Diffusion Probabilistic Models Generalize when They Fail to Memorize

TaeHo Yoon [1] **Joo Young Choi** [1] **Sehyun Kwon** [2] **Ernest K. Ryu** [1 2]

Abstract

In this work, we study the training of diffusion probabilistic models through a series of hypotheses and carefully designed experiments. We call our key finding the memorization-generalization dichotomy, and it asserts that generalization and memorization are mutually exclusive phenomena. This contrasts with the modern wisdom of supervised learning that deep neural networks exhibit "benign" overfitting and generalize well despite overfitting the data.

1. Introduction

Since the advent of some seminal works (Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Ho et al., 2020), diffusion probabilistic models have quickly assumed a dominant position within the generative model literature for its superior performance (Dhariwal & Nichol, 2021) and the ease of controlled generation (Song et al., 2021; Dhariwal & Nichol, 2021; Ho & Salimans, 2021; Ramesh et al., 2021; Rombach et al., 2022; Nichol et al., 2022). Recently, however, diffusion models have been accused of the tendency to memorize the training dataset (Somepalli et al., 2022; Carlini et al., 2023). Although this memorizing behavior seems to have significant implications on both practical (privacy issues) and theoretical sides, unfortunately, there is no satisfactory explanation on when and why diffusion models memorize or not.

In this work, we investigate the data memorization of DPMs through a series of hypotheses and controlled experiments. Our central observation, which we call the "memorization-generalization dichotomy", is that for DPMs, generalization and memorization are mutually exclusive phenomena, which contrasts with the modern wisdom of supervised learning that deep neural networks exhibit "benign" overfitting and generalize well despite overfitting the data. We experimentally demonstrate the memorization-generalization dichotomy by showing that preventing memorization (by reducing the model size or by injecting additional dummy data that the model must expend some capacity to learn) induces generalization. We furthermore show that the memorization-generalization dichotomy can manifest at the level of classes, where the model simultaneously memorizes some classes of the data while generalizing with respect to other classes.

[1]Department of Mathematical Sciences, Seoul National University, Seoul, Korea [2]Interdisciplinary Program in Artificial Intelligence, Seoul National University, Seoul, Korea. Correspondence to: TaeHo Yoon <tetrzim@gmail.com>, Ernest K. Ryu <ernestryu@snu.ac.kr>.

Accepted to ICML workshop on Structured Probabilistic Inference & Generative Modeling.

2. Related Works

Here we only provide a brief overview on diffuion probabilistic models. The remaining list of prior works is presented in Appendix A.

At a high-level, diffusion probabilistic models progressively destruct the input data distribution through a noising process, train a model to learn essential information from the process, and invert the process to regenerate the data distribution from noise. Here we provide elegant continuous description of the noising process established by Song et al. (2021).

Given d-dimensional $X_0 \sim p_{\text{true}}$, consider the SDE

$$dX_t = -\frac{\beta_t}{2} X_t \, dt + \sqrt{\beta_t} \, dW_t$$

where W_t is the Brownian motion in $\mathbb{R}^d$. Then

$$X_t | X_0 \sim \mathcal{N}\left(\sqrt{\bar{\alpha}_t} X_0, (1-\bar{\alpha}_t) I\right), \quad \bar{\alpha}_t = e^{-\int_0^t \beta_s \, ds}.$$

Let p_t denote the density of X_t for $t \geq 0$. Then along the following reverse SDE, derived using the Anderson's theorem (Anderson, 1982), the probability density flow coincides with the original forward SDE:

$$d\overline{X}_t = \left(-\frac{\beta_t}{2} - \nabla_x \log p_t(\overline{X}_t)\right) dt + \sqrt{\beta_t} d\overline{W}_t$$

Therefore, by approximating the *score* $\nabla_X \log p_t(\cdot)$ with a time-dependent neural network $s_\theta(\cdot, t)$, one can start with $\overline{X}_T \sim \mathcal{N}(0, I)$ (assuming T is large enough) and recover $\overline{X}_0$ that is approximately distributed according to p_{true} by following the reverse SDE. In practice, one often considers the *error* network $\varepsilon_\theta(x,t) = -\sqrt{1-\bar{\alpha}_t} s_\theta(x,t)$ and use a discretized version of the reverse SDE (Ho et al., 2020; Dhariwal & Nichol, 2021). One can also perform conditional generation by additionally conditioning the score or the error network on labels (or tokens) y.

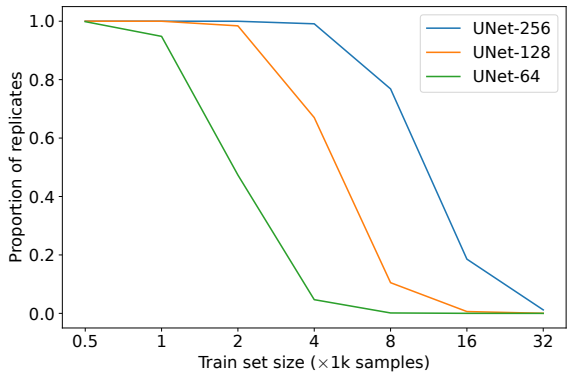


Figure 1. Proportion of generated samples replicating the training data from diffusion models of different scales, trained on different sizes of CIFAR-10 subsets.

3. Memorization-generalization dichotomy

In this work, we define *learning* to include the following two types: *rote learning* (i.e., memorization) and the *conceptual learning* (which enables generalization). A generative model performs perfect rote learning when it learns to generate from $\frac{1}{n}\sum_{X\in S}\delta_X$, where $S=\{X_1,\dots,X_N\}$ is the training set. In this case, we say the model has perfectly *memorized* the training set and say the generated samples are replicates of the training data. A generative model performs perfect conceptual learning when it learns to generate from p_{true}, where p_{true} is the true underlying distribution generating the data $X_1,\dots,X_N\sim p_{\text{true}}$. In this case, we say the model *generalizes*, and this is the desired behavior of a generative model.

In this work, we present the *memorization-generalization dichotomy* of DPMs: the phenomenon that conceptual learning happens only when rote learning fails. Of course, training can entirely fail and achieve neither type of learning, but the claim is that it is useful to view memorization as an impediment to generalization.

