Towards Memorization-Free Diffusion Models

Chen Chen Daochang Liu Chang Xu
School of Computer Science, Faculty of Engineering, The University of Sydney
{cche0711@uni., daochang.liu@, c.xu@}sydney.edu.au

Abstract

Pretrained diffusion models and their outputs are widely accessible due to their exceptional capacity for synthesizing high-quality images and their open-source nature. The users, however, may face litigation risks owing to the models' tendency to memorize and regurgitate training data during inference. To address this, we introduce Anti-Memorization Guidance (AMG), a novel framework employing three targeted guidance strategies for the main causes of memorization: image and caption duplication, and highly specific user prompts. Consequently, AMG ensures memorization-free outputs while maintaining high image quality and text alignment, leveraging the synergy of its guidance methods, each indispensable in its own right. AMG also features an innovative automatic detection system for potential memorization during each step of inference process, allows selective application of guidance strategies, minimally interfering with the original sampling process to preserve output utility. We applied AMG to pretrained Denoising Diffusion Probabilistic Models (DDPM) and Stable Diffusion across various generation tasks. The results demonstrate that AMG is the first approach to successfully eradicates all instances of memorization with no or marginal impacts on image quality and text-alignment, as evidenced by FID and CLIP scores.

1. Introduction

Diffusion models [12, 23, 34] have attracted substantial interest, given their superiority in terms of diversity, fidelity, scalability [28] and controllability [24] over previous generative models including VAEs [17], normalizing flows [29], and GANs [10, 14–16]. With guidance techniques [7, 11], diffusion models can be further improved by the strategical diversity-fidelity trade-off. State-of-the-art diffusion models trained on vast web-scale datasets are widespreadly used and have seen deployment at a commercial scale [1, 30, 31].

Such widespread adoption, however, has significantly heightened the litigation risks for companies using these models, particularly due to allegations that the models **memorize and reproduce training data during inference** without informing the data owners and the users of diffusion models. This potentially violates copyright laws and introduces ethical dilemmas, further complicated by the fact that the extensive size of training sets impedes detailed human review, leaving the intellectual property rights of the data sources largely undetermined. An ongoing example is that a legal action contends that Stable Diffusion is *a 21st-century collage tool that remixes the copyrighted works of millions of artists whose work was used as training data* [32].

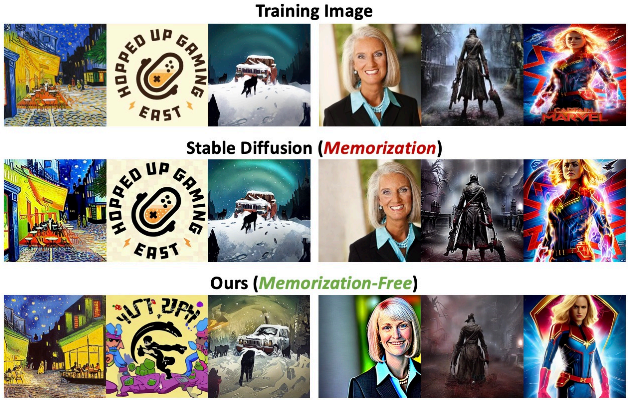

Figure 1. Stable Diffusion's capacity to memorize training data, manifested as pixel-level memorization (left) and object-level memorization (right). Our approach successfully guides pretrained diffusion models to produce memorization-free outputs.

Prior studies [4, 35, 36] have observed memorization in pretrained diffusion models, particularly during unconditional CIFAR-10 [18] and text-conditional LAION dataset [33] generations. While previous research proposed strategies to reduce memorization, these often lead to only modest improvements and fail to fully eliminate the issue. The effectiveness often come with reduced output quality and text-alignment [36], the need for retraining models [4], and extensive manual intervention [19]. Moreover, these strategies lack an automated way to differentiate potential memorization cases for targeted mitigation. For example, [19] relies on a predefined list of text prompts prone to causing memorization, and [36] applies randomization mechanisms uniformly without distinguishing between scenarios.

In this paper, we undertake the following systematic efforts to address the issue of memorization. *Firstly*, we have identified and detailed the primary causes of memorization, pinpointing image and text duplication in training datasets, along with the high specificity of user prompts for text conditioning, as key contributors. *Secondly*, we propose a novel unified framework, Anti-Memorization Guidance (AMG), which comprises three distinct guidance strategies, namely, *desspecification guidance* (G_{spe}), *caption deduplication guidance* (G_{dup}), and *dissimilarity guidance* (G_{sim}), with each meticulously crafted to address one of these identified causes. Each strategy within AMG effectively guides generations away from memorized training images, offering unique benefits. G_{spe} and G_{dup} excel in maximally preserving the quality of generated images, while G_{sim} provides a definitive assurance against memorization. The absence of any one of these strategies would compromise the delicate balance between privacy and utility, underscoring the indispensability of each of the three methods in the framework. To further enhance the privacy-utility trade-off, AMG features an automatic detection mechanism that continuously assesses the similarity between the current prediction and its nearest training data during the inference process to identify potential instances of memorization. This allows AMG to apply guidance selectively rather than uniformly, ensuring that the original sampling process of the pretrained diffusion model is maximally preserved.

