TOWARDS EFFICIENT DIFFUSION TRAINING WITH NOISY LABELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Label noise can disrupt the correspondence between images and class labels, degrading conditional alignment and generation quality in diffusion models. Transition-aware weighted Denoising Score Matching (TDSM) addresses this problem by estimating instance- and time-dependent clean-label posterior probabilities using a label transition matrix and a classifier, and using these probabilities to aggregate class-conditional denoising predictions. However, this mechanism requires a separate denoiser evaluation for each class involved in the aggregation, incurring substantial training overhead when the number of classes is large. To reduce this cost, we propose mixed-label conditioning, which combines learnable class embeddings according to the estimated clean-label posterior probabilities and conditions the denoiser on the resulting mixed representation. By encoding label uncertainty directly in the conditioning input, our method requires only a single forward pass for each mixed-label prediction, avoiding explicit class-wise prediction and output aggregation. To further improve training efficiency and effectiveness, we introduce mixed diffusion-model-guided sample selection, which compares the diffusion losses of mixed-label and reference predictions to select training examples that benefit from conditioning at the sampled noise level. Experiments on CIFAR-10 and CIFAR-100 under noisy-label settings evaluate generation quality, conditional alignment, and computational cost. The results demonstrate generation performance comparable to that of TDSM with substantially reduced training overhead.

1 INTRODUCTION

Diffusion models have achieved strong performance in image generation through learned denoising processes (Ho et al., 2020; Song et al., 2021; Karras et al., 2022). Class conditioning further enables generation from specified semantic categories by learning the correspondence between training images and their annotations. However, label noise can disrupt this correspondence, degrading conditional alignment and generation quality (Na et al., 2024). For example, training on dog images incorrectly annotated as cats can encourage the model to associate dog-related visual content with the cat condition, contaminating the learned cat-conditional distribution and the generated images may fail to match the requested class. Unlike discriminative classification, which focuses on predicting class labels, conditional generation aims to learn the image distribution associated with each clean class. Addressing label noise in this setting therefore requires mitigating the effects of unreliable annotations while retaining useful visual information from mislabeled images.

Learning with noisy labels has been extensively studied in discriminative tasks through strategies such as sample selection and re-weighting (Jiang et al., 2018; Han et al., 2018) and robust loss functions (Zhang & Sabuncu, 2018; Wang et al., 2019). These approaches primarily target classification accuracy rather than the recovery of clean class-conditional image distributions. For conditional generation, excluding an incorrectly labeled example can also discard useful information about variation within its underlying clean class, creating a trade-off between supervision reliability and data utilization. Moreover, diffusion training introduces varying Gaussian perturbations, so the visual evidence supporting a class assignment changes with the noise level (Na et al., 2024). An image that is readily identifiable under a weak perturbation may become ambiguous under stronger noise. The direct transfer of discriminative noise-handling strategies to conditional diffusion is therefore not straightforward, motivating further investigation of this comparatively less explored setting.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Transition-aware weighted Denoising Score Matching (TDSM) (Na et al., 2024) addresses label noise by explicitly aggregating class-conditional denoiser predictions. For each perturbed training image, these predictions are combined using instance- and time-dependent weights estimated from a label transition matrix and a classifier. The weights represent the posterior probabilities of the underlying clean classes given the perturbed image, observed annotation, and diffusion time. The transition matrix accounts for label corruption, while the classifier provides image-dependent evidence about class identity. The weighted combination models the noisy-label conditional score and is matched to the denoising score-matching target. When the weights are exact and the required invertibility assumptions hold, the population optimum recovers the clean class-conditional scores. However, this mechanism requires evaluating the denoiser under individual class conditions before aggregation. For a C-class problem, full aggregation requires C denoiser evaluations of the same perturbed image, making computation scale with the number of classes and incurring substantial training overhead when C is large. This motivates a more efficient way to account for label uncertainty without relying on explicit class-wise score aggregation.

