Inversion. For DDIM sampling (σ_t=0), Eq.2 becomes Ordinary Differential Equation (ODE) [6] by rewriting it as follows:

$$\sqrt{\frac{1}{\alpha_{t-1}}} x_{t-1} - \sqrt{\frac{1}{\alpha_t}} x_t = \left(\sqrt{\frac{1}{\alpha_{t-1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1} \right) \epsilon_\theta(x_t, t) \tag{4}$$

Then the forward DDIM process (DDIM inversion) can be considered as a Euler method to solve the ODE, thus can be written as:

$$x_{t+1} = \sqrt{\overline{\alpha}_{t+1}} f_\theta(x_t, t) + \sqrt{1 - \overline{\alpha}_{t+1}} \epsilon_\theta(x_t, t) \tag{5}$$

Such inversion process in Eq.5 provides a deterministic transformation between an input image (x_0) and its latent (x_T).

A.2 Details of Structural Similarity (SSIM)

SSIM [7] primarily considers three characteristics of images: luminance, contrast, and structure.

Luminance (l): SSIM assesses the luminance similarity of two images x and y by comparing their average brightness at the pixel level. If the brightness distributions of two images are similar, the SSIM value for the luminance component will be higher.

$$l(x, y) = \frac{2\mu_x \mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{6}$$

where C_1 is a constant and μ is a measure of average gray levels, which is obtained by averaging the values of all pixels:

$$\mu_x = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{7}$$

Contrast (c): Contrast measures the degree of difference between light and dark regions in an image. If the contrast distributions of two images x and y are comparable, the SSIM value for the contrast component will be higher.

$$c(x, y) = \frac{2\sigma_x \sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{8}$$

where C_2 is a constant and σ is measured by the standard deviation of gray levels:

$$\sigma_x = \left(\frac{1}{N-1} \sum_{i=1}^{N} (x_i - x_\mu)^2\right)^{\frac{1}{2}} \tag{9}$$

Structure (s): Structure focuses on the shapes and edge information of objects within an image. If two images x and y exhibit similar structural properties, such as similar object shapes and edge details, the SSIM value for the structure component will be higher.

$$s(x, y) = \frac{\sigma_{xy} + C_3}{\sigma_x \sigma_y + C_3} \tag{10}$$

where C_3 is a constant and σ_{xy} is defined as:

$$\sigma_{xy} = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mu_x)(y_i - \mu_y) \tag{11}$$

Overall, SSIM is defined as:

$$SSIM(x, y) = l(x, y)^\alpha \cdot c(x, y)^\beta \cdot s(x, y)^\gamma \tag{12}$$

where α, β, γ represent the proportions of different features in the measurement of SSIM, respectively. When α, β, γ are all set to 1, SSIM(x,y) becomes:

$$SSIM(x, y) = \frac{(2\mu_x \mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{13}$$

We follow Eq.13 to calculate the structural similarity between the output image y and the input image x. By comparing the two images' structural similarity, we obtain a membership score for x, which is used for membership inference. The more similar the two image structures are, the higher the score is. A higher score means that the input image x is more likely to be a member of the diffusion model's training set.

A.3 Details of Image Pre-processing

To preprocess images for further experiments, each image is first center-cropped to form the largest possible square using the minimal dimension of the image's height or width. This ensures the images are centered and maintain the most informative parts. After cropping, the image is resized to a resolution of 256x256 or 512x512 pixels, using bi-cubic interpolation which helps in preserving the image quality during resizing. Finally, the pixel values of the image are normalized to range between -1 and 1, a common practice for preparing inputs for neural network models. This standardization process streamlines the input data and is useful when dealing with diverse datasets, facilitating consistent and effective image analysis.

B COMPARISON TO BASELINES

B.1 Calculation Methods for Metrics

Notations. The notations regarding the metrics are demonstrated in Table 1.

Table 1: Notations regarding the metrics.

Notation	Description
TP	True Positives
FP	False Positives
TN	True Negatives
FN	False Negatives
TPR@1%FPR	True Positive Rate at 1% False Positive Rate
TPR@0.1%FPR	True Positive Rate at 0.1% False Positive Rate

Attack Success Rate (ASR). The ASR is the ratio of correctly predicated samples to the total samples, which measures the proportion of successful attacks:

$$ASR = \frac{TP + TN}{TP + TN + FP + FN} \tag{14}$$

The threshold for determining successful attacks is calculated as follows. We randomly select a subset of both members and nonmembers, comprising 20% of the total, and compute an optimal threshold from it. Then we apply this threshold to evaluate the ASR for the remaining population of members and non-members.

