ROBUST LEARNING OF DIFFUSION MODELS WITH EXTREMELY NOISY CONDITIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Conditional diffusion models have the generative controllability by incorporating external conditions. However, their performance significantly degrades with noisy conditions, such as corrupted labels in the image generation or unreliable observations or states in the control policy generation. This paper introduces a robust learning framework to address extremely noisy conditions in conditional diffusion models. We empirically demonstrate that existing noise-robust methods fail when the noise level is high. To overcome this, we propose learning pseudo conditions as surrogates for clean conditions and refining pseudo ones progressively via the technique of temporal ensembling. Additionally, we develop a Reverse-time Diffusion Condition (RDC) technique, which diffuses pseudo conditions to reinforce the *memorization effect* and further facilitate the refinement of the pseudo conditions. Experimentally, our approach achieves state-of-the-art performance across a range of noise levels on both class-conditional image generation and visuomotor policy generation tasks.

1 INTRODUCTION

Diffusion models (Ho et al., 2020; Song et al., 2021b; Karras et al., 2022; Chi et al., 2023) have significantly enhanced the generative tasks with success in multiple areas such as image generations (Kim et al., 2023; Singh & Raza, 2021; Song et al., 2021a; Karras et al., 2022; Rombach et al., 2022; Hatamizadeh et al., 2024), robotic control (Ma et al., 2024; Chi et al., 2023; Wang et al., 2023a; Li et al., 2024), and text generations (Li et al., 2022c; Gong et al., 2023; Wu et al., 2023; Gong et al., 2023) and with the continuous improvements on the generation efficiency (Song et al., 2021a; Shih et al., 2023). In particular, diffusion models gradually add random noise to the input x_0 through a forward stochastic process, resulting in increasingly noisy variants $x_t, t \in [0, T]$. A denoising model is then trained to reverse this process via score matching (Song et al., 2021b) to remove the random noise of x_t towards x_0 over multiple time steps t.

In particular, conditional diffusion models have introduced the controllability (Dhariwal & Nichol, 2021; Wang et al., 2023b; Huang et al., 2023) by adding various types of conditions to guide the generation (Ni et al., 2023; Luo et al., 2023; Zhang et al., 2023; Cao et al., 2024). For example, in the label-condition image generation (Rombach et al., 2022; You et al., 2023; Ifriqi et al., 2024), the generated images should closely match the class labels, and conditions are labels. In the generation of visuomotor policy (Chi et al., 2023; Na et al., 2024), the diffusion model can generate coherent actions based on visual observations by robots, and conditions are the visual observations (images) and the current robotic states. Specifically, conditional diffusion models embed the conditions into the learning objective of the denoising score matching across all time steps t and optimize the denoising model, given both the input x_0 and its corresponding conditions y.

However, the performing conditional diffusion models reply on high-quality conditions y that are often noisy in practice due to unreliable data sourcing (Jiang et al., 2022). Noisy conditions decrease the controllability of the generation that could cause performance degradation (as shown in Figure 1 (b)) and even hazards. For example, in image generation, incorrect labels can make the diffusion models generate the misleading images (Dufour et al., 2024); In the visuomotor policy generation, unreliable visual observation can lead to hazardous behaviors in real-world deployment, such as collapsing the entire robot system in critical applications of autonomous driving (Kahn et al.,

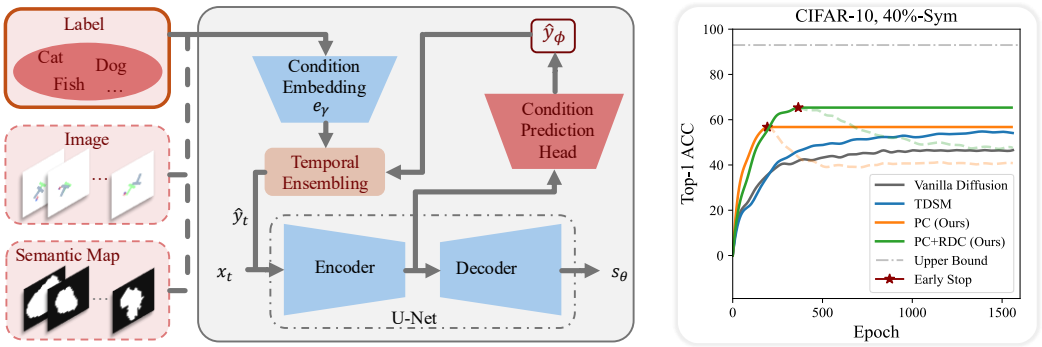


Figure 1: (a) Structures of our robust diffusion model: a lightweight prediction head that predicts pseudo conditions $\hat{y}$ is added at the output of the U-Net encoder of the diffusion model, and temporal ensembling is then adopted to update pseudo conditions. (b) Learning dynamics of conditional diffusion models on CIFAR-10 under 40% symmetric noise. Y-axis: controllability (Top-1 ACC of generated images, 1k images/class, 10 classes); X-axis: training epochs. Generations are evaluated using a pretrained CIFAR-10 classifier (Top-1 ACC 92.89%, silver dash-dot line). We compare the pseudo condition (PC) in orange curve and PC with Reverse-time Diffusion Condition (RDC) in green curve, both with early stopping (star markers), against TDSM (Na et al., 2024) in blue and the vanilla conditional diffusion (Karras et al., 2022) in gray curve.

2017; Zhu et al., 2024; Getahun & Karimoddini, 2024), robotic manipulations (Kalashnikov et al., 2018), and surgeries (Murphy & Alambeigi, 2023; Fan et al., 2024).

There is a pioneering study on the label-noise diffusion models for the image generation (Na et al., 2024), and we find it fails on the high noise level. Na et al. (2024) estimated a noisy condition transition matrix (Yao et al., 2020) that can establish a linear relationship between clean and noisy labels. Then, the obtained noisy condition transition matrix was leveraged to weigh the output of denoising model. Highly noisy conditions often create *entangled clusters* in demonstrations (Zhang et al., 2024), reducing feature consistency and hinder the diffusion model from learning the discriminative representations in early training. This is evident by the *underfitting* phenomenon of Vanilla Diffusion (Karras et al., 2022) (gray curve) and TDSM (Na et al., 2024) (blue curve) in Figure 1(b).

Targeting the problem of noisy distributions $p_0(x|\tilde{y})$, this paper introduces a pseudo condition $\hat{y}$ as a surrogate of the clean condition y. Under the classifier-free guidance framework (Ho & Salimans, 2022), we construct a lightweight prediction head that predicts pseudo conditions $\hat{y}$ of originally noisy counterparts $\tilde{y}$ (as shown in the Figure 1 (a)). During the optimization, the pseudo condition $\hat{y}$ can gradually refine itself and replace the noisy $\tilde{y}$, attributed to the memorization effect of fitting clean conditions before the noisy counterparts (Patrini et al., 2017; Han et al., 2018). Specifically, we update $\hat{y}$ using the technique of temporal ensembling (Laine & Aila, 2017) so that it gradually refines itself. Early stopping is then empirically applied to prevent $\hat{y}$ from overfitting to $\tilde{y}$, as shown by the translucent dashed orange/green curves in Figure 1 (b).

Furthermore, we propose a technique of Reverse-time Diffusion Condition (RDC) to transform the $\hat{y}$ to its diffusion process $\hat{y}_t$ ($t \in [0, T]$) that can further facilitate the refinement of conditions. Specifically, $\hat{y}_0$ represents a completely random condition, $\hat{y}_T$ represents the pseudo condition $\hat{y}$, and $\hat{y}_t$ ($t \in (0, T)$) represents $\hat{y}$ perturbed with a random noise at time t. Our robust learning objective becomes score matching between the input $(x_t, \hat{y}_t)$ and its target $(x_0, \hat{y}_T)$, and the U-Net model (Ronneberger et al., 2015) is then trained to fit this matching across various time steps $t \in [0, T]$. Specifically, RDC injects additional randomness into the already noisy conditions during the training and compels the optimization of the denoising model under these additional randomness that serves as a form of condition augmentation. In Figure 1 (b)that green line outperforms yellow line shows RDC can significantly enhance the memorization effect.

We have conducted experiments on two condition generation task: visuomotor policy generation (Chi et al., 2023) and image generation (Na et al., 2024). For visuomotor policy generation, we have conducted experiments on Push-T dataset (Florence et al., 2021) with the noisy conditions

when image observations are corrupted by camera distortion. For image generation, we have used both the CIFAR-10 and CIFAR-100 datasets (Krizhevsky et al., 2009) under two types of label-noise: symmetric noise (van Rooyen et al., 2015) and asymmetric noise (Patrini et al., 2017). All experiments corroborates our RDC-powered robust diffusion learning achieved the state-of-the-art (SOTA) performances across various levels of condition noises.

2 PRELIMINARY

2.1 SCORE MATCHING FOR DIFFUSION MODELS

Given a demonstration distribution $p_0(x)$, score matching (Hyvärinen, 2005) is to estimate the gradient of the log-density function without requiring explicit normalization of the probability distribution, i.e., $s_\theta(x) = \nabla_x \log p_0(x)$, where $x \in \mathbb{R}^d$, $\nabla_x \log p_t(x)$ is the score of x. Then score matching minimizes the expected squared difference between the estimated and score functions.

Song et al. (2021b) applied score matching for generation with transforming x to a diffusion process x_t with multiple score matching objectives at time $t \in [0, T]$. The distribution of x is first perturbed through a forward stochastic differential equation (SDE):

$$\mathrm{d}x = f(x, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}, \tag{1}$$

where $f(\cdot, t) : \mathbb{R}^d \to \mathbb{R}^d$ is the drift coefficient determining the perturbation direction of demonstration x over time t; $g(\cdot) : \mathbb{R} \to \mathbb{R}$ is the diffusion coefficient that controls the level of random noise applied to x at the time t; $\mathbf{w}$ is a standard Wiener process(Gelbrich & Römisch, 1995). In the widely adopted EDM (Karras et al., 2022) implementation, $f(\cdot, t) = 0$ and $g(\cdot) = \sqrt{2t}$. This SDE gradually transforms the distribution of x into a simple one (e.g., normal) as $t \to T$.

Then a reverse-time SDE is proposed to transform simple distribution back to the distribution of x:

$$\mathrm{d}x = \left[f(x, t) - g^2(t)\nabla_x \log p_t(x)\right] \mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}}, \tag{2}$$

where $\bar{\mathbf{w}}$ denotes another standard Wiener process (Gelbrich & Römisch, 1995).

The score $\nabla_x \log p_t(x)$ is approximated using a U-Net $s_\theta(x_t, t) : \mathcal{X} \times \mathcal{T} \to \mathcal{X}$, where $\mathcal{X} \subseteq \mathbb{R}^d$ denote the demonstration space and $\mathcal{T} \subseteq \mathbb{N}$ is the space of time steps. We minimize the objective of the denoising score matching :

$$\mathbb{E}_{t\sim\mathcal{U}(0,T)} \left[\lambda(t) \cdot \mathbb{E}_{x_t\sim p_t} \left\|s_\theta(x_t, t) - \nabla_{x_t} \log p_t(x_t)\right\|_2^2\right], \tag{3}$$

where $\mathcal{U}(0, T)$ denotes the uniform distribution over $[0, T]$, $\lambda(t)$ is a weighting function, p_0 is the demonstration x distribution, and $p_t(x_t)$ is the transition distribution defined by the forward SDE.

2.2 CLASSIFIER-FREE GUIDANCE (CFG)

CFG introduces the controllability into diffusion generation by modeling the conditional distribution $p_0(x \mid y)$ (Ho & Salimans, 2022). It allows conditional sampling without using an external classifier. Giving demonstration-condition pair (x,y), a conditional U-Net model $s_\theta(x_t, t, y) : \mathcal{X}\times\mathcal{T}\times\mathcal{Y} \to \mathcal{X}$, where $\mathcal{Y} \subseteq \mathbb{R}^d$ is the condition space, is trained to predict the conditional score of the demonstration x at time t given condition y. Besides that, an unconditional model $s_\theta(x_t, t, \varnothing)$ is also trained by randomly dropping y (i.e., setting as all-zero vector in the label condition) with a certain probability.