We conducted experiments with AMG on pretrained Denoising Diffusion Probabilistic Models (DDPM) and Stable Diffusion, spanning various generation tasks such as unconditional, class-conditional, and text-conditional generations. The outcomes, both qualitative and quantitative, demonstrate that AMG is the first method that effectively eradicates all memorization instances with minimal impact on image quality and text-alignment. In summary, our contributions through AMG are multifaceted and significant: 1) AMG introduces three guidance strategies, each meticulously designed to address one of the primary causes of memorization, providing a comprehensive solution that effectively balances privacy and utility. 2) AMG is equipped with an automatic detection system for potential memorization during each step of the inference process. This allows for the selective application of guidance strategies, maximizing the preservation of output utility. 3) Expanding upon previous research that focused only on unconditional and text-conditional generations, our study is the first to identify and address memorization in class-conditional diffusion model generations, and is the first that successfully achieves memorization-free generations with minimal compromise on image quality and text-alignment.

2. Related Work

Memorization in Diffusion Models has received increased scrutiny over the past year. [35] found that pretrained Stable Diffusions and unconditional DDPMs trained on small datasets like CelebA [22] and Oxford Flowers [25] often replicate training data. [4] reported memorization in pretrained DDPMs on CIFAR-10.Our work focus on pretrained diffusion models, which are extensively used and directly expose their users to litigation risks. We also pioneer in studying class-conditional model memorization, proposing a unified framework that successfully eradicates memorization in various generation tasks including unconditional, class-conditional, and text-conditional generations.

Mitigation Strategies. *Training data deduplication*, initially effective in language models [13, 20], was adapted for diffusion models [4], who removed 5,275 similar images from CIFAR-10 and retrained the model, achieving a reduction in memorization. Yet, it offers limited improvement and requires retraining the entire model, which is computationally intensive, especially for advanced diffusion models with large datasets. *Concept ablation* [19], implemented for Stable Diffusion, involved fine-tuning the pre-trained model on two sets of text-image pairs (one prone to memorization and the other not), curated using ChatGPT-generated paraphrases, to minimize their output disparity. While effective, this method demands extensive manual effort and relies heavily on the crafted prompts' quality. Also, it assumes the availability of a predefined list of memorization-prone prompts, which is unrealistic in many cases. *Randomizing text conditioning* [36] during training or inference can also reduce memorization but has limitations. It mitigates, rather than fully prevents, memorization and lacks guaranteed effectiveness in untested scenarios. It results in significant declines in CLIP [27] score, indicating a poorly balanced trade-off between reducing memorization and maintaining text-alignment of generated images. Furthermore, its uniform application across all text conditions, without considering their potential for causing memorization, further diminishes the images' practical value. Our approach, AMG, adopts a distinct strategy. Instead of altering text conditions to indirectly affect image generation, we aim to directly modify the generated image, thereby providing a guarantee of reduced similarity and addressing the issue of memorization more effectively. Moreover, AMG uses real-time similarity metrics to selectively apply guidance to likely duplicates during inference, ensuring a targeted approach that leaves unaffected cases unaltered, in contrast to the indiscriminate application of *randomization techniques*, also bypassing the need for manual crafting of *concept ablation*.

Other Privacy-Preservation Strategies encompass the adoption of Differential Privacy (DP) [9] in training generative models [5, 8, 40], primarily through the use of the differentially-private stochastic gradient descent (DP-SGD) algorithm [2]. Although effective on smaller datasets like MNIST and CelebA, [4] noted that implementing DP-SGD in diffusion models tends to result in divergence during training on datasets of CIFAR-10 scale. The feasibility of applying DP-SGD to even larger datasets, such as LAION, remain unexplored. In addition, methods such as dataset distillation [21, 39] offer a means to prevent raw data being directly used in the training of generative models, thereby aiding in privacy preservation. Yet, there has been no exploration of these methods on the LAION dataset to date.

3. Preliminaries

3.1. Diffusion Models

Denoising Diffusion Probabilistic Models (DDPMs) [12, 23] consist of two processes, firstly, a forward process is required to gradually add Gaussian noise to an image sampled from a real-data distribution $x_0 \sim q(x)$ over T timesteps such that $x_T \sim \mathcal{N}(0, \mathbf{I})$. The Diffusion Kernel then enables sampling x_t at arbitrary timestep t in a closed form:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}) \tag{1}$$

where x_t is the noised version of x_0 at timestep t, $\alpha_t = 1 - \beta_t$ is the noise schedule controls the amount of noise injected into the data at each step, and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. Then, in the backward process, generative modelling can be realized by learning a denoiser network ϵ_θ to predict the noise ϵ_t instead of the image x_{t-1} at any arbitrary step t:

$$\mathcal{L} = \mathbb{E}_{t \in [1,T], \epsilon \sim \mathcal{N}(0,\mathbf{I})} [\|\epsilon_t - \epsilon_\theta(x_t, t)\|_2^2] \tag{2}$$

where the denoiser network can be easily reformulated as a conditional generative model $\epsilon_\theta(x_t, y, t)$ by incorporating additional class or text conditioning y.

Score-based formulation [38] aims to construct a continuous time diffusion process, where $t \in [0, T]$ is continuous. The reverse processes can be formulated as:

$$dx_t = \left[-\frac{1}{2}\beta(t)x_t - \beta(t)\nabla_{x_t} \log q_t(x_t)\right] dt + \sqrt{\beta(t)} d\bar{w}_t \tag{3}$$

where $\beta(t)$ is a time-dependent function that allows different step sizes $\beta_t = \beta(t)\Delta t$ along the process t. A denoiser network $\nabla_{x_t} \log p_\theta(x_t)$ is then learned to approximate the score function $\nabla_{x_t} \log q_t(x_t)$ in the reverse process using a denoising score matching objective, which can be derived to be the same objective as in Eq. (2), leveraging the connection between diffusion models and score matching:

$$\nabla_{x_t} \log p_\theta(x_t) = -\frac{1}{\sqrt{1-\alpha_t}} \epsilon_\theta(x_t) \tag{4}$$

