Membership Inference Attack on Diffusion Models via Quantile Regression

Zhiwei Steven Wu *
Carnegie Mellon University
Amazon AWS AI/ML

Shuai Tang *
Amazon AWS AI/ML

Sergul Aydore
Amazon AWS AI/ML

Michael Kearns
University of Pennsylvania
Amazon AWS AI/ML

Aaron Roth
University of Pennsylvania
Amazon AWS AI/ML

Abstract

Recently, diffusion models have demonstrated great potential for image synthesis due to their ability to generate high-quality synthetic data. However, when applied to sensitive data, privacy concerns have been raised about these models. In this paper, we evaluate the privacy risks of diffusion models through a *membership inference (MI) attack*, which aims to identify whether a target example is in the training set when given the trained diffusion model. Our proposed MI attack learns a single quantile regression model that predicts (a quantile of) the distribution of reconstruction loss for each example. This enables us to identify a unique threshold on the reconstruction loss tailored to each example when determining their membership status. We show that our attack outperforms the prior state-of-the-art MI attack and avoids their high computational cost from training multiple shadow models. Consequently, our work enriches the set of practical tools for auditing the privacy risks of large-scale generative models.

1 Introduction

Diffusion models, based on generative neural networks, have gained attention in the field of image generation [9, 19]. It has been shown that diffusion models are remarkably capable of generating images that are higher-quality than previous approaches such as GANs and VAEs, while also being more scalable. However, as the size of these models has grown drastically over the last decade, so has the privacy concern that these large-scale diffusion models may reveal sensitive information on the dataset they are trained on.

One of the most popular methods to evaluate the privacy risks of ML models is *membership inference (MI) attacks* [24, 12, 11, 13, 16, 2, 10, 18] , in which an attacker aims to determine if a target example belongs to the training dataset given the trained model. MI captures the privacy risk that the presence of a data set could reveal sensitive information. For example, membership in a medical dataset may indicate a particular disease. In addition, MI attacks can be a building block for other more sophisticated attacks such as *extraction attacks* on generative models [3]. Prior work in MI attacks typically assumes that the attacker has some side information, such as auxiliary examples drawn from the same data distribution P or a similar one [2, 1, 8, 22]. In general, a successful MI attack with reasonable side information is a strong indicator of the failure of privacy protection. Finally, when applied to differentially private algorithms [7], MI attacks can serve as privacy auditing tools by

[0]Shuai and Steven are the lead authors, and other authors are ordered alphabetically. shuat@amazon.com

Workshop on Regulatable Machine Learning at the 37th Conference on Neural Information Processing Systems (RegML @ NeurIPS 2023).

providing lower bounds on the privacy parameters, which in turn assess the tightness of the privacy analyses [11, 15] and help identify potential errors in the privacy proof or implementation [21, 20].

A majority of the existing MI attacks focus on supervised learning [24, 23, 13, 16, 18, 2], and there has been significantly less development on MI attacks against generative models (e.g., [8, 22]). The goal of our work is to develop strong MI attacks against state-of-the-art diffusion models.

Our work extends the quantile-regression-based attacks in [1] for supervised learning to attacks for diffusion models. For a given trained diffusion model θ, our attack first learns a quantile regression model on public auxiliary data that predicts the α-quantile $q(z)$ of the θ's reconstruction loss on each example z (formally defined in Definition 2.1). Then for each example, we indicate that it is a member of the training set if its reconstruction loss is lower than the predicted α-quantile. By design, the attack has a false positive rate of α, that is the probability that it incorrectly declares a randomly selected point z that was not used in training to have been used in training is α. We evaluate our attack on diffusion models trained on image datasets, and demonstrate three major advantages:

1. Our quantile-regression-based attack obtains state-of-the-art accuracy on several popular vision datasets. Even though our attacks leverage the same reconstruction loss function considered in [6], their attack leverages the same *marginal approach* in [24] that applies a that applies a uniform threshold (that is, the α-quantile on the marginal distribution over the reconstruction loss) across all examples. In comparison, our attack is *conditional* since it applies a finer-grained per-example threshold when performing membership inference.
2. Compared to the prior state-of-the-art MI attacks against diffusion models [17], we achieve higher accuracy without suffering their computational cost. Similar to the LiRA attack proposed by [2], the GSA attack in [17] requires training multiple *shadow models*, each of which is obtained by running the training algorithm on a randomly drawn dataset. While the accuracy of the MI attack improves as the number of shadow models increases, their approach also becomes computationally prohibitive. In comparison, our approach only requires learning a *single* quantile regression model.
3. Since our attack does not rely on shadow models, it also requires significantly fewer details about the training algorithm, such as hyperparameters and network architecture used in training. In fact, our attack is effective even though the neural network for the quantile regression model has significantly fewer parameters than the attacked diffusion models.

2 Membership Inference Attacks on Diffusion Models

Membership inference (MI) is a common privacy attack that attempts to predict whether a given example was used to train a machine learning model [24, 12, 11, 13, 16, 2, 10, 18]. Our work focuses on performing MI attacks on diffusion models.

Problem statement. Given a training dataset $\mathbf{Z}$ drawn from an underlying distribution P, a diffusion model θ is trained on $\mathbf{Z}$. The goal of a membership inference attack is to infer whether a target example z^* was included in the training set $\mathbf{Z}$ or not.

Adversary's side information. Similar to almost all prior work on membership inference [2, 6, 1, 18], we assume the adversary has access to some public data drawn from P. In the standard terminology of MI, there are also two types of access to the algorithm's output. In a *black-box* attack, the adversary only has access to the generated synthetic dataset S. In a *white-box attack*, the adversary has access to the generative model G. In this work, we focus on white-box attacks.

The reconstruction loss function, termed as t-error [6], is used in our MI attack.

Definition 2.1. (t-error) For a given sample $z_0 \sim P$ and the deterministic reverse result $\tilde{z}_t = \Phi_\theta(z_0, t)$ at times t, the approximated posterior estimation error at step t is defined as t-error:

$$\hat{\ell}_t(\theta, z_0) = ||\psi_\theta(\phi_\theta(\tilde{z}_t, t), t) - \tilde{z}_t||^2 \tag{1}$$

Intuitively, the t-error function measures how much we change $\tilde{z}_t$ if we take one step in the deterministic diffusion process ϕ_θ and then rewind back with one step of deterministic denoising ψ_θ. While this loss function is not what the training algorithm optimizes, it provides a deterministic approximation to loss function during training [6, 9]. Thus, smaller t-error values provides evidence that z_0 was used to train the model θ.

Algorithm 1 Quantile Regression MI attacks for Diffusion Model

Require: A set of auxiliary examples D drawn from P, target example z^*, trained model from the algorithm θ, a choice of t for t-error function. Target false-positive rate α.

Require: A quantile regression learner Q (that e.g., minimize the pinball loss or fits a parametric density model).

for each $z \in D$ **do**

evaluate the score $\hat{\ell}_t(\theta, z)$

end for

Use learner Q to learn a quantile regression model q such that $q(z)$ predicts the α-quantile of the score $\hat{\ell}_t(\theta, z)$ conditioned on z. $\mathbb{P}[\hat{\ell}_t(\theta, z) \leq q(z) \mid z] \approx \alpha$.

return "IN" if $\hat{\ell}_t(\theta, z^*) \leq q(z^*)$, otherwise "NO"

3 MI Attacks with Quantile Regression

We will now describe our new membership inference attacks. Under the setting in Sec. 2, we assume that the attacker has access to a set of public examples D drawn from the underlying distribution P. Given the public dataset D, a choice of t for the t-error function, and the trained diffusion model θ, the attacker learns a quantile regressor q such that $q(z)$ predicts the α-quantile of the t-error $\hat{\ell}_t(\theta, z)$ for each example z in D, where α is a parameter that controls the false-positive rate. Then on any target example z^*, the attacker declares the example is a member of the training set if and only if the t-error $\hat{\ell}_t(\theta, z^*) \leq q(z^*)$. The formal description of the algorithm is in Algorithm 1.