On the one hand, this dichotomy runs counter to the modern view that in deep *supervised* learning, trained deep neural networks generalize well despite memorizing (interpolating) training data (Zhang et al., 2017; Belkin et al., 2019). This view, referred to as the "benign overfitting" (Belkin, 2021), claims that overparameterization is not detrimental to the generalization performance. In contrast, our memorization-generalization dichotomy implies that DPMs generalize when they are underparameterized, which is consistent with the classical statistical view that overfitting should be avoided.

This observation is interesting as it indicates that the nature of learning of DPMs (and perhaps for generative models as well) is different from that of modern deep supervised learning, and this distinction may be crucial in understanding DPMs. In the following, we present a series of hypotheses and experiments that demonstrate the value of our dichotomy in understanding the learning process of DPMs.

3.1. Hypothesis 1: Memorization capacity exists

In this section, we show how memorization (rote learning) is affected by the complexity (size) of train data and model capacity. We first define the model's capacity, motivated by prior works Nakkiran et al. (2021); Feng et al. (2021). Let G_θ be a parametrized generative model (neural network) that maps a noise (from the prior distribution p_z) to a data point in $\mathcal{X}$. Let $\mathcal{T}$ denote a training algorithm, which takes as input the train set $S=\{x_1,\dots,x_n\}\subset\mathcal{X}$ and outputs a trained $\theta^{\mathcal{T}}(S)$.

Definition 3.1. (Memorization Capacity) Define the *memorization capacity* of G_θ with respect to the data distribution $\mathcal{D}$ (with density p_{data}), a duplicate-detecting criterion (d,δ) and $\epsilon\in(0,1)$ as

$$\mathrm{MC}_{\mathcal{D},\delta,\epsilon}(G_\theta;\mathcal{T}) = \max\left\{ n \;\middle|\; \operatorname*{Prob}_{\substack{S\sim\mathcal{D}^n\\ z\sim p_z}}\left[\min_{x\in S} d(G_{\theta^{\mathcal{T}}(S)}(z),x)\le\delta\right]\ge 1-\epsilon\right\}.$$

Denote $\mathrm{MC}_{\mathcal{D},\delta,0}(G_\theta;\mathcal{T})=\inf_{\epsilon\in(0,1)}\mathrm{MC}_{\mathcal{D},\delta,\epsilon}(G_\theta;\mathcal{T})$.

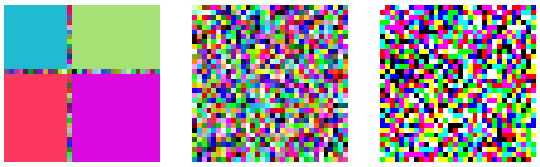
For the sake of formal definition, we assume that there is a bivariate function $d:\mathcal{X}\times\mathcal{X}\to\mathbb{R}_{\ge 0}$ such that $d(x,x')$ being small ($\le\delta$) indicates that the sample pair (x,x') is indistinguishable under human perception. We discuss a useful criterion for image data later within this section.

Now we present the main hypothesis of this section:

Hypothesis 1. For a diffusion model G_θ with non-trivial architecture, $\mathrm{MC}_{\mathcal{D},\delta,\epsilon}(G_\theta;\mathcal{T})$ is a positive number and increases with the number of parameters ($\dim\theta$) for any $\epsilon\in(0,1)$. When $|S|\le\mathrm{MC}_{\mathcal{D},\delta,0}(G_\theta;\mathcal{T})$, a diffusion model memorizes the train set. As $|S|$ grows past $\mathrm{MC}_{\mathcal{D},\delta,0}(G_\theta;\mathcal{T})$, the probability of train set replication drops quickly.

Experiments. We verify Hypothesis 1 through an experiment, of which result is summarized in Figure 1. We consider 3 variations of neural network: UNet-64, UNet-128 and UNet-256, that share the same architecture (from Nichol & Dhariwal (2021)) but have different *width* parameters. (UNet-128 is 4 times larger than UNet-64, and UNet-256 is 4 times larger than UNet-128.) We train DDPM models using each UNet model as the error network, on differently sized train sets consisting of $\{0.5,1,2,4,8,16,32\}$k randomly sampled CIFAR-10 images. We generate 10k samples from each trained model, and measure the proportion of train set replicates therein.

Criterion for replicate detection. L^2, L^∞ norms and their variants are, obviously, good candidate criteria for



(a) Mondrian (b) Dummy (c) Random ink

Figure 2. Examples of artificial data

detecting a train data replication, but we find that they are either overly lenient or conservative compared to our visual inspection. Inspired by a detection rule used in Carlini et al. (2023), we use the following slightly simplified "ratio criterion", which aligns accurately with our perception of duplication (Appendix D); given a generated sample W, find $X_1, X_2 \in S$ with smallest and second smallest L^2 norm, $\|W - X\|_2$, and mark W as a replicate if $\frac{\|W-X_1\|_2}{\|W-X_2\|_2} < \frac{1}{3}$.

Implications of the result. The experiment result indicates that each diffusion model performs rote learning up to certain point, and then display an exponential decay trend in replicating behavior with respect to increasing train set size. Additionally, we observe approximately factor 2 difference in the memorization capacity between models with factor 2 width difference. We speculate that a gigantic UNet model (UNet-8192 by the back-of-the-envelope calculation) might be capable of memorizing the entire CIFAR-10 dataset; assuming that this is the case, we can link the successful generalization of diffusion models displayed in recent works (Ho et al., 2020; Nichol & Dhariwal, 2021) to the inability to memorize (underparemetrization). However, the current result on its own is insufficient as an evidence that generalization is equivalent to the failure in memorization; we corroborate this further in Section 3.2.

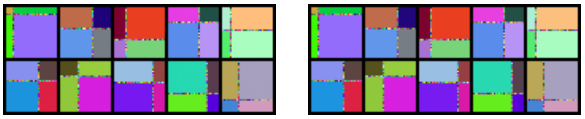
Connection to prior work. In the context of GANs, Feng et al. (2021) makes an observation similar to Hypothesis 1. However, they do not explicitly consider the model capacity as a factor affecting memorization, and they focus primarily on quantitative relationship between the data complexity and replication probability rather than explaining generalization.