3.2. Guidance in Diffusion Models

Classifier guidance (CG) and classifier-free guidance (CFG) are methods used in diffusion models to steer image generation towards higher likelihood outcomes as determined by an explicit or implicit classifier $p_\phi(y|x_t)$. In the score-based framework [38], CG and CFG involve learning the gradient of the log probability for the conditional model, $\nabla_{x_t} \log p_\theta(x_t|y)$, rather than the score of unconditional model, $\nabla_{x_t} \log p_\theta(x_t)$. The conditional score can be easily derived using Bayes' rule as the sum of the unconditional score and the gradient of the log classifier probability:

$$\nabla_{x_t} \log p_\theta(x_t|y) = \nabla_{x_t} \log p_\theta(x_t) + \nabla_{x_t} \log p_\phi(y|x_t) \tag{5}$$

Classifier Guidance (CG) [7] involves training an explicit classifier $p_\phi(y|x_t)$ on perturbed images x_t and then employing its gradients $\nabla_{x_t} \log p_\phi(y|x_t)$ to direct the diffusion sampling process towards a class label y. Inserting Eq. (4) into Eq. (5), [7] shows a new epsilon prediction $\hat{\epsilon}$ corresponds to the score presented in Eq. (5):

$$\hat{\epsilon} := \epsilon_\theta(x_t) - \sqrt{1-\bar{\alpha}_t}\nabla_{x_t} \log p_\phi(y|x_t) \tag{6}$$

Classifier-Free Guidance (CFG) [11] eliminates the need of an explicit classifier for computing Eq. (5). It requires concurrent training on conditional and unconditional objectives. At inference, the epsilon prediction is linearly directed towards the conditional prediction and away from the unconditional, and $s_0 > 1$ controls the degree of adjustment:

$$\hat{\epsilon} \leftarrow \epsilon_\theta(x_t) + s_0 \cdot (\epsilon_\theta(x_t, y) - \epsilon_\theta(x_t)). \tag{7}$$

4. Memorization in Diffusion Models

Memorization in generative models is identified when generated images exhibit extreme similarity to certain training samples. The strictest definition of memorization relates to high pixel-level similarity, often qualitatively represented as the generated image being near-copies of training samples as in Fig. 4 and left of Fig. 1. To quantify this similarity, the negative normalized Euclidean L2-norm distance (nL2) is employed as a pixel-level metric [4]. For a generated representation $\hat{x}_0$, this involves first identifying its nearest neighbor n_0 using the ℓ_2 norm, and then normalizing the norm as follows:

$$\sigma_t = -\frac{\ell_2(\hat{x}_0, n_0)}{\alpha \cdot \frac{1}{k}\sum_{z_0 \in S_{\hat{x}_0}} \ell_2(\hat{x}_0, z_0)} \tag{8}$$

where $S_{\hat{x}_0}$ is a set of $k = 50$ nearest neighbors of $\hat{x}_0$, and α is a scaling constant with a default value of 0.5.

A broader definition of memorization encompasses reconstructive memory in diffusion models, where the models reassemble various elements from memorized training images, such as foreground and background objects. These reconstructions might include transformations like shifting, scaling, or cropping. Consequently, the reconstructed outputs do not necessarily match any training image on a pixel-by-pixel basis, yet they exhibit a high degree of similarity to certain training images at the object level. In right section of Fig. 1, we observe that the outputs generated by Stable Diffusion are not pixel-level identical to the training images. However, they demonstrate significant object-level similarity. To quantify such object-level similarity, a commonly

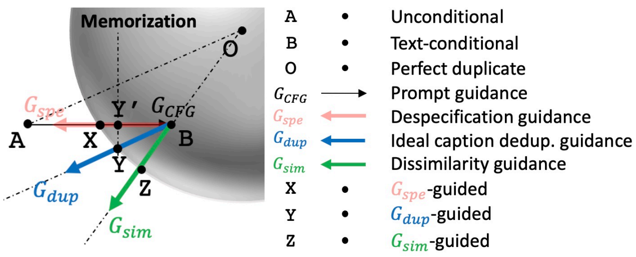

Figure 2. Geometric interpretation of different guidance methods and generations. The center O represents a scenario where the generated image is identical to the memorized training image. The distance of any point from O reflects its degree of dissimilarity to the memorized image. The surface of the sphere signifies the threshold that defines the presence of a memorization issue. The arrows represent different types of guidance strategies.

used metric is the dot product of embeddings of $\hat{x}_0$ and n_0:

$$\sigma_t = E(\hat{x}_0)^T \cdot E(n_0) \tag{9}$$

where $E(\cdot)$ represents the embedding obtained via a feature extractor, with Self-supervised Copy Detection (SSCD) [26] being the preferred method for identifying object-level memorization [35], n_0 represents the nearest neighbor of x_0, identified using this metric.

In later studies [19, 36] that examine memorization issues within the diverse LAION dataset, where numerous object-level memorization instances are identified, SSCD has been established as the standard metric. On the other hand, for the CIFAR-10 dataset, the negative nL2 is found to be an efficient measure [4], attributable to the dataset's smaller size and consistency in image presentation.

4.1. Causes of Memorization

The main causes of memorization in diffusion models are identified as follows: 1) *Overly specific user prompts* act as a "key" to the pretrained model's memory, potentially retrieving a specific training image corresponding to this "key", as observed by [36]. 2) *Duplicated training images* are more inclined to be memorized by diffusion models as noted by [4, 35], likely due to overfitting. 3) *Duplicated captions across those duplicated images* can exacerbate the memorization issue [36] by overfitting the text-image pairs to text-conditional diffusion models, turning the caption into a "key" that consistently retrieves the "value" of the associated image. Therefore, when a user employs such repetitive captions or closely related text prompts as the conditioning, the model is prone to generate the corresponding duplicated training image.