To reduce this overhead, we propose *mixed-label conditioning*, which represents label uncertainty through a weighted combination of learnable class embeddings. Instead of combining predictions after denoising, our method incorporates the estimated clean-label posterior probabilities directly into the conditioning representation. The denoising network processes this mixed representation in a single forward pass, avoiding explicit aggregation of separately evaluated class scores. This preserves contributions from multiple plausible classes while adapting to the class evidence available at each noise level. Although embedding mixing is not generally equivalent to output aggregation in a nonlinear network, we empirically observe approximate agreement between mixed-label and weighted class-wise predictions. To further improve training effectiveness, we introduce *mixed diffusion-model-guided sample selection*. Since the estimated probabilities may remain inaccurate or ambiguous, we assess the denoising utility of the mixed condition by comparing its diffusion loss with that of a zero-condition reference under the same perturbation and diffusion time. Examples whose mixed-label loss is no greater than the reference loss are selected for optimization. This decision is recomputed for each sampled perturbation rather than permanently excluding images, allowing training decisions to reflect the denoiser's current response to uncertain conditions. The complete framework combines efficient mixed-label prediction with selective training, with reference-evaluation costs included in the computational assessment.

We conduct extensive experiments on CIFAR-10 and CIFAR-100 under noisy-label settings, evaluating generation quality, conditional alignment, and computational cost. The results demonstrate that our framework achieves generation performance comparable to TDSM with substantially reduced training overhead, even after accounting for the reference evaluations required by sample selection. Our empirical analysis further reveals approximate agreement between mixed-label predictions and weighted class-wise predictions without an additional linearity regularizer, supporting mixed-label conditioning as an efficient alternative to explicit score aggregation. Our contributions are summarized as follows:

- We highlight a key computational bottleneck in TDSM: full score aggregation requires a separate denoiser evaluation for each class, leading to substantial training overhead as the number of classes grows.
- We propose a framework combining mixed-label conditioning with mixed diffusion-model-guided sample selection, avoiding explicit class-wise score aggregation and focusing optimization on examples that benefit from conditioning.
- Through extensive experiments on CIFAR-10 and CIFAR-100 under noisy-label settings, we demonstrate that the proposed framework achieves generation performance comparable to TDSM with substantially reduced training overhead.

2 Problem Formulation

2.1 Learning with Noisy Labels

Learning with noisy labels refers to tasks in which the dataset provides incorrect annotations that do not reflect the true class of its associated sample. Let $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$ denotes a clean dataset, where $X \in \mathbb{R}^d$ denote an image and $Y \in \{1, \dots, C\}$ denote its corresponding clean class

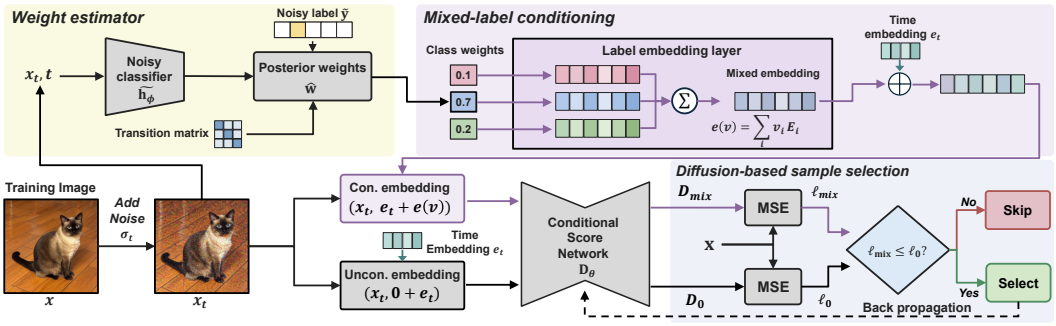


Figure 1: Overview of the proposed framework. A weight estimator accounts for label uncertainty and provides class probabilities for mixed-label conditioning. The resulting mixed embeddings guide the conditional diffusion model, avoiding explicit aggregation of class-wise predictions. Diffusion-based sample selection further evaluates the benefit of conditioning and retains suitable examples for model updates.

label. In the noisy label learning setting, this clean dataset is replaced by the unreliable dataset: $\widetilde{\mathcal{D}} = \{(x^{(i)}, \widetilde{y}^{(i)})\}_{i=1}^{N}$, where the clean labels are unobserved. In this paper, we focus exclusively on class-conditional label noise, under which the corruption process depends on the clean class but is independent of the image given that class (Natarajan et al., 2013; Patrini et al., 2017). Specifically, this noise pattern can be defined by the noise transition matrix $T \in [0,1]^{C\times C}$, which satisfies:

$$T_{ij} := \Pr(\widetilde{Y} = j \mid Y = i), \sum_{j=1}^{C} T_{ij} = 1. \tag{1}$$

Thus, each row describes how one clean class is mapped to observed annotations. The label noise rate η is defined as the probability of a clean class label being flipped into a different class. For symmetric noise, a label is retained with probability $1-\eta$ and otherwise replaced uniformly by another class, where $T_{ii} = 1-\eta$ and $T_{ij} = \eta/(C-1)$ for $i \neq j$. On the other hand, asymmetric class-conditional noise allows class-specific corruption patterns. Classical approaches combat such corruption through multiple approaches, such as noise-aware risk estimation (Natarajan et al., 2013) and transition-based loss correction (Patrini et al., 2017).