3.2. Hypothesis 2: Generalization requires both sufficient data and insufficient capacity

In this section, we demonstrate that conceptual learning indeed occurs due to the failure to retain rote learning by testing the following hypothesis:

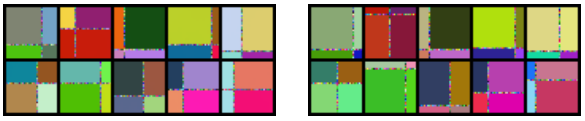
Hypothesis 2. A diffusion model G_θ performs rote learning if $n < \mathrm{MC}_{\mathcal{D},\delta,\epsilon}(G_\theta; \mathcal{T})$ ($\epsilon \approx 0$), *even if* $|S|$ *is large enough to enable generalization* with respect to $\mathcal{D}$. In this case, simply adding dummy samples drawn independently from a disjointly supported distribution $\mathcal{D}'$ to S can induce G_θ to perform conceptual learning with respect to $\mathcal{D}$.

Experiments on artificial data. For clear illustration of conceptual learning, we design a simple procedurally generated data distribution whose defining visual pattern is clear (Figure 2a). We refer to this distribution as Mondrian (abbv. "MO"); it features a cross on which the pixels' RGB values are independent and uniform random, and each quadrant shares a same color defined by independent and uniform random RGB values.



(a) Generated samples (b) Nearest training images

Figure 3. **(Left)** Samples generated by a diffusion model trained with only 2k MO images. **(Right)** The nearest train set images to each sample in the L^2 sense.



(a) Generated samples (b) Nearest training images

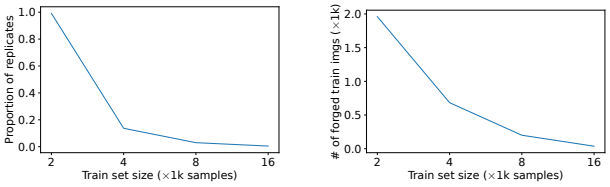
Figure 4. **(Left)** Samples generated by a diffusion model trained with 2k MO images and 6k dummy images. **(Right)** The nearest train set images to each sample in the L^2 sense.

We first train a DDPM model (UNet-128) using 2k images from the MO distribution as train set. We observe that about 92% of generated samples are replicates under the ratio criterion. Indeed, Figure 3 illustrates that the generated samples are even precisely mimicking the random pixels on the cross from training images. Next, we add 6k dummy images from a Gaussian mixture (Figure 2b) with 100 modes (abbv. "GM") to the same set of 2k MO images. Then, we train a class-conditional DDPM model (this is for controlled generation of MO images; the capacity is virtually the same as the unconditional model) on the augmented train set. We observe that only 0.2% of conditionally generated MO images are replicates; Figure 4 shows that generic samples are indeed original, faithfully obeys the MO rules, and are not the replicates of train set images.

Implications of the result. We can summarize the results as follows. **1)** 2k MO images are already sufficient to enable conceptual learning over the MO distribution. **2)** However, DDPM model using UNet-128 memorizes them, because it is capable of doing so. **3)** The model transitions to conceptual learning when supplied with additional dummy data, which only plays the role of occupying the model's capacity without providing any meaningful information. This

strongly supports the claim that conceptual learning is the result of failing to brute-force-memorize, given that the train set possesses enough complexity to enable generalization.

Experiments on CIFAR-10. In this section we demonstrate experiment of the same spirit on CIFAR-10 "car" class. We randomly choose and fix 2k "car" images, and augment this image set with GM images until the total train set size becomes $\{4, 8, 16\}$k. We train conditional DDPM models on each augmented train set. We observe progressive reduction both in the proportion of train set replicates (Figure 5a) and the number of train set images that have been replicated at least once (Figure 5b). Figure 6 shows that original images are generated from the model trained with total 8k images. More sample images are shown in Appendix C.2.



(a) Proportion of replicates (b) # of replication targets

Figure 5. Using 10k samples generated from conditional diffusion models trained on CIFAR-10 car images augmented with GM, we measure **(left)** the proportion of replicates and **(right)** the number of training images that have been at least once replicated.

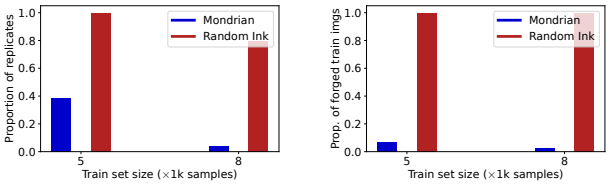
(a) Generated samples (b) Nearest training images

Figure 6. **(Left)** Conditional samples from the model trained with 6k dummies. **(Right)** L^2-nearest train set images to each sample.

3.3. Hypothesis 3: Memorization-generalization dichotomy manifests at the level of classes

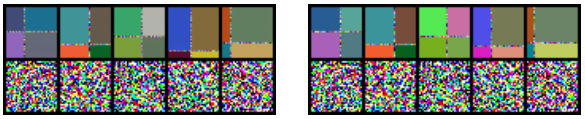
So far, we demonstrated the memorization-generalization dichotomy determined by the size of the whole training dataset. In this section, we investigate whether the dichotomy may apply class-wisely when the dataset is composed of different classes. We hypothesize:

Hypothesis 3. When training a class-conditional diffusion model on a train set containing multiple classes, the memorization-generalization dichotomy may appear separately over each class. In particular, if the train set consists of a majority (large population) and a minority (small population) class, the model memorizes the minority class while generalizing with respect to the majority class.



(a) Proportion of replicates (b) # of replication targets

Figure 7. For each class Mondrian (majority) and Random Ink (minority), we conditionally generate 5k samples from a conditional diffusion model and measure **(left)** the proportion of samples replicating the train set and **(right)** the proportion of training images that have been replicated.



(a) Generated samples (b) Nearest training images

Figure 8. **(Left)** Conditional samples from the model trained with 4.9k MO and 100 RI images. **(Right)** L^2-nearest train set images to each sample. Best viewed digitally: zoom-in is recommended.