5. Anti-Memorization Guidance

Leveraging insights into the primary causes of memorization, for diffusion models, we present *Anti-Memorization Guidance (AMG)*, a unified framework integrating a comprehensive suite of three distinct guidance strategies, namely, *dissimilarity guidance*, *desspecification guidance*, and *caption deduplication guidance*. Each strategy within AMG is meticulously crafted to address and effectively eliminate specific causes of memorization in these models. As illustrated in Fig. 2, in the original diffusion models employing classifier-free guidance (CFG), the unconditional generation (point A) is linearly guided towards its text-conditional generation (point B), which may falls inside the sphere, indicating that it has become a memorized case. In the AMG framework, all three guidance methods are capable of steering the generation process away from memorization, represented by moving the generation outside the sphere in the geometric representation. From an implementation standpoint, each guidance strategy is readily integrable with different types of pretrained diffusion models, such as conditional and unconditional Denoising Diffusion Probabilistic Models (DDPMs) and Latent Diffusion Models (LDMs), without the necessity for additional re-training or fine-tuning. This compatibility extends to different sampling methods, including DDPM sampler and accelerated sampling method such as DDIM [37]. The framework solely requires updating the epsilon prediction $\hat{\epsilon}$ in accordance with each specific guidance strategy, which is then combined with x_t to estimate the previous step representation x_{t-1} in the reverse process of diffusion models:

$$\hat{\epsilon} \leftarrow \hat{\epsilon} + 1_{\{\sigma_t > \lambda_t\}} \cdot (G_{spe} + G_{dup} + G_{sim}) \tag{10}$$

$$x_{t-1} \leftarrow \sqrt{\bar{\alpha}_{t-1}} \left(\frac{x_t - \sqrt{1-\bar{\alpha}_t}\hat{\epsilon}}{\sqrt{\bar{\alpha}_t}} \right) + \sqrt{1-\bar{\alpha}_{t-1}}\hat{\epsilon} \tag{11}$$

To minimize alterations to the original sampling process and thus preserve output utility to the greatest extent, we introduce an indicator function $1_{\{\sigma_t > \lambda_t\}}$ that activates guidance only when the current similarity σ_t exceeds a pre-set threshold λ_t. Importantly, we have designed λ_t as a dynamic, rather than static, threshold. This dynamic nature accounts for the observed fluctuations of σ_t throughout the inference process. For instance, in the early denoising stages, when t is large, the prediction $\hat{x}_0$ is generally less precise than at later stages when t approaches 0. Consequently, σ_t tends to be lower at higher t values and increases as t decreases. To effectively manage this variation, we adopt a parabolic scheduling for λ_t in alignment with the characteristics of the denoising stages. As a result, this design of conditional guidance with parabolic scheduling operates as an automatic mechanism to selectively activates guidance only when necessary. Such a configuration enables AMG to optimize the balance between reducing memorization and maintaining high-quality outputs.

Moreover, AMG ensures that the generated images, during inference time, diverge from the memorized training image at either pixel or object level. The type and strength of this divergence can be flexibly tailored based on the user's specific application and objectives.

5.1. Despecification Guidance

As previously discussed, a primary cause of memorization in text-conditional diffusion models is the overly specific nature of user prompts, acting as a "key" to the pretrained model's memory [36]. To reduce caption specificity by instructing the inference process, we firstly employ a *desspecification guidance*. Given an noised image or latent-space representation x_t and predicted noise $\hat{\epsilon}$ at time t, we can obtain its prediction of $\hat{x}_0$ using the Diffusion Kernel (Eq. (1)):

$$\hat{x}_0 = \frac{x_t - \sqrt{1 - \bar{\alpha}_t} \cdot \hat{\epsilon}}{\sqrt{\bar{\alpha}_t}} \tag{12}$$

Then, we search its nearest neighbor n_0 in the training set and compute the similarity σ_t between $\hat{x}_0$ and n_0. Depending on the user's goal to prevent pixel-level or object-level memorization, the similarity measure σ_t can be computed accordingly using nL2 (Eq. (8)) or SSCD (Eq. (9)).

This method aligns with the principles of CFG but pursues the inverse goal: to attenuate the original CFG scale to linearly adjust the epsilon prediction to be less aligned with the prompt-conditional prediction:

$$s_1 = \max(\min(c_1\sigma_t, s_0 - 1), 0) \tag{13}$$

$$G_{spe} = -s_1(\epsilon_\theta(x_t, y) - \epsilon_\theta(x_t)) \tag{14}$$

where $\epsilon_\theta(x_t, y)$ represents the pretrained diffusion model's prediction conditioned on user's text prompt, while $\epsilon_\theta(x_t)$ denotes the unconditional prediction. s_0 is the original scale of CFG, c_1 is a constant and $c_1\sigma_t$ defines the guidance scale at step t, which is directly proportional to the similarity σ_t at step t. This enables the algorithm to adaptively adjust the scale of *desspecification guidance* throughout the sampling process, corresponding to the current level of the memorization, as indicated by the value of σ_t at any step t. The function $max(\cdot, 0)$ guarantees the guidance scale $-s_1$ to be non-positive, thus diminishing caption specificity, while $min(\cdot)$ function bounds $c_1\sigma_t$ to not exceed $s_0 - 1$, safeguarding against excessive low text-image alignment.

From a geometric perspective as in Fig. 2, Despecification guidance (G_{spe}) is capable of steering the generation process away from memorization, start from point B, G_{spe} directs the prediction in the exact opposite direction of the CFG, exiting the sphere at point X on the surface.