By design, our attack has a false-positive rate of α—that is the probability that that attacker incorrectly declares a randomly selected point z that was not used in training to have been used in training is α. By varying the parameter α, we can then trace the trade-off curves of *true-positive rates* at different false-positive rates. We will now describe two ways to learn the quantile regression model.

Quantile regression learner. First, a generic way to train a quantile regression model is to optimize over the pinball loss over some function class $\mathcal{Q}$ (e.g., neural networks). Formally, for any observed t-error $\hat{\ell}$ and quantile prediction q at a target level α, the pinball loss is defined as

$$L_\alpha(\hat{\ell}, q) = (q - \hat{\ell})(\mathbf{1}[\hat{\ell} \leq q] - \alpha) \tag{2}$$

Then we can find a quantile regression model $q(\cdot)$ that minimizes the *pinball loss*:

$$\min_{q \in \mathcal{Q}} \sum_{z \in D} L_\alpha(\hat{\ell}_t(\theta, z), q(z)) \tag{3}$$

The pinball loss is minimized by the function that predicts for each z the target α-quantile of the t-error conditioned on z. However, prior work [1] show that pinball loss tends to be a difficult loss function to minimize.

To complement the pinball loss minimization approach, we also consider a parametric approach, in which we learn a parametric distribution model to each example's t-error distribution. Based on the empirical evidence shown in Figure 1 in the appendix, we choose to model the distribution over the log of t-error with a Gaussian distribution. Thus, for each example z, our learner predicts the mean $\mu(z)$ and the standard deviation $\sigma(z)$ to fit the t-error distribution conditioned on z. This allows us to derive a quantile estimate from learned Gaussian distribution. Concretely, we will use the following log-likelihood objective to fit individual Gaussian distributions over the examples z in D:

$$\min_{\mu, \sigma} \sum_{z \in D} -\log\left(\frac{1}{\sqrt{2\pi}\sigma(z)} \exp\left(-\frac{(\mu(z) - \log(\hat{\ell}_t(\theta, z)))^2}{2\sigma(z)^2}\right)\right) \tag{4}$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ are given by neural networks.

Empirically, we observe that the log-likelihood objective is easier to optimize than the pinball loss function. Thus, for learning a model that predicts quantiles, we opt to use the parametric approach. We also observe that quantiles predicted from a model learned with the parametric approach gives lower pinball loss values on the holdout set than quantiles learned by directly minimizing pinball

loss do. It suggests that the pinball loss itself is still indicative of whether an MI attack would be successful. Better optimization techniques are needed for directly optimizing the pinball loss with a neural network as the base model.

4 Experimental Details

We demonstrate the effectiveness of our MI attack via quantile regression on four de-noising diffusion probabilistic models [9] (DDPMs) trained on CIFAR-10, CIFAR-100 [14], STL-10 [5] and Tiny-ImageNet, respectively. On each dataset, data samples are split into two halves, and one half is regarded as the private samples $\mathbf{Z}$ for training a DDPM. The other half is then split into two sets, including one as the public samples D that are auxiliary information, and the other as the holdout set for testing. On public samples, we train a quantile regression model using the parametric approach.

The base for our quantile regression is a ResNet model, and it is attached with two prediction heads, of which one is for the mean parameter and the other log of the standard deviation. Compared to the standard ResNet-18 model for classifying CIFAR-10, due to the simplicity of the score function, we reduce the number of channels in each layer by a factor of 4. In our experiments, a fixed $t = 50$ is used in the t-error function. [6] suggested that the choice of t does not influence the results drastically.