2.2 Diffusion and Denoising Score Matching

Diffusion models progressively perturb data and learn to reverse the resulting process. Following the continuous-time formulation of Song et al. (2021), the forward dynamics and the corresponding reverse-time SDE for generation conditioned on class y are:

$$\mathrm{d}X_t = f(X_t, t)\,\mathrm{d}t + g(t)\,\mathrm{d}W_t, \qquad X_0 = X, \tag{2}$$

$$\mathrm{d}X_t = \left[f(X_t, t) - g(t)^2 \nabla_{X_t} \log p_t(X_t \mid Y = y)\right]\mathrm{d}t + g(t)\,\mathrm{d}\overline{W}_t. \tag{3}$$

where $t \in [0, \tau]$ is the time peroid, f is the drift, g is the diffusion coefficient, W_t and $\overline{W}_t$ are a standard Wiener process and reverse-time Wiener process (Song et al., 2021), and the unknown conditional score $\nabla_{x_t} \log p_t(x_t \mid Y = y)$ is approximated by a neural network $s_\theta(x_t, y, t)$. The terminal time τ is chosen so that the perturbed distribution is close to a tractable prior. The perturbation process is independent of the label conditional on the initial image.

Score matching learns gradients of log-densities without explicitly evaluating their normalization constants (Hyvärinen, 2005). For diffusion training, conditional denoising score matching provides a tractable objective based on the perturbation kernel (Vincent, 2011). Let $p_{t|0}(x_t \mid x)$ denote this kernel. With clean annotations, the conditional DSM objective is

$$\mathcal{L}_{\mathrm{DSM}}(\theta) = \mathbb{E}_t\Big[\lambda(t)\mathbb{E}_{(x,y)\sim p_{\mathrm{data}}}\mathbb{E}_{x_t\sim p_{t|0}(\cdot|x)}\Big[\big\|s_\theta(x_t, y, t) - \nabla_{x_t}\log p_{t|0}(x_t \mid x)\big\|_2^2\Big]\Big], \tag{4}$$

where $\lambda(t) > 0$ controls the contribution of each noise level. For the commonly used Gaussian perturbation kernel $p_{t|0}(x_t \mid x) = \mathcal{N}(x_t; \alpha_t x, \sigma_t^2 I_d)$, the target score is available in closed form.

Under standard regularity conditions and sufficient model capacity, minimizing the population DSM objective recovers the conditional data score.

2.3 Conditional Diffusion with Corrupted Labels

Conditional diffusion with corrupted labels aims to generate images belonging to a specified clean class while training only on noisy annotations. Directly using $\widetilde{Y}$ as the condition in equation 4 learns the score of the observed-label distribution, which contains contributions from multiple classes (Na et al., 2024):

$$p_t(x_t \mid \widetilde{Y} = j) = \sum_{i=1}^{C} S_{ji}\, p_t(x_t \mid Y = i). \tag{5}$$

where $S_{ji} := \Pr(Y = i \mid \widetilde{Y} = j)$ is the reverse transition matrix. Consequently, a model conditioned on noisy labels is optimized by vague and faulty supervision information, resulting in performance degradation. TDSM (Na et al., 2024) addresses this mismatch by combining predictions from different class conditions during training:

$$\nabla_{x_t} \log p_t(x_t \mid \widetilde{Y} = j) = \sum_{i=1}^{C} w_i(x_t, j, t)\, \nabla_{x_t} \log p_t(x_t \mid Y = i), \tag{6}$$

where the class conditional weight w_i is defined by:

$$w_i(x_t, j, t) := \frac{S_{ji}\, p_t(x_t \mid Y = i)}{p_t(x_t \mid \widetilde{Y} = j)} = \Pr(Y = i \mid X_t = x_t, \widetilde{Y} = j). \tag{7}$$