5.2. Caption Deduplication Guidance

As outlined in Sec. 4, duplicated captions can act as precise "keys" to retrieve memorized data from the training set, so why not turn this to our advantage? By intentionally using them as prompts for pretrained diffusion models to generate predictions that replicates these memorized images, we can then apply classifier-free guidance techniques to steer the generation away from these images:

$$s_2 = \max(\min(c_2\sigma_t, s_0 - s_1 - 1), 0) \tag{15}$$

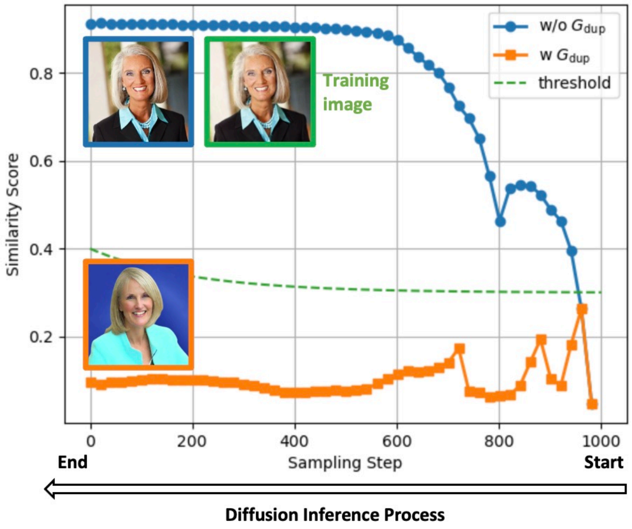

Figure 3. Comparison of similarity scores throughout the inference process, with and without the application of G_{dup}.

$$G_{dup} = -s_2(\epsilon_\theta(x_t, y_N) - \epsilon_\theta(x_t)) \tag{16}$$

where y_N denotes the caption of n_0, which is the nearest neighbor of current prediction $\hat{x}_0$ as defined in Eq. (12). In case where n_0 is a duplicated image prone to memorization and accompanied by a duplicated caption, y_N would correspond to this replicated caption. Consequently, $\epsilon_\theta(x_t, y_N)$ reflects the conditional prediction based on the duplicated caption, serving as an ideal antithesis to the prediction we aim to achieve. The function $max(\cdot, 0)$ again guarantees non-positive guidance scale $-s_2$, directs $\hat{\epsilon}$ away from conditional prediction compared to the unconditional prediction $\epsilon_\theta(x_t)$, while $\min(\cdot)$ bounds the total scale of $s_1 + c_2\sigma_t$ to not exceed $s_0 - 1$ for preserving text-image alignment.

From a geometric standpoint as in Fig. 2, caption deduplication guidance (G_{dup}) runs parallel to line OA, representing the guidance direction when using perfect text-conditioning that leads to memorization as a negative prompt. This method exits the sphere at point Y on the surface, thereby moving the generation out of the memorization zone. Fig. 3 further illustrates the efficiency of G_{dup}. It demonstrates that when the similarity score exceeds the dashed parabolic threshold line λ_t, as defined in Eq. (10), G_{dup} is activated. This activation prompts G_{dup} to guide the generations in a direction opposite to that of the training image, effectively preventing memorization.

5.3. Dissimilarity Guidance

Distinct from the first two strategies, which linearly adjust the epsilon prediction $\hat{\epsilon}$ along the vector direction between conditional and unconditional predictions, *dissimilarity guidance* identifies another dimension that offer effective guidance to reduce, or even eliminate, memorization in diffusion models. It extends the discrete class label y in classifier guidance Eq. (6) to a continuous embedding represented by the similarity score. This approach assures that our generated images are actively directed towards reducing their similarity score, thereby ensuring persistent dissimilarity from their closest counterparts in the training set, as measured by metrics such as nL2 and SSCD.

$$G_{sim} = c_3\sqrt{1-\bar{\alpha}_t} \cdot \nabla_{x_t}\sigma_t \tag{17}$$

where we use the similarity metric σ_t instead of log classifier probability $\log p_\phi(y|x_t)$ to compute gradient and guide the inference process. We also invert the sign preceding the guidance term from Eq. (6) to indicate our new objective: to minimize similarity as opposed to maximizing it. An additional scaling factor c_3 is also employed, which functions as a hyperparameter to control the intensity of the guidance, thereby managing the privacy-utility trade-off and tailoring user's specific goals. A larger c_3 permits greater deviation of the generated data from the nearest training image, but at the expense of reduced quality and text alignment.

Geometrically as in Fig. 2, dissimilarity guidance (G_{sim}) steers the generation direction directly away from point `O`, which represents the perfectly memorized case (*i.e.*, the training data), eventually exit at point `Z` on the sphere.

5.4. Unpacking AMG's Threefold Guidance

We present a detailed analysis of the indispensable role and synergies of each of the three guidance methods in AMG for achieving the optimal balance between privacy and utility.

Impact on quality and text-alignment. Similar to CFG, our G_{spe} method linearly combines the unconditional prediction $\epsilon_\theta(x_t)$ and user-prompt based conditional predictions $\epsilon_\theta(x_t, y)$. Altering the guidance scale modifies the weights of these components, but as both are derived from pretrained diffusion model's high-quality outputs, overall output quality remains largely consistent. However, text alignment decreases with lower weights on $\epsilon_\theta(x_t)$, necessitating a minimum weight of one to preserve text alignment. Similarly, for G_{dup}, assuming using a neighbor's caption y_N leads to an exact replication of training image, the output maintains high quality, thus its linear combination with $\epsilon_\theta(x_t)$ also yields quality on par with the pretrained model. Geometrically, results within the `ABY` plane, formed by lines `AB` and `BY`, maintain quality, but those closer to `A` (along line `AB`) show reduced text alignment. G_{sim}, however, does not align with the `ABY` plane, so its scale affects the quality and must be minimally set to preserve quality.