The conditional denoising score matching objective for the demonstration is $\mathrm{argmin}_\theta \mathcal{L}_{\mathrm{demo}}$, where

$$\mathcal{L}_{\mathrm{demo}} = \mathbb{E}_{t\sim\mathcal{U}(0,T)} \left[\lambda(t) \cdot \mathbb{E}_{x_t\sim p_t, y\sim p(y)} \left\|s_\theta(x_t, t, y) - \nabla_{x_t} \log p_t(x_t|y)\right\|_2^2\right]. \tag{4}$$

At sampling time, the conditional score prediction $\nabla_{x_t} \log p_{t|0}(x_t|x_0, y)$ is a weighted combination of conditional prediction and unconditional prediction:

$$s_{\mathrm{CFG}}(x_t, t, y) = s_\theta(x_t, t, \varnothing) + w \cdot (s_\theta(x_t, t, y) - s_\theta(x_t, t, \varnothing)), \tag{5}$$

where $w \geq 1$ is the guidance scale. Large w increases the controllability of the model with respect to the condition y. For consistency, all symbols appearing throughout this paper are listed in Table 4.

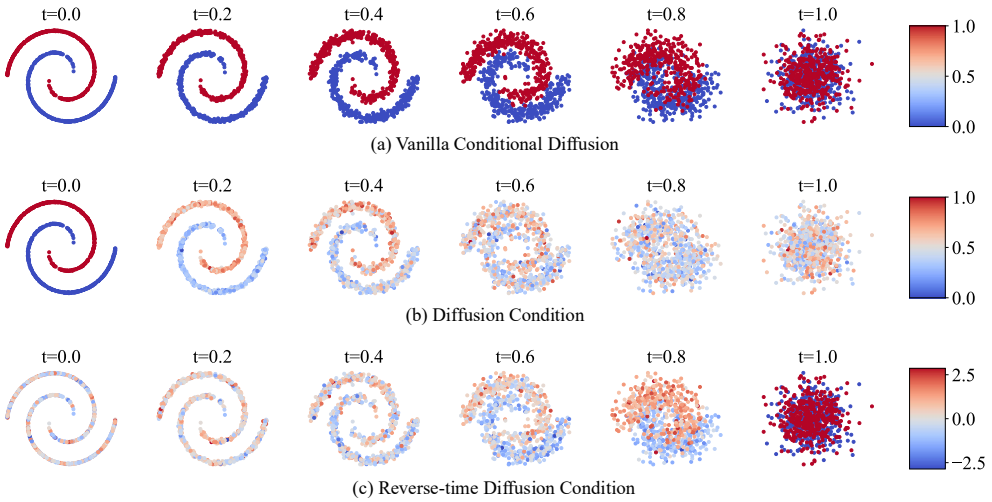


Figure 2: Visualization of three diffusion behaviors on the condition from t=0 to 1. (a) Standard conditional diffusion: y_t remains fixed. (b) Forward-diffused condition: y_t gradually becomes Gaussian. (c) Reverse-time diffusion condition (RDC): y_t evolves back toward y.

3 METHOD

3.1 CONDITIONAL DIFFUSION LEARNING UNDER NOISY CONDITIONS

3.1.1 PSEUDO CONDITION

The starting point of our method is the observation that, under noisy conditions, the conditional distribution $p_0(x|\tilde{y})$ becomes entangled, as each cluster of noisy condition $\tilde{y}$ corresponds to demonstration x originating from multiple clusters of clean conditions y. Such entanglement undermines feature consistency and makes it difficult for the diffusion model to learn valuable representations during early training. Consequently, vanilla conditional diffusion suffers from poor controllability, as shown in Figure 1(b). This raises the central motivation of our method: can noisy conditions be corrected by explicitly breaking their internal entanglement?

Our method replace this noisy clustered distribution $p_0(x|\tilde{y})$ with $p_0(x|\hat{y})$, where $\hat{y}$ is the proposed pseudo condition refined during training. As shown in Eq. 5, $s_{\text{CFG}}(x_t, t, \tilde{y})$ is switched into $s_{\text{CFG}}(x_t, t, \hat{y})$. Firstly, for each demonstration x_i in the dataset $\tilde{D}$, we construct and initialize $\hat{y}_i$ as an all-zero vector that has the same shape as $\tilde{y}_i$. Assigning the same initial pseudo-condition to all demonstrations effectively disrupts the entangled clusters in the original dataset. Then we update the pseudo condition $\hat{y}$ to approximate the clean condition y. As shown in the Figure 1 (a), we add a lightweight prediction head $q_\phi(x_t, t, \hat{y})$ at the output of the diffusion model U-Net encoder to obtain the refinement $\hat{y}_\phi = q_\phi(x_t, t, \hat{y})$ of the pseudo condition $\hat{y}$. Significantly, this additional prediction head incurs negligible additional computational overhead thanks to the parameter sharing in the U-Net encoder. We then optimize the refinement of the pseudo condition with the following learning objective $\text{argmin}_\phi \mathcal{L}_{\text{cond}}$:

$$\mathcal{L}_{\text{cond}} = \|\hat{y}_\phi - \tilde{y}\|^2. \tag{6}$$

In the early training stage, the memorization effect (Liu et al., 2020) enables demonstrations x with similar features are more likely to be clustered together, so that we can obtain a better pseudo condition $\hat{y}$ compared with the original noisy condition $\tilde{y}$ by updating $\hat{y}$ in the early training stage using the temporal ensembling(Laine & Aila, 2017):

$$\hat{y} = \alpha\hat{y} + (1-\alpha)\hat{y}_\phi, \tag{7}$$

where $0 \leq \alpha \leq 1$ is the momentum. Finally, we employ early stopping with an empirically determined criterion to prevent the clustered distribution from overfitting to the noisy distribution $p_0(x|\tilde{y})$. Please refer to Appendix 5 for more details.

3.1.2 REVERSE-TIME DIFFUSION CONDITION

Inspired by Kingma & Gao (2023), which view diffusion training as data augmentation through Gaussian noise of varying scales, we extend this idea by perturbing pseudo conditions within the diffusion process to further enhance their learning. As shown in Fig. 2, diffusing the condition along the demonstration direction (Fig. 2(a)) forces the model at $t = T$ to infer both demonstration and condition information from the fully corrupted (x_T, y_T), leading to large errors and unstable training. In contrast, our RDC (Fig. 2(b)) keeps $y_T = y$ while x_T becomes Gaussian, substantially reducing inference difficulty and stabilizing training.

Motivated by the need to stabilize training under noisy conditions, we transform pseudo condition $\hat{y}$ into a reverse-time diffusion process. First, we define the two critical state of $\hat{y}$ in the process, when $t = 0$, $\hat{y}_0 \sim \mathcal{N}(\mu, \sigma)$, ($\mu, \sigma$ are the specified mean and standard deviation), and when $t = T$, $\hat{y}_T = \hat{y}$. Then the reverse-time diffusion condition $\hat{y}_t$ is defined as:

$$\begin{cases} \text{Forward SDE: } \mathrm{d}\hat{y} = \dfrac{-f(\hat{y}, t)}{g(t)}\mathrm{d}t + \dfrac{1}{g(t)}\mathrm{d}\mathbf{w}, \\ \text{Reverse SDE: } \mathrm{d}\hat{y} = \left(\dfrac{-f(\hat{y}, t)}{g(t)} - \dfrac{1}{g(t)^2}\nabla_{\hat{y}} \log p_t(\hat{y})\right)\mathrm{d}t + \dfrac{1}{g(t)}\,\mathrm{d}\bar{\mathbf{w}}, \end{cases} \tag{8}$$

where $f(\cdot, t)$ and $g(t)$ have the same definition as Eq. 1, w and $\bar{w}$ are two standard Wiener process (Gelbrich & Römisch, 1995) as well, following the standard formulation of the SDEs, detailed derivation of RDC SDEs could be found at Appendix 4.1.

Correspondingly, we transform the prediction target of the condition prediction head from the refinement of the $\hat{y}_\phi$ into the condition score matching objective $s_\phi(x_t, t, \hat{y}_t)$, to learning the condition score function $\nabla_{\hat{y}} \log p_t(\hat{y})$ in the forward SDE. Then the prediction of pseudo condition $\hat{y}_\phi$ can be computed in the following integral form:

$$\hat{y}_\phi = \hat{y}_0 - \int_0^T \frac{s_\phi(\hat{y}_t, t)}{2t}\,\mathrm{d}t, \tag{9}$$

where the derivation for the integral form of $\hat{y}_\phi$ is given in the Appendix 4.2.

Next, we consider the optimization objective. For the demonstrations x, the objective $\mathcal{L}_{\text{demo}}$ remains unchanged as defined in Eq. 4. For the pseudo-condition $\hat{y}$, instead of following the optimization of x to estimate the condition score function $s_\phi(x_t, t, \hat{y}t)$, we directly optimize the denoised condition $\hat{y}_\phi$ in the Eq. 9 for the pseudo-condition update in the Eq. 7. Specifically, $\hat{y}_\phi$ is estimated using numerical methods (Gelbrich & Römisch, 1995), and the overall learning objective combines the conditional denoising score matching objective and the reverse-time diffusion condition learning objective $\text{argmin}_{\theta,\phi}\mathcal{L}$:

$$\mathcal{L} = \text{argmin}_\theta \mathcal{L}_{\text{demo}} + \mathbb{E}_{t\sim\mathcal{U}[0,T]}\|\hat{y}_\phi - \tilde{y}\|^2 \tag{10}$$

Note that $\hat{y}_\phi$ is obtained by integrating the learned condition score s_ϕ via the SDE-equivalent ODE (Eq. 9). Thus, the RDC loss remains a form of score matching rather than simple regression.

According to Kingma & Gao (2023), the RDC objective $\mathcal{L}_{\text{RDC}}$ is approximately equivalent to minimizing the standard conditional negative log-likelihood plus an explicit regularization term:

$$\mathcal{L}_{\text{RDC}} \approx \arg\min_\theta \left\{ -\mathbb{E}_{t,x_t}\left[\log p_\theta(x_t|\hat{y})\right] - \mathbb{E}_{t,x_t}\left[\frac{1}{2g(t)^2}\Delta t \cdot \text{Tr}(\nabla^2_{\hat{y}} \log p_\theta(x_t|\hat{y}))\right]\right\} \tag{11}$$

where the second term is the explicit time-annealed hessian regularization, with a strength inversely proportional to the square of the original diffusion coefficient. Derivation can be found in Appendix 4.3.

3.2 ALGORITHM

In this section, we extend our method to more challenging conditions such as images, where the condition embedding is implemented by a condition encoder $e_\gamma(\cdot)$ that maps noisy images to $\tilde{y}$. The main challenge is the joint optimization of the encoder e_γ and the diffusion U-Net s_θ.

The following adjustments are made to support the additional conditional encoder, with a stepwise description provided in Algorithm 1. Consistent with Section 3.1.1, we firstly construct the pseudo-condition by initializing $\hat{y}$ with the output embedding $\tilde{y}$ of the condition encoder, since uninformative embeddings at the start of training can effectively disrupt the entangled clusters present in the original dataset. A lightweight prediction head is also added to the encoder output of the U-Net to facilitate subsequent updates of the pseudo-condition. Then the pseudo-condition is converted into a RDC. we set the critical state when $t = T$ as $\hat{y}_T = \tilde{y}$, and refine the pseudo condition $\hat{y}$ with the same operations as Section 3.1 outlined. After early stopping, the condition encoder e_γ continues to refine its output embedding $\tilde{y}$ toward the learned pseudo-condition $\hat{y}$, maintaining alignment between the encoder representation and the updated condition. The learning objective of the condition encoder during the later training stage is $\mathrm{argmin}\gamma\mathcal{L}_{\mathrm{enc}}$, where $\mathcal{L}_{\mathrm{enc}} = \|\tilde{y} - \hat{y}\|^2$.