Experiments on artificial data. We train a conditional DDPM model with total sample set size of 5k and 8k, where 100 images from each set are the Random Ink images (Figure 2c; abbv. "RI"), all of whose RGB values are independent Bernoulli variables, and the rest are the MO images. The setup makes intuitively clear that the RI images constitute minority. We generate 5k samples conditioned on each class, and measure **1)** the proportion of train set replicates within the samples, and **2)** the proportion of at least once replicated images within the train set, for each class.

Implications of the result. We observe that for the minority class RI, the learning pattern is closer to rote learning. On the other hand, the same model tends to perform conceptual learning with respect to the majority class MO. The result supports Hypothesis 3, indicating that a model can perform different types of (rote or conceptual) learning over different classes. We conjecture that this insight can further extend to the dichotomy at the level of abstract concepts, explaining e.g., a text-conditioned diffusion model well-generalizing with respect to concepts like "sky" or "woods" while memorizing "Yann LeCun".

4. Conclusion

We propose to view generalization of diffusion models as a failure to memorize the training data. We speculate that the memorizing trait of diffusion models is intimately related to their impressive capability as generative models, and a more in-depth study of their memorization will pave the way toward understanding the source of their performance.

References

Anderson, B. D. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

Belkin, M. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numerica*, 30:203–248, 2021.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramèr, F., Balle, B., Ippolito, D., and Wallace, E. Extracting training data from diffusion models. *arXiv preprint arXiv:2301.13188*, 2023.

Chen, D., Yu, N., Zhang, Y., and Fritz, M. Gan-leaks: A taxonomy of membership inference attacks against generative models. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, pp. 343–362, 2020.

den Burg, G. J. V. and Williams, C. On memorization in probabilistic deep generative models. *Advances in Neural Information Processing Systems*, 2021.

Dhariwal, P. and Nichol, A. Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems*, 2021.

Duan, J., Kong, F., Wang, S., Shi, X., and Xu, K. Are diffusion models vulnerable to membership inference attacks? *arXiv preprint arXiv:2302.01316*, 2023.

Feldman, V. Does learning require memorization? A short tale about a long tail. *Symposium on Theory of Computing*, pp. 954–959, 2020.

Feng, Q., Guo, C., Benitez-Quiroz, F., and Martinez, A. M. When do gans replicate? on the choice of dataset size. *International Conference on Computer Vision*, 2021.

Hayes, J., Melis, L., Danezis, G., and De Cristofaro, E. Logan: Membership inference attacks against generative models. *arXiv preprint arXiv:1705.07663*, 2017.

Hilprecht, B., Härterich, M., and Bernau, D. Monte carlo and reconstruction membership inference attacks against generative models. *Proc. Priv. Enhancing Technol.*, 2019 (4):232–249, 2019.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. *Advances in Neural Information Processing Systems Workshop on Deep Generative Models*, 2021.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 2020.

Hu, H. and Pang, J. Membership inference of diffusion models. *arXiv preprint arXiv:2301.09956*, 2023.

Jia, J., Salem, A., Backes, M., Zhang, Y., and Gong, N. Z. Memguard: Defending against black-box membership inference attacks via adversarial examples. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pp. 259–274, 2019.

Li, Z. and Zhang, Y. Membership leakage in label-only exposures. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pp. 880–895, 2021.

Long, Y., Bindschaedler, V., Wang, L., Bu, D., Wang, X., Tang, H., Gunter, C. A., and Chen, K. Understanding membership inferences on well-generalized learning models. *arXiv preprint arXiv:1802.04889*, 2018.

Meehan, C., Chaudhuri, K., and Dasgupta, S. A non-parametric test to detect data-copying in generative models. *International Conference on Artificial Intelligence and Statistics*, 2020.

Melis, L., Song, C., De Cristofaro, E., and Shmatikov, V. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE symposium on security and privacy (SP)*, pp. 691–706. IEEE, 2019.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021.

Nasr, M., Shokri, R., and Houmansadr, A. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE symposium on security and privacy (SP)*, pp. 739–753. IEEE, 2019.

Nichol, A. Q. and Dhariwal, P. Improved denoising diffusion probabilistic models. *International Conference on Machine Learning*, 2021.

Nichol, A. Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., Mcgrew, B., Sutskever, I., and Chen, M. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. *International Conference on Machine Learning*, 2022.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation. *International Conference on Machine Learning*, 2021.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. *Conference on Computer Vision and Pattern Recognition*, 2022.

Sablayrolles, A., Douze, M., Schmid, C., Ollivier, Y., and Jégou, H. White-box vs black-box: Bayes optimal strategies for membership inference. In *International Conference on Machine Learning*, 2019.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., Salimans, T., Ho, J., Fleet, D. J., and Norouzi, M. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 2022.

Salem, A., Zhang, Y., Humbert, M., Fritz, M., and Backes, M. ML-Leaks: Model and data independent membership inference attacks and defenses on machine learning models. In *Network and Distributed Systems Security Symposium 2019*. Internet Society, 2019.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models. *Symposium on Security and Privacy*, pp. 3–18, 2017.

Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. *ICML*, 2015.

Somepalli, G., Singla, V., Goldblum, M., Geiping, J., and Goldstein, T. Diffusion art or digital forgery? Investigating data replication in diffusion models. *arXiv preprint arXiv:2212.03860*, 2022.

Song, L. and Mittal, P. Systematic evaluation of privacy risks of machine learning models. In *USENIX Security Symposium*, volume 1, pp. 4, 2021.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 2019.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *International Conference on Learning Representations*, 2021.

Truex, S., Liu, L., Gursoy, M. E., Yu, L., and Wei, W. Demystifying membership inference attacks in machine learning as a service. *IEEE Transactions on Services Computing*, 14(6):2073–2089, 2021.

Webster, R., Rabin, J., Simon, L., and Jurie, F. This person (probably) exists. identity membership attacks against gan generated faces. *arXiv preprint arXiv:2107.06018*, 2021.