Importance of dissimilarity guidance (G_{sim}). G_{sim}, as shown in Fig. 2, is vital despite its possible quality impact. The necessity stems from the fact that G_{spe} and G_{dup}'s combined scale, capped at $s_0 - 1$, cannot assure moving the generation outside the sphere, potentially leaving it within the `BXY` sector. G_{sim}, on the other hand, reliably ensures the generation is guided out of the sphere, effectively addressing memorization.

Importance of caption deduplication guidance (G_{dup}). Text-conditioning heightens specificity, thus reducing diversity in generations. In severe cases, such as when point `B` near center `O`, G_{sim} would need a high scale to prevent memorization without G_{spe} and G_{dup}, risking artifacts. G_{spe} and G_{dup} mitigate this by lowering the needed scale for G_{sim}, thus preserving output quality. Comparing G_{spe} and G_{dup}, both maintain quality within the `ABY` plane, but G_{dup} involves less text alignment sacrifice due to the shorter linear projection of `BY` on the `AB` line (*i.e.*, `BY'` $<$ `BX`), thus more beneficial than G_{spe}.

Importance of despecification guidance (G_{spe}). The inclusion of G_{spe}, despite G_{dup}'s apparent advantages, is justified by G_{dup}'s practical limitations as it idealizes deduplication guidance, requiring access to perfect "keys" for pretrained model memories, approximated here by captions of memorized training images. Such ideal "keys" are uncommon, and caption-based approximations may not always be as precise as prompt-based methods in G_{spe}, particularly when memorized images' captions are not duplicated in the dataset. Consequently, our approach in AMG integrates both G_{spe} and G_{dup} to emulate these ideal "keys" for effective negative guidance.

Unconditional and class-conditional generations. Our research reveals that in the absence of text-conditioning, memorization in diffusion models is reduced but not eliminated, aligning with our prior findings. This significantly lessens memorization, leaving image duplication as the only remaining issue. In such cases, G_{spe} is highly effective in eliminating memorization due to two factors: 1) *Early Detection*: During initial stages of reverse diffusion, our conditional guidance with a parabolic schedule efficiently identifies potential replication. 2) *Increased Diversity*: Without specific text-conditioning, the generation process yields greater diversity. This enables G_{dis} to effectively steer generations away from memorized modes to un-memorized ones in the initial stages, ensuring they don't revert. Once memorization ceases to be detected, further guidance application is discontinued. As a result, this method predominantly impacts the coarse structure, with guidance typically applied only during the early stages of the denoising process. The finer details in later stages remain intact, thus ensuring the overall quality of the output is preserved.

6. Experiments

6.1. Experimental Setup

Scope. In the realm of pretrained diffusion models, studies [4, 35, 36] have identified memorization in unconditional DDPMs on CIFAR-10 and text-conditional Stable Diffusion on LAION datasets. While [35] found no memoriza-

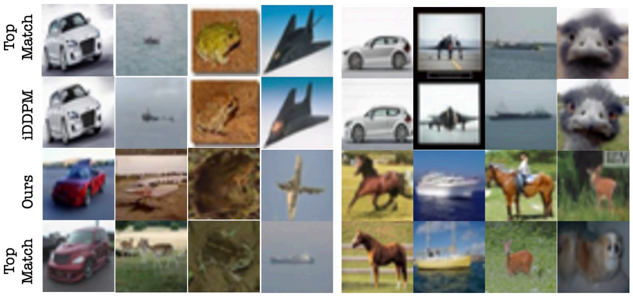

Figure 4. Applying AMG to iDDPM on CIFAR-10. Left: Class-conditional generation. Right: Unconditional generation.

tion in latent diffusion models using ImageNet, our analysis of DDPMs on ImageNet [6] and LSUN Bedroom [41] also showed no memorization cases. Beyond these scenarios, we explored class-conditional generation and identified memorization cases on CIFAR-10. Our findings confirm the successful elimination of memorization in all tested scenarios, highlighting the wide applicability of our approach.

Evaluation metrics. To assess text-conditional generations (Tab. 1), we employ three metric types: memorization, quality, and text-alignment. Memorization metrics include the 95th percentile [36] of similarity scores of all generated images, determined using pixel-level nL2 norm (Eq. (8)) or object-level SSCD embedding similarity (Eq. (9)). We contend that relying solely on this metric might be misleading, especially if the distribution of the generated data exhibits a heavy upper tail beyond the 95th percentile. In such scenarios, the similarity score could be significantly understated. We propose to additionally examine the maximum similarity score within the distribution, to gauge the worst-case scenario regarding memorization. Additionally, the proportion of images exceeding certain similarity thresholds, thus flagged as memorized, is a key metric [4]. Notably, thresholds vary; for CIFAR-10, an nL2 below 1.4 normally indicates pixel-level memorization, while for LAION, an SSCD above 0.5 suggests object-level memorization following the convention of previous work [19, 35, 36]. We use FID to measure the fidelity and diversity of generated images, and CLIP score to measure the generated images' alignment with the input text prompts.