Algorithm 1 Robust Conditional Diffusion Learning with RDC

Input: Dataset $\tilde{\mathcal{D}} = \{(x, \tilde{y})\}$; condition encoder e_γ (optional); score network s_θ; predictor q_ϕ.
Output: Trained θ, and ϕ; (if e_γ used) trained γ.

if *using encoder* **then**
  initialize pseudo condition $\hat{y}$ with the output $\tilde{y}$ of the condition encoder e_γ;
  else initialize pseudo condition $\hat{y}$ with 0;
end
while *not early stopping* **do**
  sample $t \sim \mathcal{U}(0, T)$; $x_t \sim p_t(x_t)$; $y_t \sim p_t(y_t)$;
  update pseudo condition $\hat{y}_\phi \leftarrow q_\phi(x_t, t, y_t)$; $y_t \leftarrow (1-\alpha)y_t + \alpha\hat{y}_\phi$;
  update θ, ϕ with $\mathcal{L}_{\mathrm{cond}}$ (6) and $\mathcal{L}_{\mathrm{demo}}$ (4);
end
if *using encoder* **then**
  update γ with $\mathcal{L}_{\mathrm{enc}} = \|\tilde{y} - \hat{y}\|^2$;
  replace $\hat{y}_t$ with $\tilde{y}$;
end
else replace $\hat{y}_t$ with $\hat{y}$;
Update θ with $\mathcal{L}_{\mathrm{demo}}$ (Eq. 4);

4 EXPERIMENT

We evaluated our method on several tasks, including 2-D data generation (Section 4.1), image generation conditioned on labels (experiments on symmetric noise in Section 4.2, and experiments on asymmetric noise in Appendix 5.5), visuomotor policy generation conditioned on images (Section 4.3), and image generation conditioned on semantic maps (Section 5.6). Our approach achieves SOTA performance in most of tasks.

4.1 2-D TOY CASE

Figure 3 shows a toy example of 2-D synthetic data generation, which is to visualize the conditional generation performance of our method. The synthetic dataset consists of four classes with 2k 2-D synthetic data for each class. We then compare our method with the vanilla diffusion model EDM (Karras et al., 2022), and the SOTA robust conditional learning diffusion model, TDSM (Na et al., 2024), under 20%, 40%, 60%, and 80% symmetric noise. We use the x and y axes in all these sub-figures in the Figure 3 to represent the two feature dimensions of the 2-D data. See more experimental setup in the Appendix 5.

In 2-D generation tasks, the Mean Absolute Error (MAE) quantifies how closely the generated data match the corresponding training data (noise level 0%) across both coordinate axes. A smaller MAE indicates closer alignment between the generated data and the clean data.

The noise level increases from left to right in each row of Figure 3, and we can see that our method surpass TDSM (Na et al., 2024) at all noise level from 20% to 80%. At 80% noise, TDSM shows

no improvement over the baseline, while our method still reduces MAE by about 0.5, indicating that TDSM method based on noisy condition transition matrix estimation fails under high noise.

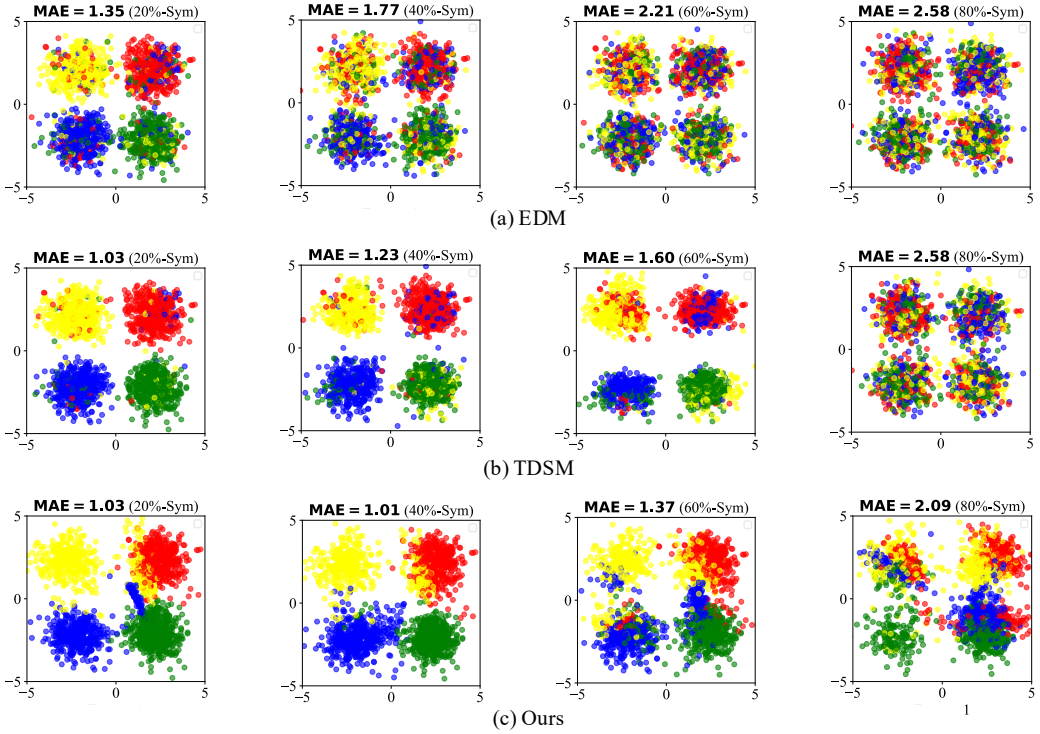


Figure 3: 2-D Toy Case. Comparison of robust condition learning methods using 2-D synthetic data in four class (red for class 1, blue for class 2, green for class 3, yellow for class 4). From left to right, three methods are trained and generated on 2-D data with symmetric noise, and the noise level equals 20%, 40%, 60%, and 80%.

4.2 Image Generation Conditioned on Labels

We conducted class-label image generation experiments on the noisy variants of CIFAR-10 and CIFAR-100 datasets, where the class-label is condition, and the image is the demonstration. In this task, we experimented with two types of label noise separately. For symmetric noise, class labels were uniformly mislabeled as any other class, and experiments were conducted at noise levels $\eta = 20\%$, 40%, 60%, and 80%. For asymmetric noise, class labels were flipped to similar classes, with experiments conducted at noise levels $\eta = 20\%$ and 40%. Metrics contain Fréchet Inception Distance (FID) (Heusel et al., 2017), Inception Score (IS) (Salimans et al., 2016), Density, Coverage (Naeem et al., 2020) and Class-Wise (CW) metrics (Chao et al., 2022) (evaluate the average value of the above metric separately for each class). The experimental settings almost the same with Karras et al. (2022). See more details of hyperparameter selection such as early stopping rule in Appendix 5.

Table 1 presents the experimental results under symmetric noise conditions. We compared our method with two models: the vanilla diffusion model EDM (Karras et al., 2022) and the SOTA robust image generation model TDSM (Na et al., 2024) on the CIFAR-10 and CIFAR-100 datasets. The results show that our method significantly outperforms the current SOTA model across all noise levels. Specifically, our method achieves more than 10% better performance than TDSM (Na et al., 2024) on all class-wise metrics. The advantage of our method remains notable at high noise levels. When the noise level exceeds 60%, our method maintains far better performance compared to other methods. Notably, TDSM (Na et al., 2024) often collapses during training on CIFAR-100 with symmetric noise levels above 60%, so we omit their results for this setting.

Table 1: Conditional Generation Performance Comparison with EDM and TDSM on the CIFAR-10 and CIFAR-100 datasets under Symmetric Noise.

Dataset	Noise Level	Method	FID (↓)	IS (↑)	Density (↑)	Coverage (↑)	CW-FID (↓)	CW-Density (↑)	CW-Coverage (↑)
CIFAR-10	0%	EDM	1.92	10.03	103.08	81.90	10.23	102.63	81.57
	20%	EDM	**2.00**	9.91	100.03	81.13	16.21	88.45	77.80
		TDSM	2.06	9.97	106.13	81.89	12.16	99.52	80.29
		Ours	2.02	**10.05**	**107.90**	**94.28**	**10.24**	**106.24**	**93.84**
	40%	EDM	**2.07**	9.83	100.94	80.93	30.45	73.02	71.63
		TDSM	2.43	9.96	**111.63**	82.03	15.92	97.80	78.65
		Ours	2.17	**10.04**	105.88	**93.80**	**10.64**	**102.96**	**93.18**
	60%	EDM	3.67	9.70	99.14	83.99	51.69	53.47	74.12
		TDSM	**3.22**	9.67	**100.19**	86.74	41.56	62.63	80.48
		Ours	3.23	**9.68**	99.33	**86.84**	**33.53**	**68.00**	**81.56**
	80%	EDM	5.84	9.45	99.36	61.73	79.42	38.40	51.18
		TDSM	5.47	**9.70**	102.52	61.96	78.98	39.91	53.80
		Ours	**4.25**	9.58	**103.53**	**78.73**	**68.39**	**47.35**	**56.70**
CIFAR-100	0%	EDM	2.51	12.80	87.98	77.63	66.97	82.58	75.78
	20%	EDM	**2.96**	12.28	83.01	75.02	79.91	66.47	70.11
		TDSM	4.26	12.29	85.66	74.90	78.71	70.62	70.77
		Ours	3.18	**12.95**	**98.32**	**93.52**	**71.57**	**90.49**	**91.51**
	40%	EDM	**3.36**	11.86	81.70	73.92	100.04	49.77	60.64
		TDSM	6.85	12.07	**88.45**	72.12	93.24	60.60	63.89
		Ours	4.60	**12.73**	84.75	**89.25**	**76.56**	**75.74**	**87.90**
	60%	EDM	7.07	**12.54**	**93.55**	83.53	117.75	42.92	74.37
		TDSM	-	-	-	-	-	-	-
		Ours	**5.57**	12.03	91.89	**87.45**	**104.34**	**67.39**	**84.03**
	80%	EDM	**11.13**	**12.66**	**92.09**	71.53	146.97	25.02	52.57
		TDSM	-	-	-	-	-	-	-
		Ours	11.50	10.94	83.08	**73.03**	**133.09**	**35.64**	**59.98**

For more experimental results on the asymmetric noise setting on both CIFAR-10 and CIFAR-100 datasets. Please refer to Appendix 5 for more details.

4.3 Visuomotor Policy Generation Conditioned on Images

We conducted visuomotor policy generation experiments conditioned on images from the noisy Push-T dataset (Florence et al., 2021). The task is to push a gray T-shaped block from a random position to a green target using image observations. To simulate condition noise, we apply two camera distortions: radial, which magnifies the image center, and tangential, which stretches regions due to camera misalignment. Distortions are applied with probability η within predefined intensity thresholds. Policies are evaluated via the Target Area Coverage (TAC) metric (Chi et al., 2023), measuring the IoU between the block and target. Results are averaged over three training seeds and 500 random environment initializations. More implementation details can be found in Appendix 5.

We compare our method with Diffusion Policy (DP) (Chi et al., 2023) as the baseline, and further introduce the SOTA image distortion correction method MOWA (Liao et al., 2025) into the image pre-processing stage of DP (Chi et al., 2023), which is called "MOWA + DP". The latter compari-

Table 2: TAC Comparison with DP and MOWA+DP on the Push-T datasets under camera distortions.

Method	Noise Level			
	20%	40%	60%	80%
DP	76.64±1.67	73.02±2.53	68.35±5.00	68.46±3.31
MOWA + DP	77.80±**0.58**	73.32±3.21	71.89±3.42	71.67±**2.09**
Ours	**80.26**±1.07	**73.44**±**1.42**	**72.74**±**1.21**	**71.78**±3.24

son is designed to examine how far simply applying advanced denoising can achieve robust policy learning from noisy visual observations.

As shown in Table 2, our method demonstrates significant improvements over the DP (Chi et al., 2023) and MOWA + DP method across noise levels from 20% to 80% on Push-T dataset (Florence et al., 2021). Moreover, compared to the two-stage paradigm "MOWA + DP", our end-to-end approach is more streamlined and computationally efficient.