Wu, B., Zhao, S., Chen, C., Xu, H., Wang, L., Zhang, X., Sun, G., and Zhou, J. Generalization in generative adversarial networks: A novel perspective from privacy protection. *Advances in Neural Information Processing Systems*, 32, 2019.

Wu, Y., Yu, N., Li, Z., Backes, M., and Zhang, Y. Membership inference attacks against text-to-image generation models. *arXiv preprint arXiv:2210.00968*, 2022.

Yeom, S., Giacomelli, I., Fredrikson, M., and Jha, S. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pp. 268–282. IEEE, 2018.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. *International Conference on Learning Representations*, 2017.

A. Other Related Works

Memorization and generalization. In the supervised learning setup, the concept of memorization is commonly viewed as equivalent of overfitting, i.e., achieving perfect accuracy on the train set. Zhang et al. (2017) shows that neural networks successfully generalize even when they are overparameterized and thus are capable of memorizing the entire train set. Feldman (2020) argues that memorization is actually a necessary component for generalization. Belkin et al. (2019) reports the double descent phenomenon, explaining the relationship between model capacity and overfitting. Nakkiran et al. (2021) proposes the concept of effective model complexity, and further generalizes the double descent phenomenon.

In the context of generative models, Wu et al. (2019) investigates generalization of GANs from the perspective of privacy protection. Meehan et al. (2020) proposes a statistical test methodology on whether a generative model is overfitting the train set in the sense of "data-copying". For VAEs, den Burg & Williams (2021) provides an designed empirically memorization score. Feng et al. (2021) probes the quantitative relationship between the dataset complexity and memorization for GANs.

A closely related concept to memorization is the membership inference attack (MIA), which aims to determine whether a given sample originates from the training dataset or not (Shokri et al., 2017). MIA has been studied extensively both in classification tasks (Shokri et al., 2017; Yeom et al., 2018; Long et al., 2018; Salem et al., 2019; Jia et al., 2019; Sablayrolles et al., 2019; Melis et al., 2019; Nasr et al., 2019; Truex et al., 2021; Li & Zhang, 2021; Song & Mittal, 2021) and generation tasks (Hayes et al., 2017; Hilprecht et al., 2019; Chen et al., 2020; Webster et al., 2021). However, the threat model for MIA is concerned with a broader concept of attack whose possibility persists even when the trained neural network is not responsible for direct reconstruction of the training data.

Memorization in DPMs While some prior works (Saharia et al., 2022) state that overfitting is not a significant issue for training diffusion models, recent works report cases where diffusion models do memorize the training data and replicate them during the sampling process (Somepalli et al., 2022; Carlini et al., 2023). Other works (Wu et al., 2022; Hu & Pang, 2023; Duan et al., 2023) show that diffusion models can be susceptible to the membership inference attacks. The prior works listed above, however, are mainly concerned about privacy issues or detection of train set replication within large-scale diffusion models. We a different viewpoint from them; our goal is to understand the generalization of diffusion models based on the memorization-generalization dichotomy hypothesis and controlled experiments within relatively simple setups.

B. Experiment Details

All experiments use the official PyTorch implementation[1] of IDDPM (Improved Denoising Diffusion Probabilistic Models) (Nichol & Dhariwal, 2021). We use the same model hyperparameters and training configurations as the default setup for the CIFAR-10 dataset provided by the official repository, except possibly for the width of residual blocks, batch size and training iteration (Table 1).

	UNet-64	UNet-128	UNet-256
num_channels	64	128	256
num_res_blocks	3		
learn_sigma	True		
drop_out	0.3		
diffusion_steps	4000		
noise_scheduler	cosine		
optimizer	Adam		
lr	0.0001		

Table 1. Model hyperparameters and training configuration for experiments

[1] https://github.com/openai/improved-diffusion

B.1. Experiments of Section 3.1

For all models in this experiment, we use batch size 128 and train for 500k iterations.

B.2. Experiments of Section 3.2

All networks in this experiment are the UNet-128 models. We use batch size 256 and train for 100k iterations.

We generate the Gaussian Mixture (GM) dummy images according to the following rule: **1)** We first fix a set of 100 images with independent and uniform random pixel values (normalized within $[0, 1]$), then **2)** randomly select one of the 100 fixed images X and sample an image from the isotropic Gaussian distribution with mean X and standard deviation 0.5 for each pixel, and finally **3)** clip the sampled image to take values within $[0, 1]$.

We use unconditional error network $\varepsilon_\theta(x, t)$ when training with 2k images (Mondrian and CIFAR-10 cars), which has 52.5M trainable parameters. We use conditional error network $\varepsilon_\theta(x, t, y)$ (which takes class label y as input and combines the label embedding of y with the timestep embedding) when training with GM dummy images. The conditional network has 53.0M trainable parameters.