Implementational details. Experimental results depend on variables such as text prompts, number of generated images, training image scope (e.g., LAION-10k to LAION-5B), choice of diffusion model, and sampling steps. Since baseline methods often employ different settings, we have reimplemented these baselines to ensure a fair and comparable evaluation. In our text-conditional generation experiments on the LAION5B dataset, we utilized [3]'s system for replicates identification. They used a CLIP embedding-based index for LAION5B, with an efficient retrieval system for identifying k-nearest neighbors. Our approach differed in seeking the singular nearest neighbor based on SSCD embedding similarity. To approximate this, we first identified 1,000 images with the lowest CLIP embedding similarities, then computed SSCD similarities to find the highest match, using it for memorization metrics. Stable Diffusion v1.4 with a DDIM sampler and 50 sampling steps was our model of choice. For unconditional and class-conditional generations on CIFAR-10, which comprises 50,000 images, we calculated SSCD similarities for each generated image with the entire training set. OpenAI's iDDPM [23] with a DDPM sampler and 250 sampling steps was used. Further details are in the supplementary material.

	Memorization Metrics by SSCD ↓					
	Top5%	Top1	%>0.5	%>0.4	FID ↓	CLIP ↑
SD [30]	0.91	0.93	44.85	59.23	106.41	**28.04**
Ablation [19]	-	-	0.30*	-	-	-
GNI [36]	0.91	0.94	42.75	58.18	97.81	27.79
RT [36]	0.61	0.84	15.07	26.75	101.69	22.63
CWR [36]	0.79	0.85	26.45	40.93	**96.25**	25.96
RNA [36]	0.75	0.82	17.78	29.05	99.68	23.37
Ours(Main)	0.41	0.47	**0.00**	7.07	99.12	26.98
Ours(Strong)	**0.34**	**0.39**	**0.00**	**0.00**	100.45	26.72

Table 1. Comparisons on text-conditional generation of LAION5B based on SSCD similarity. AMG successfully eliminates memorization with minimal impact on quality and text-alignment.

	Memorization Metrics by nL2				
	Top5%↑	Top1↑	%<1.4↓	%<1.6↓	FID↓
iDDPM [23]	1.58	0.51	0.93	5.78	7.44
Ours(Main)	1.61	1.47	**0.00**	4.34	7.25
Ours(Strong)	**1.71**	**1.68**	**0.00**	**0.00**	**6.98**

Table 2. Comparisons on unconditional generation of CIFAR-10 based on nL2 similarity. AMG effectively eliminates memorization without affecting image quality.

	Memorization Metrics by nL2				
	Top5%↑	Top1↑	%<1.4↓	%<1.6↓	FID↓
iDDPM [23]	1.53	0.51	1.53	9.77	11.81
Ours(Main)	1.56	1.46	**0.00**	8.70	11.54
Ours(Strong)	**1.71**	**1.68**	**0.00**	**0.00**	**11.44**

Table 3. Comparisons on class-conditional generation of CIFAR-10 based on nL2 similarity. AMG effectively eliminates memorization without affecting image quality.

6.2. Comparison with Baselines

Text-conditional generations on LAION. Table 1 shows that AMG outperforms all baselines in memorization metrics by a huge margin, eliminating all memorization cases defined by a similarity score over 0.5. It also shows a minimal loss in text-alignment, evidenced by having the second-highest CLIP score among mitigation strategies, thus maintaining strong alignment with user intentions. Notably, the strategy with the highest CLIP score, GNI, performs poorly in memorization metrics, closely resembling the original Stable Diffusion without mitigation. AMG matches baselines in FID, indicating comparable quality. Overall, AMG leads to memorization-free generations and stands out in balancing quality and utility. AMG's flexibility allows users

	Mem. by SSCD ↓			
	Top5%	%>0.5	FID ↓	CLIP ↑
Baseline [30]	0.9133	44.85	106.41	**28.04**
$G_{sim} + G_{spe}$	0.4072	**0.00**	119.13	26.67
$G_{sim} + G_{dup}$	0.4073	**0.00**	120.48	26.17
$G_{spe} + G_{dup}$	0.7396	31.62	**87.10**	27.18
Full	**0.4066**	**0.00**	99.12	26.98

Table 4. Ablation studies on text-conditional generation based on SSCD. Grey-colored font denotes areas of sacrifice.

	Mem. by nL2		
	Top5%↑	%<1.4↓	FID ↓
Baseline [23]	1.58	0.93	7.44
w/o conditional guidance	1.49	**0.00**	257.27
w constant schedule	1.59	0.04	7.44
Full	**1.61**	**0.00**	**7.25**

Table 5. Ablation studies on unconditional generation based on nL2. Grey-colored font denotes areas of sacrifice.

to adjust guidance strength based on their definition of memorization. While the standard threshold is 0.5, increasing AMG's guidance scale (*i.e.*, the strong version of AMG) effectively prevents memorization even at a 0.4 threshold, at a minimal extra cost of quality and text-alignment.

Unconditional and class-conditional generations on CIFAR-10. Tab. 2 and Tab. 3 illustrate AMG's effectiveness in transitioning from text-conditional to class-conditional or unconditional generation tasks, and from preventing object-level to pixel-level memorization. AMG consistently outperforms, even when compared to [4], who reported a 23% reduction in memorization by retraining diffusion models on a deduplicated CIFAR-10 dataset. AMG ensures memorization-free outputs and even slightly exceeds the original diffusion model's quality, as reflected in FID scores, likely due to the increased diversity of its generated images compared to the original model's replicated outputs. This success is attributed to two main factors: 1) AMG's early memorization detection during reverse sampling, utilizing a conditional guidance with a parabolic schedule; 2) The absence of text-conditioning eliminates key memorization causes, like specific user prompts and caption duplication, enhancing output diversity. Solely using *dissimilarity guidance* in AMG can be very effective in preventing memorization whilst preserving output quality, since it only alters the coarse structure of images in early stages of reverse sampling when potential memorization is detected. Guidance ceases once memorization is no longer detected, thus preserving sample quality.