Figure 4: Results on Push-T with 80% camera distortion. Each row shows one policy with nine key frames sampled at equal intervals.

To better illustrate our task, Figure 4 shows visuomotor policies trained on the Push-T dataset with 80% camera distortion noise. Each row shows one generated policy from a random initial state, with nine equally spaced frames representing the main steps. These results show that our model can still learn and perform complex tasks under extreme observation noise, highlighting its robustness.

4.4 Ablation Study

Table 3: Ablation results comparing vanilla diffusion (EDM) with(out) our PC and RDC modules on CIFAR-10 with 40% symmetric noise.

Method	CW-FID (↓)	CW-Density (↑)	CW-Coverage (↑)
EDM	30.45	73.02	71.63
Ours(PC)	37.09	54.04	64.38
Ours(PC+RDC)	**10.64**	**102.96**	**93.18**

Ablation Study on the Effectiveness of pseudo condition and RDC Mechanisms. To evaluate the contribution of the proposed pseudo condition and RDC method, we conducted a set of ablation experiments to compare three configurations: (1) the vanilla diffusion model, (2) the vanilla diffusion model with the proposed pseudo condition, and (3) the vanilla diffusion model with both the proposed pseudo condition and RDC mechanism. We share the same experimental setting across all three experiment, the results shown in the Table 3 demonstrate that inappropriate implementation of adding pseudo condition can even do harm to the vanilla diffusion model EDM (Karras et al., 2022), as shown by the orange curve in Figure 1 (b) during the later stages of training. After we add RDC on top of pseudo condition, we can see a notable conditional generation improvement compared with the vanilla diffusion model EDM (Karras et al., 2022). These findings validate the combined effectiveness of pseudo condition and RDC in improving the model's generation capabilities.

Conclusion

This paper addresses the problem of extremely noisy conditions in conditional diffusion models. We propose to learn pseudo conditions as surrogates for clean conditions and refine them progressively via the technique of temporal ensembling. Moreover, We improve pseudo condition learning using the RDC technique, which enhances memorization and facilitates the refinement of pseudo conditions through the reverse-time diffusion process. Experiments on both class-conditional image generation and visuomotor policy generation tasks shows that our method achieves SOTA performance across a range of noise levels.

REPRODUCIBILITY STATEMENT

We are committed to ensuring the reproducibility of our results. To this end, we outline the following key components we provide for replication:

- **Datasets:** Appendix 5.1 provides the detailed processing steps used to synthesize noisy condition from the publicly available datasets. Besides that, all datasets used in our experiments will be released after the blind-review phase.
- **Experimental Setup:** All hyperparameters, such as learning rate, batch size, optimizer type, and training schedule, are documented in Appendix 5.3.
- **Code:** The implementation of our method, including training/evaluation scripts and configuration files, will be released after the blind-review phase.

REFERENCES

Yingbin Bai, Erkun Yang, Bo Han, Yanhua Yang, Jiatong Li, Yinian Mao, Gang Niu, and Tongliang Liu. Understanding and improving early stopping for learning with noisy labels. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 24392–24403, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/cc7e2b878868cbae992d1fb743995d8f-Abstract.html.

Hanqun Cao, Cheng Tan, Zhangyang Gao, Yilun Xu, Guangyong Chen, Pheng-Ann Heng, and Stan Z. Li. A survey on generative diffusion models. *IEEE Trans. Knowl. Data Eng.*, 36(7): 2814–2830, 2024. doi: 10.1109/TKDE.2024.3361474. URL https://doi.org/10.1109/TKDE.2024.3361474.

Chen-Hao Chao, Wei-Fang Sun, Bo-Wun Cheng, Yi-Chen Lo, Chia-Che Chang, Yu-Lun Liu, Yu-Lin Chang, Chia-Ping Chen, and Chun-Yi Lee. Denoising likelihood score matching for conditional score-based data generation. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=LcF-EEt8cCC.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In Kostas E. Bekris, Kris Hauser, Sylvia L. Herbert, and Jingjin Yu (eds.), *Robotics: Science and Systems XIX, Daegu, Republic of Korea, July 10-14, 2023*, 2023. doi: 10.15607/RSS.2023.XIX.026. URL https://doi.org/10.15607/RSS.2023.XIX.026.

Noel C. F. Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen W. Dusza, David A. Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael A. Marchetti, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). *CoRR*, abs/1902.03368, 2019. URL http://arxiv.org/abs/1902.03368.

Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 8780–8794, 2021.

Nicolas Dufour, Victor Besnier, Vicky Kalogeiton, and David Picard. Don't drop your samples! coherence-aware training benefits conditional diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 6264–6273. IEEE, 2024. doi: 10.1109/CVPR52733.2024.00599. URL https://doi.org/10.1109/CVPR52733.2024.00599.

Ke Fan, Ziyang Chen, Giancarlo Ferrigno, and Elena De Momi. Learn from safe experience: Safe reinforcement learning for task automation of surgical robot. *IEEE Trans. Artif. Intell.*, 5(7): 3374–3383, 2024. doi: 10.1109/TAI.2024.3351797. URL https://doi.org/10.1109/TAI.2024.3351797.

Pete Florence, Corey Lynch, Andy Zeng, Oscar A. Ramirez, Ayzaan Wahid, Laura Downs, Adrian Wong, Johnny Lee, Igor Mordatch, and Jonathan Tompson. Implicit behavioral cloning. In Aleksandra Faust, David Hsu, and Gerhard Neumann (eds.), *Conference on Robot Learning, 8-11 November 2021, London, UK*, volume 164 of *Proceedings of Machine Learning Research*, pp. 158–168. PMLR, 2021. URL https://proceedings.mlr.press/v164/florence22a.html.

Fahimeh Fooladgar, Minh Nguyen Nhat To, Parvin Mousavi, and Purang Abolmaesumi. Manifold dividemix: A semi-supervised contrastive learning framework for severe label noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4012–4021, 2024.

Matthias Gelbrich and Werner Römisch. Numerical solution of stochastic differential equations (peter e. kloeden and eckhard platen). *SIAM Rev.*, 37(2):272–275, 1995. doi: 10.1137/1037073. URL https://doi.org/10.1137/1037073.

Tesfamichael Getahun and Ali Karimoddini. An integrated vision-based perception and control for lane keeping of autonomous vehicles. *IEEE Trans. Intell. Transp. Syst.*, 25(8):9001–9015, 2024. doi: 10.1109/TITS.2024.3376516. URL https://doi.org/10.1109/TITS.2024.3376516.

Aritra Ghosh, Himanshu Kumar, and P. S. Sastry. Robust loss functions under label noise for deep neural networks. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pp. 1919–1925. AAAI Press, 2017. doi: 10.1609/AAAI.V31I1.10894.

Jacob Goldberger and Ehud Ben-Reuven. Training deep neural-networks using a noise adaptation layer. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. Diffuseq: Sequence to sequence text generation with diffusion models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=jQj-_rLVXsj.

Xian-Jin Gui, Wei Wang, and Zhang-Hao Tian. Towards understanding deep learning from noisy labels with small-loss criterion. In Zhi-Hua Zhou (ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pp. 2469–2475. ijcai.org, 2021. doi: 10.24963/IJCAI.2021/340. URL https://doi.org/10.24963/ijcai.2021/340.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W. Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 8536–8546, 2018.

Ali Hatamizadeh, Jiaming Song, Guilin Liu, Jan Kautz, and Arash Vahdat. Diffit: Diffusion vision transformers for image generation. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part VIII*, volume 15066 of *Lecture Notes in Computer Science*, pp. 37–55. Springer, 2024. doi: 10.1007/978-3-031-73242-3_3. URL https://doi.org/10.1007/978-3-031-73242-3_3.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 6626–6637, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *CoRR*, abs/2207.12598, 2022. doi: 10.48550/ARXIV.2207.12598.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Han Huang, Leilei Sun, Bowen Du, and Weifeng Lv. Conditional diffusion based on discrete graph structures for molecular graph generation. In Brian Williams, Yiling Chen, and Jennifer Neville (eds.), *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pp. 4302–4311. AAAI Press, 2023. doi: 10.1609/AAAI.V37I4.25549. URL https://doi.org/10.1609/aaai.v37i4.25549.

Jinchi Huang, Lie Qu, Rongfei Jia, and Binqiang Zhao. O2u-net: A simple noisy label detection approach for deep neural networks. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 3325–3333. IEEE, 2019. doi: 10.1109/ICCV.2019.00342. URL https://doi.org/10.1109/ICCV.2019.00342.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.*, 6:695–709, 2005. URL https://jmlr.org/papers/v6/hyvarinen05a.html.

Tariq Berrada Ifriqi, Pietro Astolfi, Melissa Hall, Reyhane Askari Hemmat, Yohann Benchetrit, Marton Havasi, Matthew J. Muckley, Karteek Alahari, Adriana Romero-Soriano, Jakob Verbeek, and Michal Drozdzal. On improved conditioning mechanisms and pre-training strategies for diffusion models. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/18023809c155d6bbed27e443043cdebf-Abstract-Conference.html.

Liangxiao Jiang, Hao Zhang, Fangna Tao, and Chaoqun Li. Learning from crowds with multiple noisy label distribution propagation. *IEEE Trans. Neural Networks Learn. Syst.*, 33(11):6558–6568, 2022. doi: 10.1109/TNNLS.2021.3082496. URL https://doi.org/10.1109/TNNLS.2021.3082496.

Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2309–2318. PMLR, 2018.

Gregory Kahn, Adam Villaflor, Vitchyr Pong, Pieter Abbeel, and Sergey Levine. Uncertainty-aware reinforcement learning for collision avoidance. *CoRR*, abs/1702.01182, 2017. URL http://arxiv.org/abs/1702.01182.

Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, and Sergey Levine. Scalable deep reinforcement learning for vision-based robotic manipulation. In *2nd Annual Conference on Robot Learning, CoRL 2018, Zürich, Switzerland, 29-31 October 2018, Proceedings*, volume 87 of *Proceedings of Machine Learning Research*, pp. 651–673. PMLR, 2018. URL http://proceedings.mlr.press/v87/kalashnikov18a.html.

Takuhiro Kaneko, Yoshitaka Ushiku, and Tatsuya Harada. Label-noise robust generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 2467–2476. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00257.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

Jeongho Kim, Gyojung Gu, Minho Park, Sunghyun Park, and Jaegul Choo. Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. *CoRR*, abs/2312.01725, 2023. doi: 10.48550/ARXIV.2312.01725. URL https://doi.org/10.48550/arXiv.2312.01725.

Diederik P. Kingma and Ruiqi Gao. Understanding diffusion objectives as the ELBO with simple data augmentation. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/ce79fbf9baef726645bc2337abb0ade2-Abstract-Conference.html.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Seong Min Kye, Kwanghee Choi, Joonyoung Yi, and Buru Chang. Learning with noisy labels by efficient transition matrix estimation to combat label miscorrection. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXV*, volume 13685 of *Lecture Notes in Computer Science*, pp. 717–738. Springer, 2022. doi: 10.1007/978-3-031-19806-9_41. URL https://doi.org/10.1007/978-3-031-19806-9_41.

Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=BJ6oOfqge.

Haoran Li, Yaocheng Zhang, Haowei Wen, Yuanheng Zhu, and Dongbin Zhao. Stabilizing diffusion model for robotic control with dynamic programming and transition feasibility. *IEEE Trans. Artif. Intell.*, 5(9):4585–4594, 2024. doi: 10.1109/TAI.2024.3387401. URL https://doi.org/10.1109/TAI.2024.3387401.

Junnan Li, Richard Socher, and Steven C. H. Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

Shikun Li, Xiaobo Xia, Shiming Ge, and Tongliang Liu. Selective-supervised contrastive learning with noisy labels. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 316–325. IEEE, 2022a. doi: 10.1109/CVPR52688.2022.00041. URL https://doi.org/10.1109/CVPR52688.2022.00041.