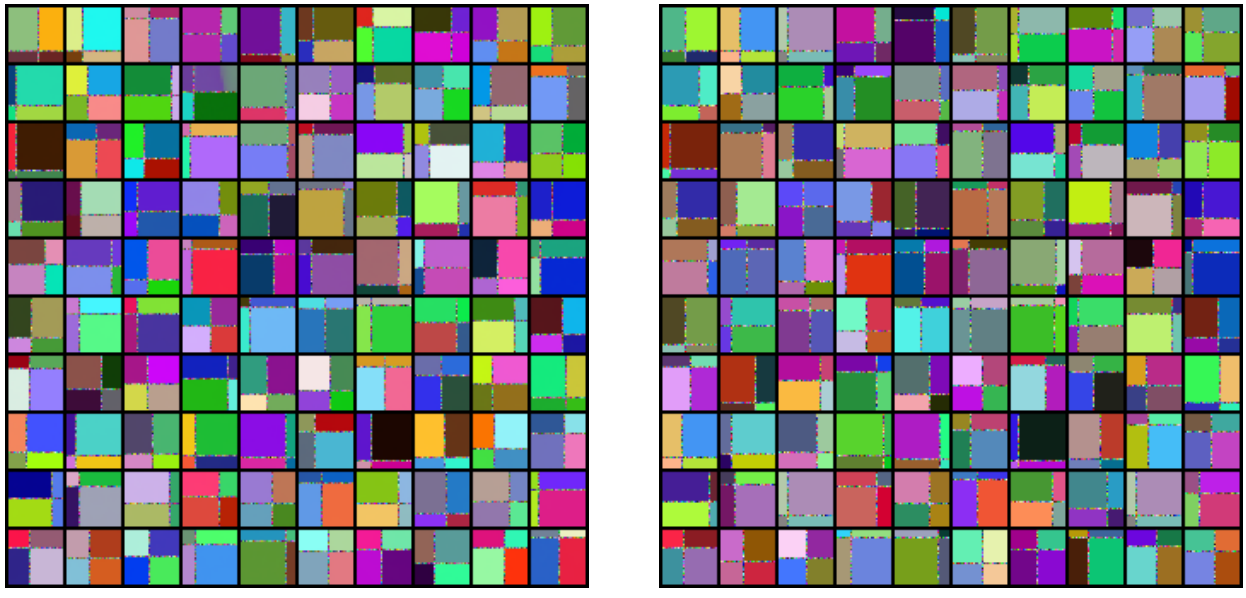
B.3. Experiments of Section 3.3

All networks in this experiment are the class-conditional UNet-128 models. We use batch size 256 and train for 100k iterations.

C. Additional Experiment Results for Section 3.2

C.1. Using unconditional networks for training on dummy-augmented data

Although conditional networks do not seem to have significantly different memorization capacity from unconditional networks (considering that the difference in their number of parameters is less than 1%), for the sake of completeness, we also train an *unconditional* UNet-128 DDPM model on the mixture of 2k MO and 6k GM images. We unconditionally generate 10k images and find that 2,626 images out of them are MO, using a separately trained binary classifier discriminating MO images from GM images. The ratio criterion detects 0 replicates within the 2,626 MO images, which is clearly consistent with our previous experiments (Figure 9).



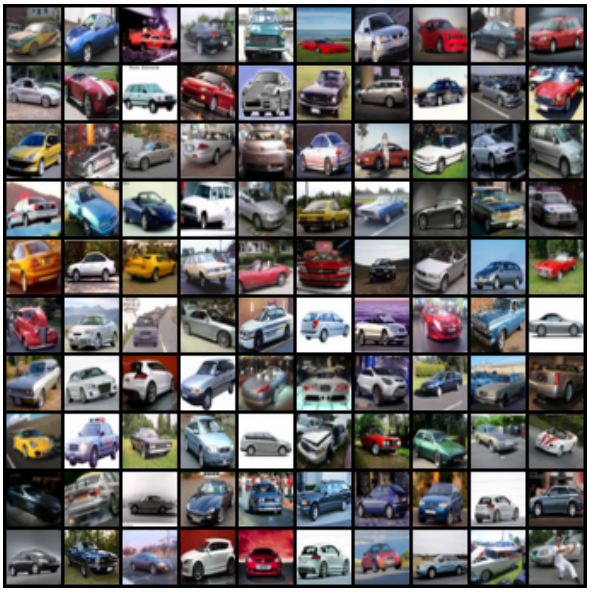
(a) MO samples from unconditional generation

(b) Nearest training images

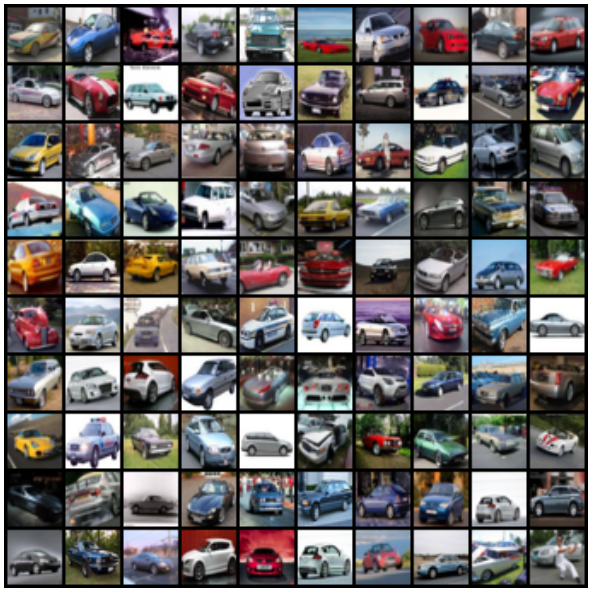
Figure 9. **(Left)** MO class samples extracted from unconditionally generated samples from a diffusion model trained with 2k MO images and 6k dummy images. **(Right)** The nearest train set images to each sample in the L^2 sense.

C.2. Additional samples from the CIFAR-10 experiment

We display 100 "car" samples each from the unconditional model and the conditional models trained with dummies, and the train set images that are nearest to each sample in the L^2 sense.