We adopt the same evaluation metric as prior work [1, 2]. Specifically, we are interested in the True Positive Rates (TPRs) at very low False Positive Rates (FPRs). Intuitively, a successful membership inference attack should identify true members with high accuracy, and in the meantime, make few mistakes on accusing nonmembers as members.

Comparison Partners. We mainly compare our MI attacks via quantile regression with two approaches. The first one is a simple **marginal** baseline, and for a target FPR value α, it computes the quantile on t-errors of the public samples, and then the performance of this marginal baseline is evaluated on the private samples and the holdout set. Thus, the marginal baseline only produces a single threshold for a target FPR, and it does not condition on the input images, whereas ours learns to predict the threshold for a given image, thus each images has a different threshold for a target FPR.

The other comparison partner is also a white-box attack using LiRA with gradient information [17], namely **GSA**. The LiRA attack formulates MI as hypothesis testing. Let θ denote the trained generative model, then the two competing hypotheses are $H_0 : \theta \sim A(\mathbf{Z}) \mid z^* \notin \mathbf{Z}$ and $H_1 : \theta \sim A(\mathbf{Z}) \mid z^* \in \mathbf{Z}$. These hypotheses correspond to whether or not the input dataset $\mathbf{Z}$ includes the target example z^*. Despite the simple formulation, estimating the two distributions requires training shadow models using random subsets from the same data domain, and in our case, each shadow model is a diffusion model, which may take days to train. The advantages of our attack is that, firstly, our hypothesis testing setup takes the condition on the target model, which makes our attack model-specific; secondly, our attack only requires learning a single model, which is much more computationally-efficient.

Results are presented in Table 1 and Table 2. Our attack on CIFAR10, besides being much more efficient, outperforms GSA attacks when we focus on lower TPR (0.1%). We also have demonstrated the effectiveness of our attack on diffusion models trained on other image datasets in Table 1 and 2. Besides the performance improvement over prior work, our algorithm is also computationally efficient and requires no knowledge about the training algorithm of the diffusion models.

Table 1: Performance of MI Attacks on CIFAR10 and CIFAR100.

Dataset	CIFAR-10		CIFAR-100	
MI Attack	TPR @ 1% FPR	TPR @ 0.1% FPR	TPR @ 1% FPR	TPR @ 0.1% FPR
Quantile (Single Model)	99.18%	98.956%	99.48%	99.26%
GSA_1 (Shadow Models)	99.70%	82.90%	-	-
GSA_2 (Shadow Models)	97.88%	58.57%	-	-
Marginal Baseline	9.6%	0.7%	11.06%	5.76%

Table 2: Performance of MI Attacks on Tiny-ImageNet and STL-10

Dataset	Tiny-ImageNet		STL-10	
MI Attack	TPR @ 1% FPR	TPR @ 0.1% FPR	TPR @ 1% FPR	TPR @ 0.1% FPR
Quantile (Single Models)	99.998%	99.998%	99.92%	99.85%
Marginal Baseline	8%	0.32%	5.78%	0.55%

References

[1] M. A. Bertran, S. Tang, A. Roth, M. Kearns, J. Morgenstern, and S. Wu. Scalable membership inference attacks via quantile regression. *pre-print*, 2023.

[2] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramèr. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1897–1914, 2022.

[3] N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson, A. Oprea, and C. Raffel. Extracting training data from large language models. *arXiv preprint arXiv:2012.07805*, 2020.

[4] N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson, A. Oprea, and C. Raffel. Extracting training data from large language models. *ArXiv*, abs/2012.07805, 2020.

[5] A. Coates, A. Ng, and H. Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011.

[6] J. Duan, F. Kong, S. Wang, X. Shi, and K. Xu. Are diffusion models vulnerable to membership inference attacks? In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 8717–8730. PMLR, 23–29 Jul 2023.

[7] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Conference on Theory of Cryptography*, TCC '06, pages 265–284, New York, NY, USA, 2006.

[8] J. Hayes, L. Melis, G. Danezis, and E. De Cristofaro. Logan: Membership inference attacks against generative models. *Proceedings on Privacy Enhancing Technologies*, 2019:133–152, 01 2019.