Precision. The Precision is the ratio of correctly predicted positives to the total predicted positives, which measures the method's exactness:

$$Precision = \frac{TP}{TP + FP} \tag{15}$$

Recall. The Recall is the ratio of correctly predicted positives to all actual positive samples, which measures the method's completeness:

$$Recall = \frac{TP}{TP + FN} \tag{16}$$

Area Under the ROC Curve (AUC). The AUC is used to evaluate the ability to discriminate between positive and negative samples. It is calculated by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings, and then computing the area under the resulting curve.

TPR@1%FPR, TPR@0.1%FPR. TPR@1% and TPR@0.1% measure the true positive rate when the false positive rate is fixed at 1% and 0.1%, which signify the accuracy in identifying true positives at high confidence.

B.2 Detailed Descriptions of Baselines

PIA [2] utilizes the training loss of diffusion models as a metric for membership inference, specifically by comparing the pixel-level difference between added noise and the predicted noise. The diffusion model's output at time t=0 is used as the added noise. The membership score is computed at timestep 200. A higher score indicates that the probability that the image belongs to the training set is higher:

$$\begin{aligned} score = -||\bar{\epsilon}_0 - \epsilon_\theta(\sqrt{\bar{\alpha}_{200}}x_0 + \sqrt{1-\bar{\alpha}_{200}}\bar{\epsilon}_0, t = 200)||_4^2 \\ \bar{\epsilon}_0 = \epsilon_\theta(x_0, t = 0) \end{aligned} \tag{17}$$

Naive Loss [3] also uses the training loss of diffusion models as a metric for membership inference. What sets it apart from PIA is that it incorporates random Gaussian noise to images and compares it with the predicted noise at the pixel-level. The membership score is computed at timestep 350. A higher score indicates that the image is more likely to be a member of the training set:

$$score = -||\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_{350}}x_0 + \sqrt{1-\bar{\alpha}_{350}}\epsilon, t = 350)||_2^2 \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}) \tag{18}$$

SecMI [1] compares the distance between two adjacent noisy images, which are generated through the diffusion process and the denoising process respectively. Specifically, it first applies DDIM inversion with a fixed interval ($t_i = 1$) to an image x and obtains x_{100} at timestep 100. Then it diffuses x_{100} one step further then reverses one step to get the reconstructed result x_{100}^r. The membership score is defined as the pixel-level distance between x_{100} and x_{100}^r. A higher score indicates that the image is more likely to be a member of the training set:

$$score = -||x_{100}^r - x_{100}||_2^2 \tag{19}$$

C DISTORTED SAMPLES FOR ROBUSTNESS EVALUATION

To evaluate the robustness of our method, various distortions, including additional noise, rotation, saturation change and brightness fluctuation, are applied to both member and nonmember images. Examples are shown in Figure 1.

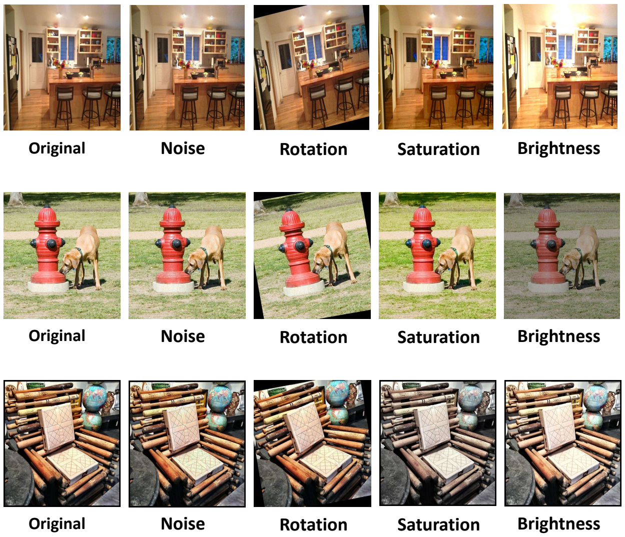
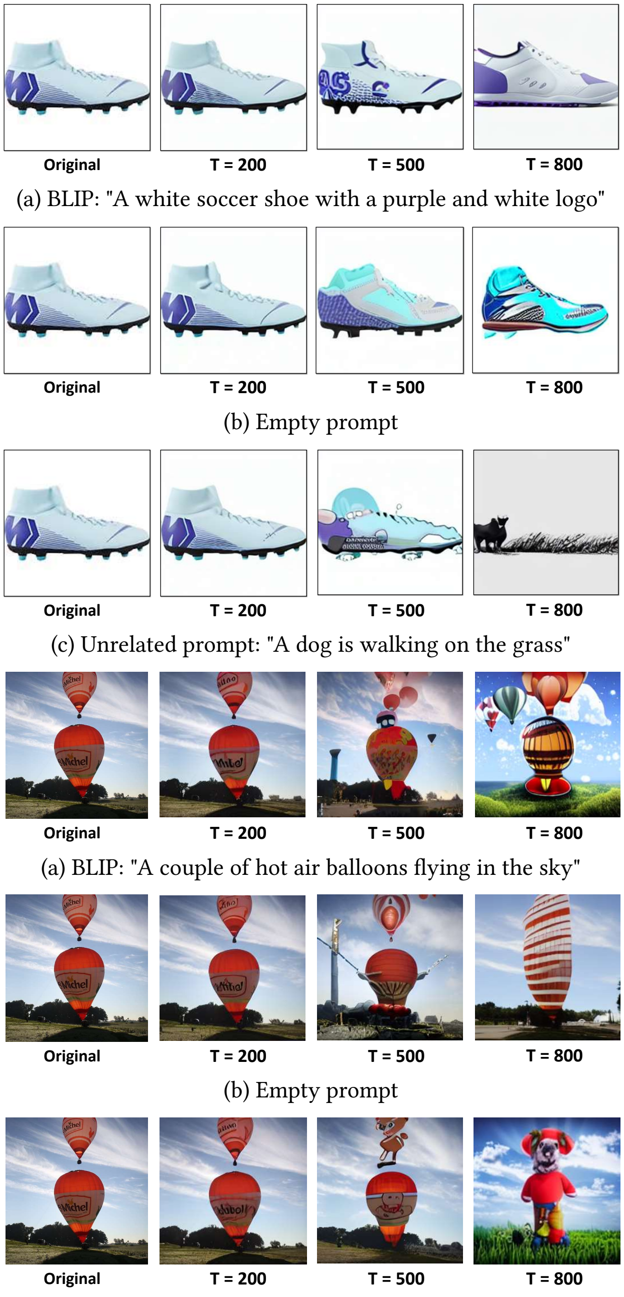