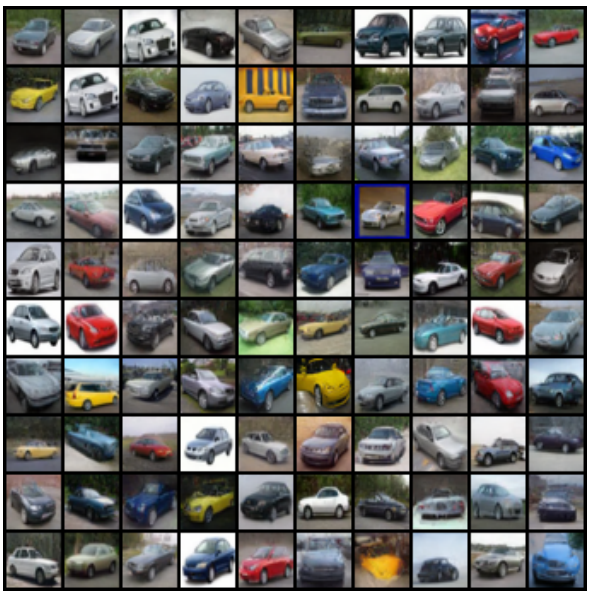


(a) Generated samples

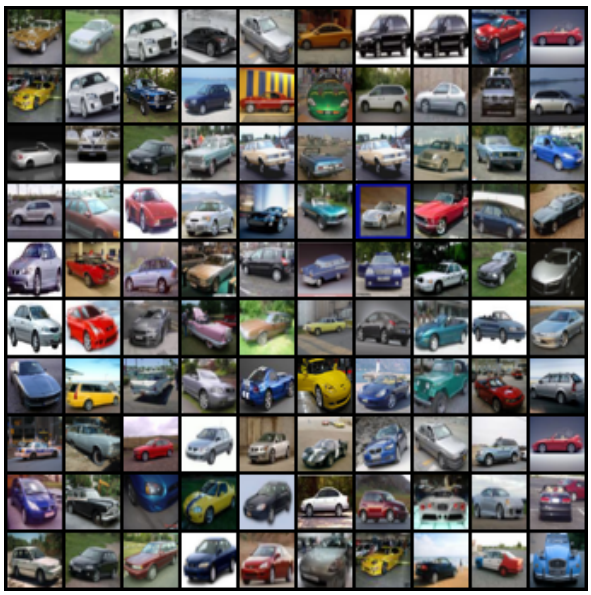


(b) Nearest training images

Figure 10. **(Left)** Unconditional samples from the model trained without dummies. **(Right)** L^2-nearest train set images to each sample.

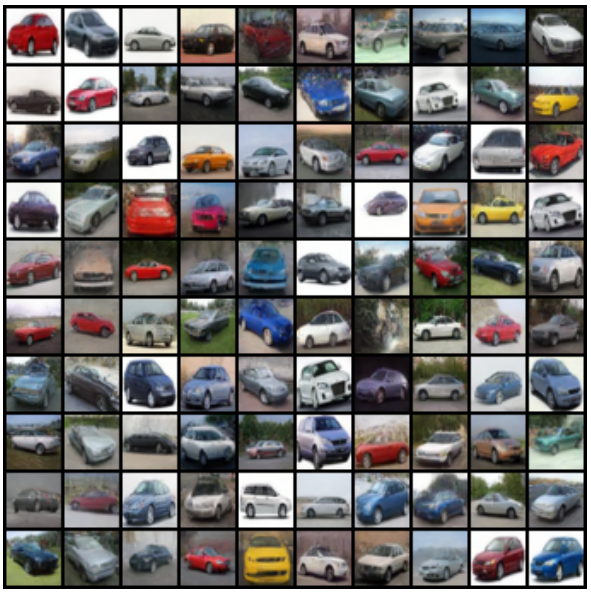


(a) Generated samples

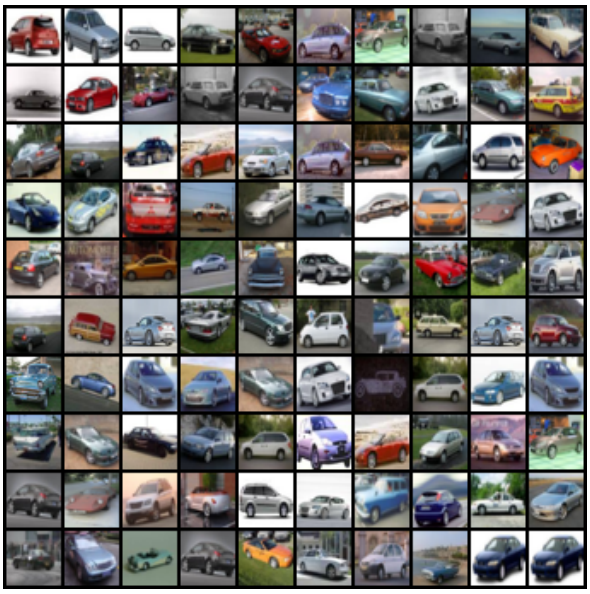


(b) Nearest training images

Figure 11. **(Left)** Conditional samples from the model trained with 2k dummies. **(Right)** L^2-nearest train set images to each sample.

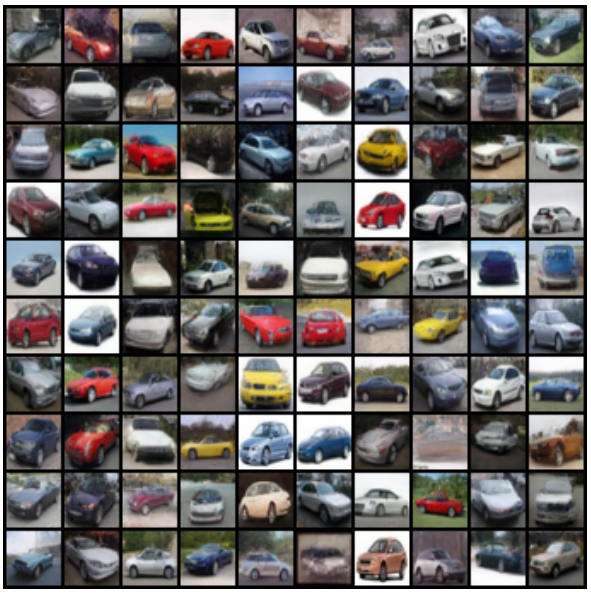


(a) Generated samples

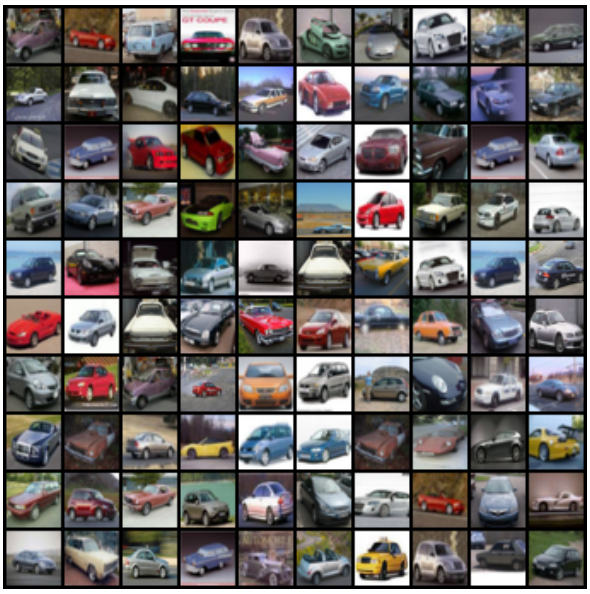


(b) Nearest training images

Figure 12. **(Left)** Conditional samples from the model trained with 6k dummies. **(Right)** L^2-nearest train set images to each sample.