Shikun Li, Xiaobo Xia, Hansong Zhang, Yibing Zhan, Shiming Ge, and Tongliang Liu. Estimating noise transition matrix with label correlations for noisy multi-label learning. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022b.

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B. Hashimoto. Diffusion-lm improves controllable text generation. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022c. URL http://papers.nips.cc/paper_files/paper/2022/hash/1be5bc25d50895ee656b8c2d9eb89d6a-Abstract-Conference.html.

Kang Liao, Zongsheng Yue, Zhonghua Wu, and Chen Change Loy. MOWA: multiple-in-one image warping model. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(9):7369–7382, 2025. doi: 10.1109/TPAMI.2025.3567465. URL https://doi.org/10.1109/TPAMI.2025.3567465.

Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Jianjie Luo, Yehao Li, Yingwei Pan, Ting Yao, Jianlin Feng, Hongyang Chao, and Tao Mei. Semantic-conditional diffusion networks for image captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 23359–23368. IEEE, 2023. doi: 10.1109/CVPR52729.2023.02237. URL https://doi.org/10.1109/CVPR52729.2023.02237.

Xiao Ma, Sumit Patidar, Iain Haughton, and Stephen James. Hierarchical diffusion policy for kinematics-aware multi-task robotic manipulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 18081–18090. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01712. URL https://doi.org/10.1109/CVPR52733.2024.01712.

Xingjun Ma, Yisen Wang, Michael E. Houle, Shuo Zhou, Sarah M. Erfani, Shu-Tao Xia, Sudanthi N. R. Wijewickrema, and James Bailey. Dimensionality-driven learning with noisy labels. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3361–3370. PMLR, 2018.

Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah M. Erfani, and James Bailey. Normalized loss functions for deep learning with noisy labels. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6543–6553. PMLR, 2020. URL http://proceedings.mlr.press/v119/ma20c.html.

Eran Malach and Shai Shalev-Shwartz. Decoupling "when to update" from "how to update". In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 960–970, 2017.

Takeru Miyato and Masanori Koyama. cgans with projection discriminator. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

Braden P. Murphy and Farshid Alambeigi. A surgical robotic framework for safe and autonomous data-driven learning and manipulation of an unknown deformable tissue with an integrated critical space. *J. Medical Robotics Res.*, 8(1&2):2340001:1–2340001:14, 2023. doi: 10.1142/S2424905X23400019. URL https://doi.org/10.1142/S2424905X23400019.

Byeonghu Na, Yeongmin Kim, HeeSun Bae, Jung Hyun Lee, Se Jung Kwon, Wanmo Kang, and Il-Chul Moon. Label-noise robust diffusion models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7176–7185. PMLR, 2020. URL http://proceedings.mlr.press/v119/naeem20a.html.

Haomiao Ni, Changhao Shi, Kai Li, Sharon X. Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 18444–18455. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01769. URL https://doi.org/10.1109/CVPR52729.2023.01769.

Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2642–2651. PMLR, 2017.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 2233–2241. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.240. URL https://doi.org/10.1109/CVPR.2017.240.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4331–4340. PMLR, 2018.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 10674–10685. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01042.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells III, and Alejandro F. Frangi (eds.), *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*, volume 9351 of *Lecture Notes in Computer Science*, pp. 234–241. Springer, 2015. doi: 10.1007/978-3-319-24574-4_28. URL https://doi.org/10.1007/978-3-319-24574-4_28.

Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 2226–2234, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/8a3363abe792db2d8761d6403605aeb7-Abstract.html.

Andy Shih, Suneel Belkhale, Stefano Ermon, Dorsa Sadigh, and Nima Anari. Parallel sampling of diffusion models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/0d1986a61e30e5fa408c81216a616e20-Abstract-Conference.html.

Nripendra Kumar Singh and Khalid Raza. Medical image generation using generative adversarial networks: A review. In *Health Informatics*, pp. 77–96. 2021. doi: 10.1007/978-981-15-9735-0_5.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021a.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021b.

Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. Joint optimization framework for learning with noisy labels. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 5552–5560. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00582.

Feilong Tang, Zhongxing Xu, Qiming Huang, Jinfeng Wang, Xianxu Hou, Jionglong Su, and Jingxin Liu. Duat: Dual-aggregation transformer network for medical image segmentation. In Qingshan Liu, Hanzi Wang, Zhanyu Ma, Weishi Zheng, Hongbin Zha, Xilin Chen, Liang Wang, and Rongrong Ji (eds.), *Pattern Recognition and Computer Vision - 6th Chinese Conference, PRCV 2023, Xiamen, China, October 13-15, 2023, Proceedings, Part V*, volume 14429 of *Lecture Notes in Computer Science*, pp. 343–356. Springer, 2023. doi: 10.1007/978-981-99-8469-5_27. URL https://doi.org/10.1007/978-981-99-8469-5_27.

Ryutaro Tanno, Ardavan Saeedi, Swami Sankaranarayanan, Daniel C. Alexander, and Nathan Silberman. Learning from noisy labels by regularized estimation of annotator confusion. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 11244–11253. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.01150.

Kiran Koshy Thekumparampil, Ashish Khetan, Zinan Lin, and Sewoong Oh. Robustness of conditional gans to noisy labels. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 10292–10303, 2018.

Brendan van Rooyen, Aditya Krishna Menon, and Robert C. Williamson. Learning with symmetric label noise: The importance of being unhinged. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 10–18, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/45c48cce2e2d7fbdea1afc51c7c6ad26-Abstract.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Tsun-Hsuan Johnson Wang, Juntian Zheng, Pingchuan Ma, Yilun Du, Byungchul Kim, Andrew Spielberg, Joshua B. Tenenbaum, Chuang Gan, and Daniela Rus. Diffusebot: Breeding soft robots with physics-augmented generative diffusion models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023a. URL http://papers.nips.cc/paper_files/paper/2023/hash/8b1008098947ad59144c18a78337f937-Abstract-Conference.html.

Zhizhong Wang, Lei Zhao, and Wei Xing. Stylediffusion: Controllable disentangled style transfer via diffusion models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 7643–7655. IEEE, 2023b. doi: 10.1109/ICCV51070.2023.00706. URL https://doi.org/10.1109/ICCV51070.2023.00706.

Qi Wei, Haoliang Sun, Xiankai Lu, and Yilong Yin. Self-filtering: A noise-aware sample selection for label noise with confidence penalization. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXX*, volume 13690 of *Lecture Notes in Computer Science*, pp. 516–532. Springer, 2022. doi: 10.1007/978-3-031-20056-4_30. URL https://doi.org/10.1007/978-3-031-20056-4_30.

Tong Wu, Zhihao Fan, Xiao Liu, Hai-Tao Zheng, Yeyun Gong, Yelong Shen, Jian Jiao, Juntao Li, Zhongyu Wei, Jian Guo, Nan Duan, and Weizhu Chen. Ar-diffusion: Auto-regressive diffusion model for text generation. In Alice Oh, Tristan Naumann, Amir

Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/7d866abba506e5a56335e4644ebe18f9-Abstract-Conference.html.

Xiaobo Xia, Tongliang Liu, Nannan Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. Are anchor points really indispensable in label-noise learning? In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 6835–6846, 2019.

Xiaobo Xia, Tongliang Liu, Bo Han, Nannan Wang, Mingming Gong, Haifeng Liu, Gang Niu, Dacheng Tao, and Masashi Sugiyama. Part-dependent label noise: Towards instance-dependent label noise. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Xiaobo Xia, Tongliang Liu, Bo Han, Chen Gong, Nannan Wang, Zongyuan Ge, and Yi Chang. Robust early-learning: Hindering the memorization of noisy labels. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=Eql5b1_hTE4.

Fan Yang, Kai Wu, Shuyi Zhang, Guannan Jiang, Yong Liu, Feng Zheng, Wei Zhang, Chengjie Wang, and Long Zeng. Class-aware contrastive semi-supervised learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 14401–14410. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01402. URL https://doi.org/10.1109/CVPR52688.2022.01402.

Yu Yao, Tongliang Liu, Bo Han, Mingming Gong, Jiankang Deng, Gang Niu, and Masashi Sugiyama. Dual T: reducing estimation error for transition matrix in label-noise learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Xichen Ye, Yifan Wu, Weizhong Zhang, Xiaoqiang Li, Yifan Chen, and Cheng Jin. Optimized gradient clipping for noisy label learning. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pp. 9463–9471. AAAI Press, 2025. doi: 10.1609/AAAI.V39I9.33025. URL https://doi.org/10.1609/aaai.v39i9.33025.

Zebin You, Yong Zhong, Fan Bao, Jiacheng Sun, Chongxuan Li, and Jun Zhu. Diffusion models and semi-supervised learners benefit mutually with few labels. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/8735753cc18f6baa92d1f069fd8b14a0-Abstract-Conference.html.

Jingfeng Zhang, Bo Song, Haohan Wang, Bo Han, Tongliang Liu, Lei Liu, and Masashi Sugiyama. Badlabel: A robust perspective on evaluating and enhancing label-noise learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(6):4398–4409, 2024. doi: 10.1109/TPAMI.2024.3355425. URL https://doi.org/10.1109/TPAMI.2024.3355425.

Yivan Zhang, Gang Niu, and Masashi Sugiyama. Learning noise transition matrix from only noisy labels via total variation regularization. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12501–12512. PMLR, 2021. URL http://proceedings.mlr.press/v139/zhang21n.html.

Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 10146–10156. IEEE, 2023. doi: 10.1109/CVPR52729.2023.00978. URL https://doi.org/10.1109/CVPR52729.2023.00978.

Zhilu Zhang and Mert R. Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 8792–8802, 2018.

Guoqing Zheng, Ahmed Hassan Awadallah, and Susan T. Dumais. Meta label correction for noisy label learning. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pp. 11053–11061. AAAI Press, 2021. doi: 10.1609/AAAI.V35I12.17319.

Xiong Zhou, Xianming Liu, Chenyang Wang, Deming Zhai, Junjun Jiang, and Xiangyang Ji. Learning with noisy labels via sparse regularization. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 72–81. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00014. URL https://doi.org/10.1109/ICCV48922.2021.00014.

Zeyu Zhu, Shuai Wang, and Huijing Zhao. Uncertainty-aware deep imitation learning and deployment for autonomous navigation through crowded intersections. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2024, Abu Dhabi, United Arab Emirates, October 14-18, 2024*, pp. 13980–13987. IEEE, 2024. doi: 10.1109/IROS58592.2024.10801536. URL https://doi.org/10.1109/IROS58592.2024.10801536.

APPENDIX

1 LLM USAGE STATEMENT

We confirm that LLMs were used only for language polishment of this manuscript. They did not contribute to the research ideation and/or writing to the extent.

2 RELATED WORK

2.1 CONDITION-NOISE LEARNING

Research on noisy-condition has been extensive, and prior works have approached this problem from multiple directions, including data-centric strategies that filter or correct noisy conditions, model-based approaches that explicitly characterize noise distributions, and optimization techniques that introduce regularization or noise-robust training objectives.

From the data perspective, sample selection or reweighting methods (Han et al., 2018; Jiang et al., 2018; Ren et al., 2018; Li et al., 2020; 2022a) mitigate the impact of condition noise by selecting training demonstrations with clean conditions or reducing the weights of noisy demonstrations. Empirical studies on the memorization effect (Malach & Shalev-Shwartz, 2017; Liu et al., 2020) show that DNNs tend to first learn simple and clean conditions before gradually overfitting to all noisy ones. Many methods exploit this phenomenon to filter out demonstrations with clean conditions that are easier to fit (Gui et al., 2021; Wei et al., 2022). However, determining the proper timing of early learning is challenging (Xia et al., 2021; Bai et al., 2021), and non-stop iterative selection or reweighting may accumulate errors.

To address this limitation, condition correction methods (Tanaka et al., 2018; Ma et al., 2018; Liu et al., 2020; Zheng et al., 2021) adopt a gentler approach that considers all demonstrations and iteratively refines noisy conditions with soft/hard network predictions. Many of these methods leverage semi-supervised learning and contrastive learning, treating noisy demonstrations as unconditional ones while using classification predictions as pseudo conditions (Yang et al., 2022; Fooladgar et al., 2024).