[9] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239*, 2020.

[10] N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS genetics*, 4(8):e1000167, 2008.

[11] M. Jagielski, J. Ullman, and A. Oprea. Auditing differentially private machine learning: How private is private sgd? In *Advances in Neural Information Processing Systems*, NeurIPS '20, 2020. `https://arxiv.org/abs/2006.07709`.

[12] B. Jayaraman and D. Evans. Evaluating differentially private machine learning in practice. In *USENIX Security Symposium*, 2019.

[13] B. Jayaraman, L. Wang, D. Evans, and Q. Gu. Revisiting membership inference under realistic assumptions. *arXiv preprint arXiv:2005.10881*, 2020.

[14] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[15] M. Nasr, J. Hayes, T. Steinke, B. Balle, F. Tramèr, M. Jagielski, N. Carlini, and A. Terzis. Tight auditing of differentially private machine learning. *CoRR*, abs/2302.07956, 2023.

[16] M. Nasr, S. Song, A. Thakurta, N. Papernot, and N. Carlini. Adversary instantiation: Lower bounds for differentially private machine learning. In *IEEE Symposium on Security & Privacy*, IEEE S&P '21, 2021. `https://arxiv.org/abs/2101.04535`.

[17] Y. Pang, T. Wang, X. Kang, M. Huai, and Y. Zhang. White-box membership inference attacks against diffusion models. *arXiv preprint arXiv:2308.06405*, 2023.

[18] R. Shokri, M. Stronati, C. Song, and V. Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy (S&P), Oakland*, 2017.

[19] Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

[20] T. Stadler, B. Oprisanu, and C. Troncoso. Synthetic data - anonymisation groundhog day. In K. R. B. Butler and K. Thomas, editors, *31st USENIX Security Symposium, USENIX Security 2022, Boston, MA, USA, August 10-12, 2022*, pages 1451–1468. USENIX Association, 2022.

[21] F. Tramèr, A. Terzis, T. Steinke, S. Song, M. Jagielski, and N. Carlini. Debugging differential privacy: A case study for privacy auditing. *CoRR*, abs/2202.12219, 2022.

[22] B. van Breugel, H. Sun, Z. Qian, and M. van der Schaar. Membership inference attacks against synthetic data through overfitting detection. In F. J. R. Ruiz, J. G. Dy, and J. van de Meent, editors, *International Conference on Artificial Intelligence and Statistics, 25-27 April 2023, Palau de Congressos, Valencia, Spain*, volume 206 of *Proceedings of Machine Learning Research*, pages 3493–3514. PMLR, 2023.

[23] L. Wang, B. Jayaraman, D. Evans, and Q. Gu. Efficient privacy-preserving nonconvex optimization. *arXiv preprint arXiv:1910.13659*, 2019.

[24] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *IEEE Computer Security Foundations Symposium*, CSF '18, pages 268–282, 2018. `https://arxiv.org/abs/1709.01604`.

A Diffusion Models

Before describing our attack, it is helpful to briefly describe how diffusion models work at a high level, following the notation of [9]. For a real image, a diffusion model provides a stochastic path from the image to noise. A diffusion model consists of two processes: (i) a T-step diffusion process (denoted as q below) that iteratively adds Gaussian noise to an image, and (ii) a denoising process (denoted as p below) that gradually reconstructs the image from noise.