(a) Generated samples

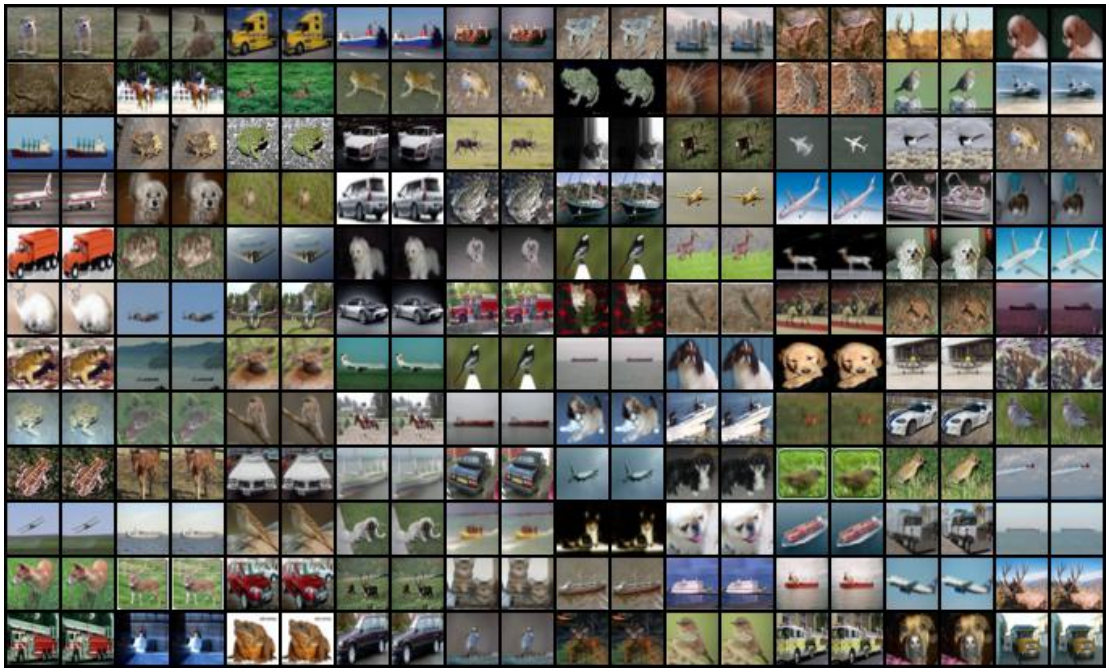


(b) Nearest training images

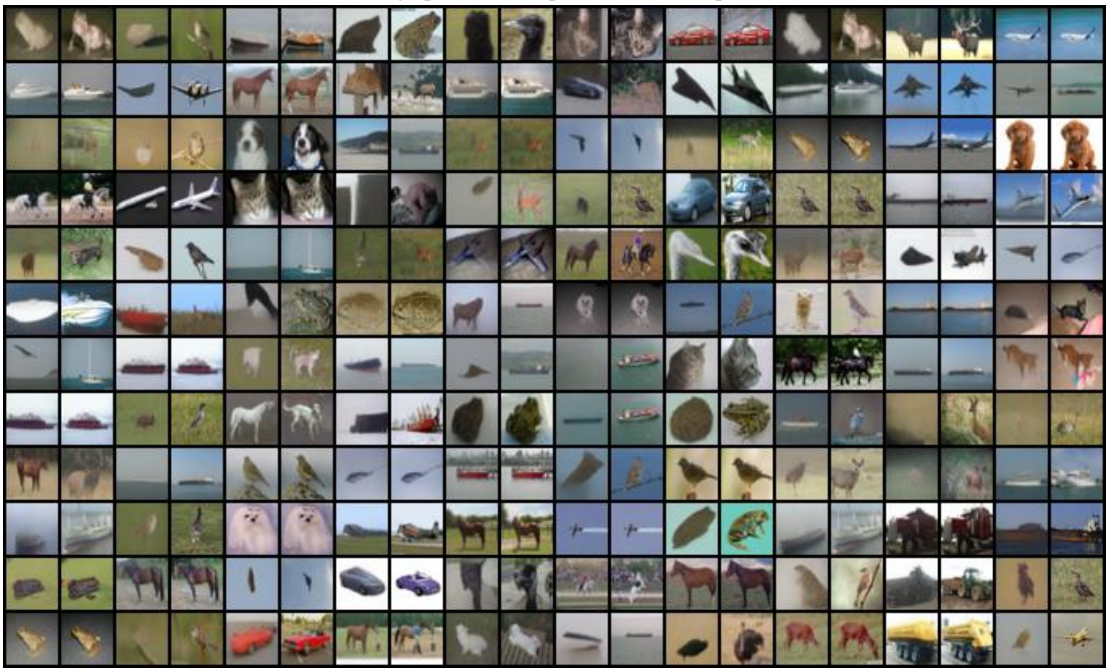
Figure 13. **(Left)** Conditional samples from the model trained with 14k dummies. **(Right)** L^2-nearest train set images to each sample.

D. Validation of the Ratio Criterion

We provide a simple visual justification for the ratio criterion we use for replicate detection in our experiments. Below, we display multiple pairs of images from the setup of Section 3.1, each consisting of a generated sample image and train set image nearest to it in the L^2 sense. In the first set, we only gather samples that are marked as replicates according to the ratio criterion (i.e., $\frac{\|W-X_1\|_2}{\|W-X_2\|_2} < \frac{1}{3}$), while in the second set, we only gather samples that are marked as non-duplicates.



(a) Image pairs for samples marked as replicates



(b) Image pairs for samples makred as non-replicates

Figure 14. Image pairs consisting of a generated sample image and the L^2-nearest train set image to it, for **(top)** samples makred as replicates by the ratio criterion and **(bottom)** samples marked as non-duplicates.