From the model perspective, methods based on the noisy condition transition matrix aim to learn the mapping between the distributions of clean and noisy conditions by estimating the noisy condition transition matrix (Goldberger & Ben-Reuven, 2017; Zhang et al., 2021; Li et al., 2022b; Kye et al., 2022). However, most of these methods follow the impractical assumption of instance-independent

noise (Xia et al., 2020; Kye et al., 2022), and they usually fail under extremely noisy conditions due to large estimation errors of the transition matrix (Yao et al., 2020).

From the optimization perspective, noise-robust regularization prevents model parameters from overfitting noisy conditions by adding regularization terms to the loss function (Ghosh et al., 2017; Zhang & Sabuncu, 2018; Zhou et al., 2021; Ye et al., 2025). Beyond typical regularization techniques such as dropout and data augmentation, more advanced approaches have been proposed to handle higher levels of noise (Tanno et al., 2019; Xia et al., 2019). The main challenge of this line of work lies in the delicate design of optimization strategies (Huang et al., 2019; Ma et al., 2020).

2.2 Condition-noise Learning in Generation

In generation tasks, conditional information or priors (such as class labels or text descriptions) are often introduced to enable controllable generation (Odena et al., 2017; Miyato & Koyama, 2018), which also brings challenges related to condition noise. Mismatched demonstration-condition pairs can degrade the quality of conditional generation, motivating research on robust methods.

One line of work focuses on conditional GANs. RCGAN (Thekumparampil et al., 2018) considers both known (RCGAN) and unknown (RCGAN-U) noise distributions. RCGAN introduces a known noise channel in the generator's output conditions and combines it with a projection discriminator to enhance adaptability to noise. RCGAN-U jointly optimizes the generator and a dynamically learned noise model (confusion matrix) to approximate the true conditional distribution under unknown noise. Similarly, Kaneko et al. (2019) proposed AC-GAN (with an auxiliary classifier) and cGAN (without any auxiliary classifier), incorporating a noise transition matrix into the classifiers and discriminators and adding noise-robust regularization to the loss functions. While effective, these methods struggle under high noise levels, and computationally expensive techniques like mutual information regularization can limit efficiency on high-dimensional or large-scale datasets.

Conditional diffusion models extend this idea by incorporating extra conditional information to guide generation. Meanwhile, robust learning for conditional diffusion is less explored. Na et al. (2024) represent the conditional score of noisy conditions as a linear combination of clean condition scores, weighted by instance-specific and time-dependent condition transition matrices. By minimizing the distance between the weighted score network output and the noisy data score, the diffusion model is guided to produce outputs closer to clean conditions.

Another recent study (Dufour et al., 2024) introduces a coherence score to represent the consistency between condition and demonstration, incorporating it into the diffusion model to dynamically adjust reliance on conditions. It is important to note that this method is based on a different setting: it assumes the availability of additional robust classifiers, while our work addresses robust conditional diffusion pre-training without requiring any auxiliary condition information (clean conditions in the entire training phase).

3 Notation

We summarize the main symbols used throughout the paper in Table 4 for better readability.

4 Derivation of formulas

4.1 RDC SDEs

Based on the standard forward and reverse-time SDEs in diffusion introduced in Section 2.1 (Eq. 1 and Eq. 2), we construct a reverse-time diffusion process for the pseudo condition $\hat{y}$ that shares the same schedule but in opposite direction as x. We let x become $\hat{y}$, set $f(\hat{y}, t) = f(x, t)$ and keep $g(t)$ unchanged, and exchange the start and end states, where the start and end states correspond to a Gaussian noise and $\hat{y}$, respectively.

From this, we can write the forward SDE of the reverse-time diffusion condition as:

$$\mathrm{d}\mathbf{w} = f(\hat{y}, t)\mathrm{d}t + g(t)\mathrm{d}\hat{y}. \tag{12}$$

By rearranging the terms, we obtain our forward SDE for RDC:

$$\text{Forward SDE: } \mathrm{d}\hat{y} = \frac{-f(\hat{y}, t)}{g(t)}\mathrm{d}t + \frac{1}{g(t)}\mathrm{d}\mathbf{w}. \tag{13}$$

Table 4: Summary of all notations.

Symbol	Meaning
$x \in \mathbb{R}^d$	Demonstration
$y \in \mathbb{R}^d$	Clean condition
$\tilde{y} \in \mathbb{R}^d$	Noisy condition
$\hat{y} \in \mathbb{R}^d$	Pseudo condition (refined from $\tilde{y}$)
$\widetilde{D} = \{(x^{(i)}, \tilde{y}^{(i)})\}_{i=1}^n$	Noisy training dataset
$p_0(x)$	Distribution of demonstrations x
$p_0(x\|y)$	Clean conditional distribution
$\tilde{p}_0(x, \tilde{y})$	Noisy joint distribution
$p_t(x_t)$	Transition distribution of x at time t
$s_\theta(x_t, t)$	Score network (U-Net) w.r.t. x_t and t
$s_\theta(x_t, t, y)$	Conditional score network given condition y
$s_{\text{CFG}}(x_t, t, y)$	Classifier-Free Guidance score
$s_\phi(x_t, t, \hat{y}_t)$	Condition score network for pseudo condition
$q_\phi(y\|x)$	Auxiliary network inferring pseudo condition
$e_\gamma(\cdot)$	Condition encoder (for complex conditions such as images)
$f(x, t)$	Drift coefficient in SDE
$g(t)$	Diffusion coefficient in SDE
$\mathbf{w}, \bar{\mathbf{w}}$	Standard Wiener processes
T	Diffusion time horizon
$\lambda(t)$	Weighting function for score matching loss
w	Guidance scale in CFG

In this forward SDE, the drift and diffusion coefficients are:

$$f'(\hat{y}, t) = \frac{-f(\hat{y}, t)}{g(t)}, \quad g'(t) = \frac{1}{g(t)}.$$

Finally, substituting these redefined coefficients into the standard reverse SDE (Eq. 2) yields the reverse SDE for RDC:

$$\text{Reverse SDE: } \mathrm{d}\hat{y} = \left(\frac{-f(\hat{y}, t)}{g(t)} - \frac{1}{g(t)^2} \nabla_{\hat{y}} \log p_t(\hat{y}) \right) \mathrm{d}t + \frac{1}{g(t)} \mathrm{d}\bar{\mathbf{w}}. \tag{14}$$

4.2 INTEGRAL FORM OF $\hat{y}_\phi$ (EQ. 9)

We start from the reverse-time SDE given in Eq. 8:

$$\mathrm{d}\hat{y}_t = \left(-\frac{f(\hat{y}_t, t)}{g(t)} - \frac{1}{g(t)^2} \nabla_{\hat{y}} \log p_t(\hat{y}_t) \right) \mathrm{d}t + \frac{1}{g(t)} \mathrm{d}\bar{\mathbf{w}}_t. \tag{15}$$

To obtain a deterministic probability-flow ODE we drop the Brownian term $\mathrm{d}\bar{\mathbf{w}}_t$ and replace the score by a neural network $s_\phi(\hat{y}_t, t) \approx \nabla_{\hat{y}} \log p_t(\hat{y}_t)$:

$$\frac{\mathrm{d}\hat{y}_t}{\mathrm{d}t} = -\frac{f(\hat{y}_t, t)}{g(t)} - \frac{s_\phi(\hat{y}_t, t)}{g(t)^2}. \tag{16}$$

As mentioned before, our RDC shares the same schedule but in opposite direction with the demonstration x, following Karras et al. (2022), both use the implementation of $f(\cdot, t) = 0$ and $g(t) = \sqrt{2t}$. The drift simplifies to

$$\frac{\mathrm{d}\hat{y}_t}{\mathrm{d}t} = -\frac{s_\phi(\hat{y}_t, t)}{2t}. \tag{17}$$

Integrating from 0 to T yields the final integral form of $\hat{y}_\phi$

$$\hat{y}_\phi = \hat{y}_0 - \int_0^T \frac{s_\phi(\hat{y}_t, t)}{2t}\,\mathrm{d}t, \tag{18}$$

where $\hat{y}_0 \sim p_T(\cdot)$ is the initial noise. In practice the integral is approximated by a numerical solver such as Euler or Heun's method.

4.3 Derivation of Time-Annealed Condition Regularization for RDC

In this section, we provides the detailed derivation of the explicit Hessian regularization term, where the RDC condition noise is generated by a process inspired by the reverse-time SDE.

Given forward SDE (Eq.8) of RDC, the RDC loss $\mathcal{L}_{\text{RDC}}$ is theoretically equivalent to minimizing the negative expected log-likelihood(Kingma & Gao, 2023):

$$\mathcal{L}_{\text{RDC}} \approx \arg\min_\theta \left\{ \mathbb{E}_{t,x_t} \left[-\mathbb{E}_{\hat{y}_t \sim q(\cdot|\hat{y})} \left[\log p_\theta(x_t|\hat{y}_t) \right] \right] \right\} \tag{19}$$

To simplify notations in Eq.8, we define a new drift term $F(\hat{y}, t) = \frac{-f(\hat{y},t)}{g(t)}$ and a new diffusion coefficient $G(t) = \frac{1}{g(t)}$. Then we consider the infinitesimal change $\Delta\hat{y} = F(\hat{y}, t)\Delta t + G(t)\Delta\mathbf{w}$ over a small time step Δt. We apply a second-order Taylor expansion to $\log p(x_t|\hat{y} + \Delta\hat{y})$ around $\hat{y}$:

$$\log p(x_t|\hat{y} + \Delta\hat{y}) \approx \log p(x_t|\hat{y}) + (\Delta\hat{y})^T \nabla_{\hat{y}} \log p(x_t|\hat{y}) + \frac{1}{2}(\Delta\hat{y})^T \mathbf{H}_{\hat{y}} (\Delta\hat{y}) \tag{20}$$

where $\mathbf{H}_{\hat{y}} = \nabla^2_{\hat{y}} \log p(x_t|\hat{y})$. To calculate the negative expectation with respect to the SDE increment $\Delta\hat{y}$, we will consider each item separately:

First term This term is simply the conditional log-likelihood objective evaluated at the unperturbed condition $\hat{y}$: $-\log p(x_t|\hat{y})$.

Second term The expectation of this term contributes a first-order correction term:

$$-\mathbb{E}[(\Delta\hat{y})^T] \nabla_{\hat{y}} \log p \approx -\left(F(\hat{y})\Delta t\right)^T \nabla_{\hat{y}} \log p \tag{21}$$

This term is a first-order correction resulting from the SDE's drift $F(\hat{y}, t)$. We conceptually group this term with the zeroth-order term ($-\log p(x_t|\hat{y})$) because they together constitute the primary conditional diffusion objective.

Third term This term provides the explicit Hessian regularization. We calculate the negative expectation of the second-order Taylor term: $-\frac{1}{2}\mathbb{E}[(\Delta\hat{y})^T \mathbf{H}_{\hat{y}} (\Delta\hat{y})]$.