Let z_0 be the real image without noise and z_T be the noisy image with the largest amount of noise. The transitions of diffusion and denoising are mathematically described as:

$$q(z_t|z_{t-1}) = \mathcal{N}(z_t; \sqrt{1-\beta_t}z_{t-1}, \beta_t I) \tag{5}$$

$$p_\theta(z_{t-1}|z_t) = \mathcal{N}(z_{t-1}; \mu_\theta(z_t, t), \Sigma_\theta(z_t, t)) \tag{6}$$

where $q_{z_t|z_{t-1}}$ is the probability distribution of the diffused image z_t given the previous image z_{t-1}, $p_\theta(z_{t-1}|z_t)$ is the probability distribution of the denoised image z_{t-1} given the noisy image z_t, $\mu_\theta(\cdot)$ and $\Sigma_\theta(\cdot)$ are the mean and covariance of the denoised image, respectively, as parameterized by the model parameters θ, β_t is a noise schedule that controls the amount of noise added at each step. Moreover, the marginal distribution at any time step t given the example z_0 can be written as

$$q(z_t \mid z_0) = \mathcal{N}(z_t; \sqrt{\overline{\alpha}_t}z_0, (1-\overline{\alpha}_t)I), \tag{7}$$

where $\alpha_t = 1 - \beta_t$ and $\overline{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. We work with the following re-parameterization of μ_θ with

$$\mu_\theta(z_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(z_t - \frac{\beta_t}{1-\overline{\alpha}_t}\epsilon_\theta(z_t, t)\right) \tag{8}$$

where ϵ_θ is a predictor (given by θ) that predicts the noise component given z_t.

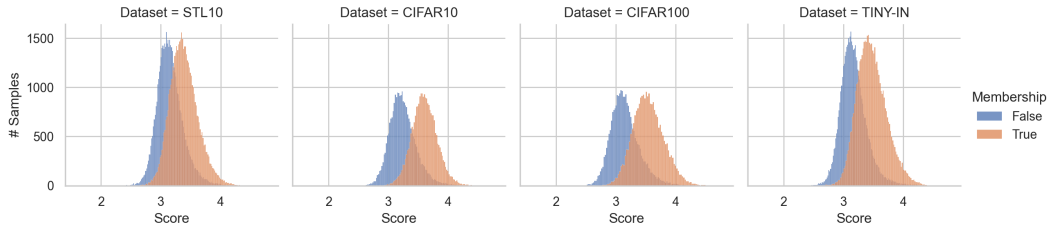

Figure 1: The distrbutions of the (negative) log transformation of the t-error on the private set and the public set of the dataset. It is clear that on each dataset, members and nonmembers have slightly different marginal score distributions, however, they are not drastically different from each other, which explains why the marginal baselines are not optimal, and also motivates our approach that conditions the score prediction on the input sample.

Many MI attacks proceed by identifying a loss function, and make membership inference by comparing the loss on the target example with a threshold. Intuitively, if the loss is unusually low, then there is evidence that the example was part of the training set. For supervised learning models, MI attacks typically leverage the classification loss (e.g., the cross-entropy loss). For diffusion models, existing work has proposed candidates of loss functions that measure the reconstruction error at different time steps of the diffusion process [4, 6]. We leveraged the t-error function defined in [6], which has the compelling advantage that it is deterministic and avoids repeated sampling from the diffusion process. Consider the following deterministic approximation of the diffusion and denoising processes:

$$z_{t+1} = \phi_\theta(z_t, t) = \sqrt{\overline{\alpha}_{t+1}} f_\theta(z_t, t) + \sqrt{1 - \overline{\alpha}_{t+1}} \epsilon_\theta(z_t, t) \tag{9}$$

$$z_{t-1} = \psi_\theta(z_t, t) = \sqrt{\overline{\alpha}_{t-1}} f_\theta(z_t, t) + \sqrt{1 - \overline{\alpha}_{t-1}} \epsilon_\theta(z_t, t) \tag{10}$$

where $f_\theta(z_t, t) = \frac{z_t - \sqrt{1-\overline{\alpha}_t}\epsilon_\theta(z_t, t)}{\sqrt{\overline{\alpha}_t}}$ is the estimate of z_0 given the z_t and the prediction $\epsilon_\theta(z_t, t)$. Then we could also define the deterministic reverse result as

$$\Phi(z_0, t) = \phi_\theta(\cdots \phi_\theta(\phi_\theta(z_0, 0), 1) \ldots, t-1) \tag{11}$$