6.3. Ablation Studies

Table 4 underlines the crucial role of AMG's tripartite guidance in optimizing the privacy-utility trade-off. Key observations include: 1) All AMG versions notably enhance memorization metrics, particularly those incorporating G_{sim}, which eradicate memorization entirely with proper guidance scale tuning. This underscores G_{sim}'s theoretical guarantee against memorization, albeit with slight impacts on quality and text-alignment. Ablation of G_{sim} improves FID and CLIP scores but results in 31.62% of generated images being marked as memorized. Thus, G_{sim} inclusion significantly boosts privacy with minimal utility loss. 2) Comparisons between the full AMG version and variants lacking either G_{dup} or G_{spe} demonstrate that while achieving similar privacy levels, the ablated versions yield inferior FID and CLIP scores. This confirms the importance of both G_{dup} and G_{spe} in the guidance ensemble.

Table 5 highlights the efficacy of our conditional guidance with parabolic scheduling. Applying guidance indiscriminately during sampling, rather than selectively based on potential memorization, guarantees elimination of memorization but degrades output quality, evidenced by significantly higher FID scores. This is due to excessive alteration of both coarse structures (early sampling stages) and finer details (later stages). The parabolic schedule aligns with denoising stages: initially, predictions are highly noised and dissimilar to training images, becoming more accurate and revealing potential memorization cases with higher similarity as denoising progresses. This schedule enables early detection and effective resolution of memorization issues. A constant schedule would fail to provide this early detection, leading to 0.04% of generations being memorized, which could be eliminated by increasing the guidance scale but at the cost of quality. Therefore, our conditional guidance strategy enhances the privacy-utility trade-off by negating the need for this additional quality compromise.

7. Conclusion

We introduce *AMG*, a unified framework featuring three specialized guidance strategies, each addressing a specific cause of memorization in diffusion models. Theoretical analysis and empirical ablation studies confirm the essential role of each strategy in achieving an optimal privacy-utility trade-off. AMG's strategic guidance scheduling and innovative automatic detection enable conditional application, further refining this balance. Our experiments demonstrate that AMG reliably generates images during inference that are distinct from memorized training images, maintaining high quality and text-alignment. Furthermore, AMG offers the flexibility to adapt to various user requirements by allowing customization in the type of memorization prevented (pixel-level or object-level) through adjustments in the similarity metrics employed in its guidance. Additionally, it provides options for guidance intensity (main or strong version) by adjusting the guidance scale, catering to a wide range of applications and user preferences.

8. Acknowledgement

This work was supported in part by the Australian Research Council under Projects DP210101859 and FT230100549.

References

[1] Midjourney. https://www.midjourney.com/. 1

[2] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016. 2

[3] Romain Beaumont. Clip retrieval: Easily compute clip embeddings and build a clip retrieval system with them. https://github.com/rom1504/clip-retrieval, 2022. 7

[4] Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *USENIX Security Symposium*, pages 5253–5270, 2023. 1, 2, 3, 4, 6, 7, 8

[5] Chen Chen, Daochang Liu, Siqi Ma, Surya Nepal, and Chang Xu. Private image generation with dual-purpose auxiliary classifier. In *CVPR*, 2023. 2

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009. 7

[7] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, pages 8780–8794, 2021. 1, 3

[8] Tim Dockhorn, Tianshi Cao, Arash Vahdat, and Karsten Kreis. Differentially Private Diffusion Models. *Transactions on Machine Learning Research*, 2023. 2

[9] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014. 2

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 1

[11] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv:2207.12598*, 2022. 1, 3

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, pages 6840–6851, 2020. 1, 3

[13] Nikhil Kandpal, Eric Wallace, and Colin Raffel. Deduplicating training data mitigates privacy risks in language models. In *ICML*, 2022. 2

[14] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 1

[15] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019.

[16] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *CVPR*, pages 8110–8119, 2020. 1

[17] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 1

[18] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. 1

[19] Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *ICCV*, 2023. 1, 2, 4, 7

[20] Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. In *ACL*, pages 8424–8445, 2022. 2

[21] Songhua Liu, Kai Wang, Xingyi Yang, Jingwen Ye, and Xinchao Wang. Dataset distillation via factorization. *Advances in neural information processing systems*, 35:1100–1113, 2022. 3

[22] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015. 2

[23] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, 2021. 1, 3, 7, 8

[24] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. 2021. 1

[25] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing*, pages 722–729, 2008. 2

[26] Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze. A self-supervised descriptor for image copy detection. In *CVPR*, pages 14532–14542, 2022. 4

[27] 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. *arXiv preprint arXiv:2103.00020*, 2021. 2

[28] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. 2022. 1

[29] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *ICML*, pages 1530–1538, 2015. 1

[30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 1, 7, 8

[31] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 1

[32] Joseph Saveri and Butterick Matthew. Stable diffusion litigation, 2023. 2023. 1

[33] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models. 2022. 1

[34] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, pages 2256–2265, 2015. 1

[35] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *CVPR*, pages 6048–6058, 2023. 1, 2, 4, 6, 7

[36] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and mitigating copying in diffusion models. In *NeurIPS*, 2023. 1, 2, 4, 5, 6, 7

[37] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 4

[38] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021. 3

[39] Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018. 3

[40] Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. Differentially private generative adversarial network. *arXiv preprint arXiv:1802.06739*, 2018. 2

[41] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. In *arXiv preprint arXiv:1506.03365*, 2015. 7