To compute the expectation of this quadratic form, we use the trace identity $\mathbb{E}[(\Delta\hat{y})^T \mathbf{H} (\Delta\hat{y})] = \mathrm{Tr}(\mathbf{H} \cdot \mathrm{Cov}(\Delta\hat{y}))$. The covariance of the SDE increment is determined by the diffusion coefficient: $\mathrm{Cov}(\Delta\hat{y}) = G(t)^2 \mathbf{I}\Delta t = \frac{1}{g(t)^2}\mathbf{I}\Delta t$. Applying the trace identity yields:

$$\begin{aligned}
-\frac{1}{2}\mathbb{E}[(\Delta\hat{y})^T \mathbf{H}_{\hat{y}} (\Delta\hat{y})] &\approx -\frac{1}{2}\mathrm{Tr}(\mathbf{H}_{\hat{y}} \cdot \mathrm{Cov}(\Delta\hat{y})) \\
&= -\frac{1}{2}\mathrm{Tr}\left(\mathbf{H}_{\hat{y}} \cdot \frac{1}{g(t)^2}\mathbf{I}\Delta t \right) \\
&= -\frac{1}{2g(t)^2}\Delta t \cdot \mathrm{Tr}(\mathbf{H}_{\hat{y}})
\end{aligned}$$

We can finally define the full RDC objective as:

$$\mathcal{L}_{\text{RDC}} \approx \arg\min_\theta \left\{ -\mathbb{E}_{t,x_t} \left[\log p_\theta(x_t|\hat{y}) \right] + \mathbb{E}_{t,x_t} \left[\frac{1}{2g(t)^2}\Delta t \cdot \mathrm{Tr}(\nabla^2_{\hat{y}} \log p_\theta(x_t|\hat{y})) \right] \right\}, \tag{22}$$

where the second term is a time-annealed regularization penalizing the sensitivity of the conditional log-likelihood to condition noise(sum of second-order derivatives over all condition dimensions).

5 More Details of Experiments

5.1 Noisy Condition Setting

1) **Visuomotor Policy Generation**

We conduct experiments on this task with Push-T dataset. Push-T is an object pushing task for robots: The robotic arm needs to push a gray T-shaped block at a random position to the green target position based on the image observations. This dataset contains 206 video clips (25650 frames).

In this task, the image observation is the condition, radial distortion and tangential distortion are two typical forms of lens distortion for image data, which are widely exist in the actual imaging process. We simulate these two types of image noise with the following two expressions. Set (x, y) as the coordinates of the image.

- Radial Distortion. Radial distortion is mainly caused by the non ideal characteristics of the lens optical components, resulting in varying degrees of distortion of the image edges relative to the center, presenting as barrel or pillow distortion.

$$\begin{cases} x' = x(1 + k_1 r^2 + k_2 r^4), \\ y' = y(1 + k_1 r^2 + k_2 r^4), \end{cases} \tag{23}$$

  where k_1 and k_2 are the first- and second-order radial distortion coefficients that control the distortion intensity. In close-up tasks such as robot grasping of objects, we set $k_1 \in [0,\ 0.2]$ and $k_2 = 0$ to simulate barrel distortion during the close-up process.

- Tangential Distortion. Tangential distortion is caused by improper installation between the lens and imaging sensor, manifested as a deviation of the image in a certain direction.

$$\begin{cases} x' = x + 2p_1 xy + p_2(r^2 + 2x^2), \\ y' = y + 2p_1(r^2 + 2y^2) + p_2 xy, \end{cases} \tag{24}$$

  where p_1 controls the distortion along the x-direction, and p_1 controls distortion along the y-direction. We set $p_1 \in [0,\ 0.1]$ and $p_2 \in [0,\ 0.1]$.

  We further introduce spatial noise to simulate minor geometric disturbances of the camera sensor position. Specifically, we perturb the coordinates by adding independent pixel-wise offsets to the horizontal and vertical directions:

$$\begin{cases} x' = x + \Delta x, \\ y' = y + \Delta y, \end{cases} \tag{25}$$

  where Δx and Δy represent the added noise in the x and y directions, respectively. These values are randomly sampled within the ranges $\Delta x \in [0,\ 10]$ and $\Delta y \in [0,\ 10]$.

As shown in Figure 5, we visualize several examples from the Push-T dataset before and after the application of the above noise. Each pair of images illustrates how camera distortion and spatial perturbations affect the image observation of the same scene. These visualizations help demonstrate challenges in robotic perception under the close-range operation.

Figure 5: Visualization examples of noisy Push-T dataset. Each subfigure shows the effect of condition noise, where the left image is the original image observations and the right image is the corresponding version with added camera distortion.

2) **Image Generation**

We conduct image generation experiments on both CIFAR-10 and CIFAR-100 datasets. CIFAR-10 consists of 60000 32x32 pixel color images, divided into 10 classes, with each class containing 6000 images. These images cover classes such as airplanes, cars, birds, cats, deer, dogs, frogs, horses, boats, and trucks. CIFAR-10 is suitable for label-condition image generation task. CIFAR-100 is similar to CIFAR-10, but contains 100 classes, each class containing 600 images, for a total of 60000 images. These classes include various fine-grained objects, such as different types of fish, insects, flowers, etc. In our setting, CIFAR-100 is used to evaluate the robust condition generation performance of the model when dealing with more complex categories.

In the image generation task, the class-label is the condition. In order to better cope with label uncertainty in real-world scenarios, we introduced symmetric noise and asymmetric noise into both the CIFAR-10 and CIFAR-100 dataset.

- Symmetric Noise: symmetric noise refers to the situation where the label of each class is mislabeled as another class with the same probability η, simulating a random, unbiased label error scenario.
- Asymmetric Noise: asymmetric noise refers to the tendency of certain classes to be mistakenly labeled as other specific classes based on their similarity and other characteristics. For example, images of "cats" are mistakenly labeled as "dogs" instead of random other classes. This type of noise can better reflect label confusion caused by visual similarity and other factors in actual scenes. For CIFAR-10, labels are flipped by truck $\leftrightarrow$ automobile, bird $\leftrightarrow$ airplane, deer $\leftrightarrow$ horse, cat $\leftrightarrow$ dog. For CIFAR-100, classes are randomly flipped within the same superclass.

5.2 Details of Prediction Head

In the label-conditioned image generation task, the prediction head consists of two convolutional layers with batch normalization and a residual connection, followed by a global average pooling layer and a final fully connected linear classifier.

In the image-condition visuomotor policy generation task, the prediction head processes the 1D image embedding via a 1D convolution followed by temporal pooling and flattening.

5.3 Details of Experimental Setup

2-D Data Generation For the 2-D toy case, a 1-D U-Net is adopted as the score matching network. The model is trained with a batch size of 512 for a total of 10,000 iterations. For the early stopping, we set $0 \sim 500$ iterations. All other hyperparameters are kept consistent across different noise levels. For the sampling process, we use deterministic sampling with Heun 2^{nd}-order integrator; noise schedule $\sigma(t) = t$, signal scaling $s(t) = 1$; step-density exponent $\rho = 7$; 35 Number of Function Evaluations (NFE); $\sigma_{\text{min}} = 0.002$, $\sigma_{\text{max}} = 80$ (Karras et al., 2022).

Conditional Image Generation In this section, we utilized 8 NVIDIA Tesla 4090 GPUs and employed CUDA 11.8 and PyTorch 2.0 versions in our experiments. Our model framework and code are based on EDM (Karras et al., 2022). For all experiments, we used DDPM++ network architecture with a U-net backbone, which is originally proposed by Song et al. (2021b) and modified by Karras et al. (2022). The training setting is the same with Karras et al. (2022). Besides that, we set the α in the Eq. 7 as 0.1. For the early stopping, we set $0 \sim 25,000$ iterations for CIFAR-10 and $0 \sim 30,000$ iterations for CIFAR-100 under the symmetric noise setting.

For the sampling process, we use deterministic sampling with Heun 2^{nd}-order integrator; noise schedule $\sigma(t) = t$, signal scaling $s(t) = 1$; step-density exponent $\rho = 7$; 35 NFE; $\sigma_{\text{min}} = 0.002$, $\sigma_{\text{max}} = 80$, which is exactly the setting of Karras et al. (2022). Besides that, we apply the following metrics to evaluate our method:

- **FID (↓)** evaluates the quality of generated images by computing the Fréchet distance between the feature distributions of generated and given reference images. Smaller FID indicates higher generation quality.
- **IS (↑)** measures the quality and diversity of the generated image by using a pretrained Inception model to predict the generated image, and calculating the KL divergence of the predicted distribution of the generated image.

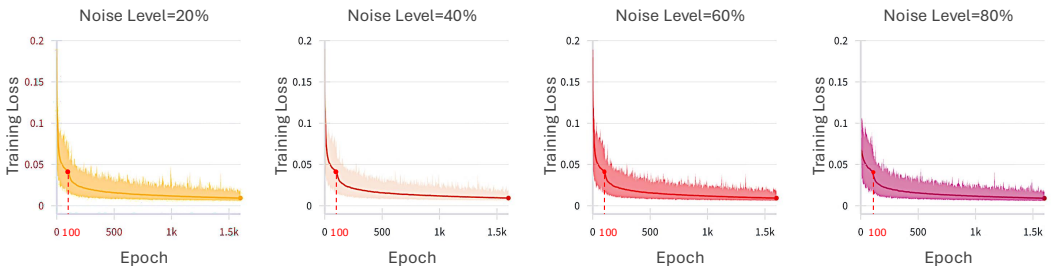


Figure 6: Visuomotor policy training loss over epochs under 20%–80% camera distortion. The loss drops sharply before entering a plateau, corresponding to the transition from fitting the clean observation to memorizing noisy one. Early stopping is applied at the beginning of this plateau.

- **Density** (↑) reflects the distribution density of generated images in the feature space, revealing the concentration and coverage of generation. High density value reveals a concentrated distribution in the feature space, with better consistency and representativeness.
- **Coverage** (↑) measures the extent a generative model covers the distribution of the given reference distribution. High coverage value indicates that the model can comprehensively cover the distribution of the given reference distribution and generated images are more diverse and representative.
- **CW (Class-wise) Metrics** compute the average value of the metric within each class. For example, **CW-FID** (↓) reflects the generation quality of the model on each class, and **CW-Density** (↑) and **CW-Coverage** (↑) evaluate the distribution density and coverage within each class, respectively.

Visuomotor Policy Generation In this section, all experiments are conducted on 3*A100 GPUs using the AdamW optimizer. The learning rate is set to 1e-4, with a weight decay rate of 1e-6. The model is trained for 3000 epochs with a batch size of 64, and we repeat our experiments on the random seeds of 42, 43, and 44. Target Area Coverage (TAC) metric is used to measure the quality of visuomotor policy, which is to compute the IoU between the T-shaped block and the green target area after all the operation of the robotic arm. Besides that, we set the α in the Eq. 7 as 0.2, and the early stopping is set as $0 \sim 100$ epochs.

5.4 Analysis of Early Stopping for Pseudo-Condition Learning

Early stopping is applied to prevent pseudo-conditions from overfitting to noisy observations. Due to the memorization effect, models first fit clean condition samples (easy) and later overfit noisy ones (hard), producing a loss curve that drops sharply before plateauing (Figure **??**). We freeze pseudo-condition updates once the loss stabilizes; temporal ensembling further smooths the updates, making the diffusion model largely insensitive to the exact stopping point.

Table 5 reports visuomotor policy generation performance (TAC) under 80% camera distortion with different early stopping epochs. The results are stable across a range of epochs, demonstrating the robustness of this procedure.

Table 5: Visuomotor policy generation performance of our method with different early stopping epochs under 80% camera distortion.

Early Stopping Epoch	TAC
80	72.55 ± 1.78
100	71.78 ± 3.24
120	72.65 ± 1.49

5.5 Image Generation Conditioned on Labels with Asymmetric Noise

In this section, we tested the label-condition image generation performance of our method under a more challenging setting, which is to simulate condition confusion between similar classes. Specifically, we introduced asymmetric noise of 20% and 40% on both the CIFAR-10 and CIFAR-100 datasets. Here, we define the similar classes the same with Na et al. (2024), which is clearly stated in Appendix 5.1

For all the experiments under asymmetric noise, we follow the training and sampling settings of Karras et al. (2022) as well. Besides that, we set the α in the Eq. 7 as 0.3 for CIFAR-10 and 0.5 for CIFAR-10. For the early stopping, we set $0 \sim 25,000$ iterations for CIFAR-10 and CIFAR-100.

As shown in the Table 6, even though under the case of 40% asymmetric noise on CIFAR-100, our method demonstrates comparable performance to the baseline method, the performance of our method far exceeds that of SOTA method TDSM in all other settings.

Table 6: Conditional Generation Performance Comparison with EDM and TDSM on the CIFAR-10 and CIFAR-100 datasets under Asymmetric Noise.