Figure 1: Samples under various distortions: additional noise, rotation, saturation change and brightness fluctuation.



(a) BLIP: "A white soccer shoe with a purple and white logo"

(b) Empty prompt

(c) Unrelated prompt: "A dog is walking on the grass"

(a) BLIP: "A couple of hot air balloons flying in the sky"

(b) Empty prompt

(c) Unrelated prompt: "A dog is walking on the grass"

Figure 2: (a) Reconstruction results with a prompt extracted by BLIP model. (b) Reconstruction results with an empty prompt. (c) Reconstruction results with an unrelated prompt.

D MORE EXAMPLES OF TEXT GUIDED IMAGE RECONSTRUCTION

To evaluate the impact of texts on our method's performance, we first corrupt a image in the diffusion process to a certain timestep T. Then we restore it in the denoising process under three conditions: captions from the BLIP model, empty prompts, and unrelated texts. More results are shown in Figure 2.

E FUTURE WORK

A promising direction for future work involves exploring higher levels of diffusion model's memorization beyond the structural aspects, such as semantic-level. By analyzing the semantic representations learned by the model, we can enhance the understanding and manipulation of deeper, more abstract associations between texts and images. This line of inquiry could lead to the development of more sophisticated attack strategies that target the semantic understanding of the model, potentially revealing vulnerabilities that are less explored in current research. Additionally, exploring the intersection of semantic-level memorization and diffusion model's structure could yield insights into designing more robust defense mechanisms against membership inference attacks. This research not only broadens the applicative scope of text-to-image models but also deepens the theoretical foundations regarding how artificial systems interpret and generate human-like semantic outputs.

F LIMITATIONS

Our method depends on the encoder and decoder components of text-to-image models, as it involves introducing noise into the latent space and extracting image structures from the pixel space to avoid noise interference. Consequently, our approach is not compatible with diffusion models that lack these encoder and decoder elements, i.e. DDPM[5].

G ETHICAL STATEMENT

This study enhances the fairness and inclusivity of AI by identifying biases in text-to-image generative models. We undertake our research with a firm commitment to ethical standards and transparent practices. The aim of our method is to discern whether a specific sample belongs to the training set. This functionality can safeguard privacy rights by spotting unauthorized utilization of personal information for training. Nevertheless, it's necessary to acknowledge that our approach may also pose a privacy threat. For example, privacy could be compromised when anonymous data is categorized by determining its membership in the training set.

REFERENCES

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

[2] Fei Kong, Jinhao Duan, RuiPeng Ma, Hengtao Shen, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. 2023. An Efficient Membership Inference Attack for the Diffusion Model by Proximal Initialization. *arXiv preprint arXiv:2305.18355* (2023).

[3] Tomoya Matsumoto, Takayuki Miura, and Naoto Yanai. 2023. Membership inference attacks against diffusion models. *arXiv preprint arXiv:2302.03262* (2023).

[4] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).

[5] Yang Song and Stefano Ermon. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems* 32 (2019).

[6] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456* (2020).

[7] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.