Dataset	Noise Level	Method	FID (↓)	IS (↑)	Density (↑)	Coverage (↑)	CW-FID (↓)	CW-Density (↑)	CW-Coverage (↑)
CIFAR-10	0%	EDM	1.92	10.03	103.08	81.90	10.23	102.63	81.57
	20%	EDM	**2.02**	10.06	100.66	81.36	11.97	96.10	79.95
		TDSM	**1.95**	10.04	104.15	81.81	10.89	101.77	80.99
		Ours	2.17	**9.97**	**105.59**	**93.72**	**6.8**	**103.31**	**93.32**
	40%	EDM	**2.23**	10.09	101.25	81.10	15.18	92.13	78.12
		TDSM	**2.06**	10.02	**105.19**	81.90	12.54	99.21	79.98
		Ours	2.14	**10.02**	103.15	**93.02**	**12.47**	**99.25**	**91.79**
CIFAR-100	0%	EDM	2.51	12.80	87.98	77.63	66.97	82.58	75.78
	20%	EDM	**2.76**	12.49	87.36	77.04	75.39	33.31	72.14
		TDSM	2.64	12.79	88.41	77.46	**69.83**	78.92	74.01
		Ours	3.34	**12.92**	**98.43**	**91.98**	66.89	**84.95**	**90.57**
	40%	EDM	**2.73**	12.51	**87.06**	**76.56**	89.13	60.27	64.19
		TDSM	2.81	**12.57**	87.01	76.27	**73.13**	**74.30**	**71.48**
		Ours	2.83	12.51	86.74	75.88	89.34	59.69	63.33

5.6 Image Generation Conditioned on Semantic Maps

We evaluate our method on semantic synthesis (Rombach et al., 2022) to explore its potential for augmenting medical image semantic datasets. In medical imaging, precise mask annotations are crucial, as boundary noise from ambiguous pathological regions can mislead model training and even contribute to misdiagnosis.

To simulate such noisy condition, we perturb the mask annotations of the ISIC 2018 (Codella et al., 2019) dataset with a two-pixel erosion or dilation, introducing varying levels of boundary noise. Our generation model is trained on these noisy masks and later tested with clean masks as input. For comparison, we additionally use DuAT (Tang et al., 2023), a advanced medical image mask prediction method with 86% mIoU accuracy on the ISIC 2018 (Codella et al., 2019) dataset, to provide realistic predicted masks for evaluation.

We assess both the photorealism and the mask consistency of the generated results using mean IoU (mIoU) and mean Dice coefficient (mDice). Specifically, mIoU measures the average overlap between predicted and clean mask regions across classes, while mDice captures the harmonic similarity with greater sensitivity to boundary alignment.

Table 7 summarizes the quantitative results, showing that our current implementation underperforms the baseline methods. This indicates that the model in its current form is not yet capable of effectively handling boundary noise. To better understand this behavior, we further plot the Mask Fitting Speed Comparison during Training (Figure 7). The figure reveals that, in the early stages of training, our method exhibits a desirable effect: the pseudo-condition aligns more closely with the clean mask than with the noisy mask. However, the model quickly overfits to the noisy annotations, and this rapid convergence prevents us from leveraging the early-stage advantage for effective correction.

Table 7: Mean IoU Comparison with EDM on the ISIC 2018 datasets under Boundary Noise.

Noise Level	Method	mIoU	mDice
0%	EDM	**86.56**	**92.67**
20%	EDM	**84.82**	**91.62**
	Ours	84.20	91.24
40%	EDM	**83.86**	**91.03**
	Ours	83.21	90.63
60%	EDM	82.14	89.96
	Ours	**82.94**	**90.47**
80%	EDM	**80.73**	**89.07**
	Ours	80.11	88.68

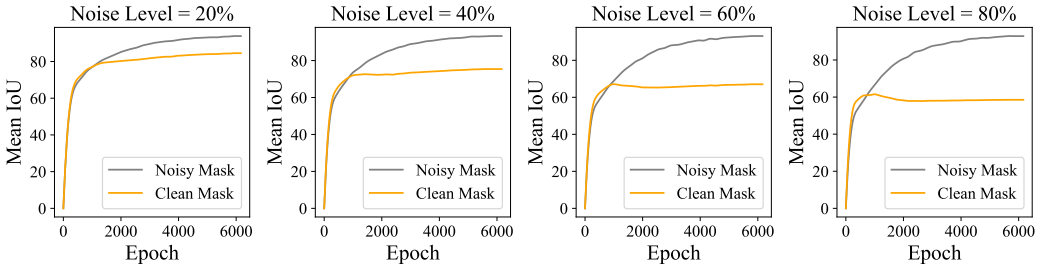


Figure 7: Mask Fitting Speed Comparison during Training. Each subplot shows the mean IoU between our predicted Pseudo Condition and the observed noisy mask (orange line) or the unobserved clean mask (gray line) over training epochs. From left to right, the subplots correspond to different noise levels from 20% to 80%.

5.7 Visuomotor Policy Generation Conditioned on Image with Gaussian Noise

In this section, we provide additional experiments to evaluate the robustness of our method under more realistic noise conditions. Specifically, we inject extreme (60% and 80%) Gaussian noise into the image observations. As shown in Table 8, our method maintains strong performance under both noise levels, highlighting its robustness when handling gaussian noise images.

Table 8: TAC Comparison with DP on the Push-T dataset under gaussian noise.

Method	Noise Level	
	60%	80%
DP	80.32±2.66	79.58±1.51
Ours	**81.47±1.36**	**81.76±0.62**

5.8 Image Generation Conditioned on Label under Transformer Backbone

In this section, we further examine the generality of our method by replacing the UNet backbone with a Transformer-based architecture. For the Transformer backbone, we strictly followed the implementation of DP (Vaswani et al., 2017). The design of the condition prediction head remains unchanged from the implementation described in Appendix 5.2. This experiment evaluate whether our method remains effective under a different diffusion model backbone. As shown in Table 9, our method consistently outperforms DP when conditioned on 60% and 80% image observation distortion, demonstrating that the improvements of our method are not tied to a specific network architecture.

Table 9: TAC Comparison with DP on the Push-T dataset under Transformer Backbone.

Method	Noise Level	
	60%	80%
DP	60.45±2.70	60.54±**1.94**
Ours	**62.72**±**0.87**	**62.19**±3.47

5.9 Visualization Results on Label-condition Image Generation

To demonstrate the superiority of our method more intuitively, under the CIFAR-10 training setting of 40% symmetric noise, we randomly sampled a noise and generated it under exactly the same conditions using the current SOTA method TDSM (Na et al., 2024) and our method, and visually compared the generation effects of the two methods. As shown in the Figure 8, we randomly present two example results. It can be seen that under 40% symmetric noise, conditional image generation using TDSM (Na et al., 2024) sometimes fails to generate certain classes of images and produces some unrecognizable content, while our method can robustly generate images that meet the requirements of each class.

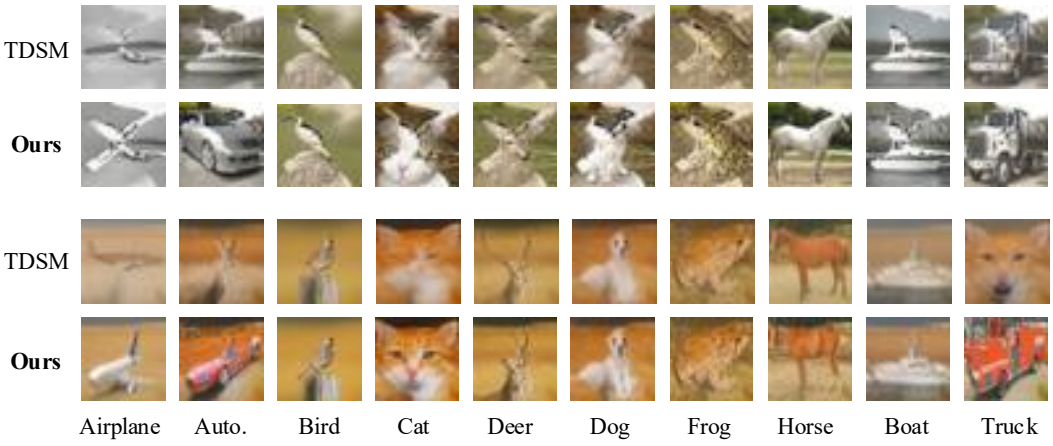
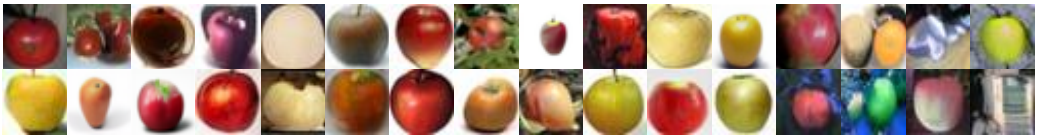
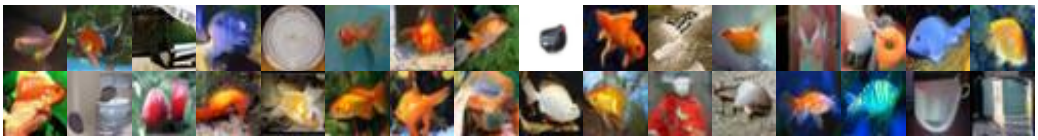


Figure 8: Visualization examples on CIFAR-10 compared with TDSM under 40% symmetric noise, where "auto." is short for "automobile".

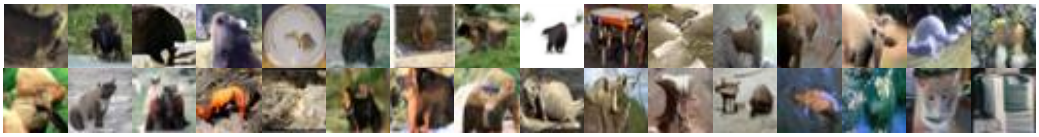
Besides the above results, we provide more visualization results of our method on CIFAR-100 under 60% symmetric noise. As shown in Figure 9, we trained our model under 60% symmetric noise training set, and then sampled seven different typical classes (apple, goldfish, bear, bed, beetle, bicycle, and bottle) starting from the same 32 random noise images. This figure shows the label-conditioned image generation performance of our method under different conditions given the same initial state. Figure 9 shows that our method can maintain robust conditional generation performance even under extreme 60% symmetric noise.



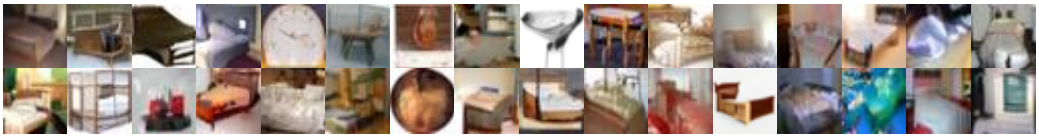
(a) Apple



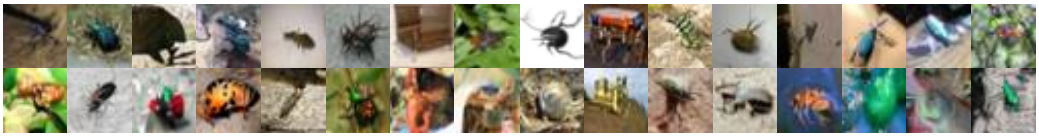
(b) Aquarium Fish



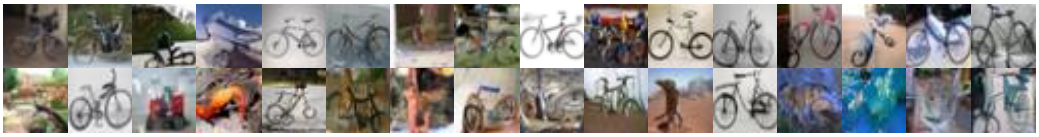
(c) Bear



(d) Bed



(e) Beetle



(f) Bicycle

(g) Bottle

Figure 9: Visualization examples of our method on CIFAR-100 under 60% symmetric noise.