These weights form a probability distribution over clean classes and depend on both the perturbed image and diffusion time. Accordingly, TDSM replaces the single conditional prediction in DSM with a corrected weighted sum:

$$\mathcal{L}_{\mathrm{TDSM}}(\theta) = \mathbb{E}_t\Big[\lambda(t)\mathbb{E}_{(x,\widetilde{y})\sim\widetilde{p}_{\mathrm{data}}}\mathbb{E}_{x_t\sim p_{t|0}(\cdot|x)}\Big[\Big\|\sum_{i=1}^{C} w_i(x_t, \widetilde{y}, t)\, s_\theta(x_t, i, t) - \nabla_{x_t}\log p_{t|0}(x_t \mid x)\Big\|_2^2\Big]\Big]. \tag{8}$$

TDSM provides theoretical proof that under class-conditional label noise and an invertible reverse transition matrix, the TDSM objective with the transition-aware weights in equation 8 has the same optimal score function as the clean-label DSM objective in equation 4. Specifically, with sufficient model capacity, its population minimizer recovers the clean class-conditional scores. In practice, TDSM estimates $w_i(x_t, j, t)$ using a time-dependent noisy-label classifier $\widetilde{h}_\phi(x_t, t)_k$ and a reverse transition matrix. The weight estimator is given by:

$$\widehat{w}_i(x_t, j, t) = \frac{\widehat{S}_{ji}\, n_j}{\widetilde{h}_\phi(x_t, t)_j} \sum_{k=1}^{C} (\widehat{S}^{-1})_{ik} \frac{\widetilde{h}_\phi(x_t, t)_k}{n_k}, \tag{9}$$

where $\widehat{S}$ estimates S, n_k denotes the number of training examples with observed label k, and the expression assumes an invertible $\widehat{S}$ and positive denominators. When the transition matrix is unknown, TDSM adapts VolMinNet (Li et al., 2021) to jointly learn a time-dependent noisy-label classifier and a forward transition matrix. The classifier and transition matrix are optimized before diffusion model training. Once this preparatory stage is complete, both are frozen and used to estimate the transition-aware weights throughout diffusion model training.

3 Methodology

TDSM handles noisy annotations through transition-aware aggregation of class-conditional scores, but obtaining these predictions requires additional class-wise denoiser evaluations. We introduce mixed-label conditioning to represent the same label uncertainty in the embedding space, producing a mixed-label prediction without explicitly evaluating each contributing class. We further employ diffusion-model-guided selective sampling to identify examples for which the mixed condition improves denoising relative to a zero-condition reference.

3.1 MIXED-LABEL CONDITIONING

As described in Section 2.3, TDSM evaluates the denoising network under individual class conditions and aggregates the resulting scores using the transition-aware weights $\widehat{w}(x_t, \widetilde{y}, t)$. These weights encode uncertainty about the underlying clean label, but combining predictions at the output level requires multiple class-conditioned network evaluations. We retain this uncertainty information while moving the combination to the conditioning space: class embeddings are mixed before the denoising network is evaluated.

Let $E \in \mathbb{R}^{C \times d_e}$ be a learnable class embedding matrix, with row E_i representing class i. For a perturbed training example $(x_t, \widetilde{y}, t)$, we use the TDSM weight estimate defined in equation 9 as its conditioning vector: $v = \widehat{w}(x_t, \widetilde{y}, t) \in \Delta^{C-1}$.The mixed-label embedding is then:

$$e(v) = \sum_{i=1}^{C} v_i E_i^\top. \tag{10}$$

This construction preserves contributions from plausible clean classes without committing to a single hard assignment. Since the weights depend on the perturbed image and diffusion time, the embedding also adapts to the class evidence available at each noise level. The mixed-label embedding is added to the time embedding and supplied to the denoising network's conditioning blocks:

$$z(v, t) = e_{\text{time}}(t) + e(v), \qquad s_\theta(x_t, v, t) := F_\theta(x_t, z(v, t)), \tag{11}$$

where the time and class embeddings have matching dimensions, and θ includes both the network parameters and E. For condition vectors arranged as rows of a matrix V, the corresponding embeddings are computed through VE. Unlike TDSM, which combines separately evaluated scores, our mixed-label conditioning combines class information before prediction, requiring only one denoiser evaluation for each mixed condition.

However, as the denoising network is nonlinear, the two forms of aggregation are not generally equivalent:

$$s_\theta(x_t, v, t) \neq \sum_{i=1}^{C} v_i s_\theta(x_t, \mathbf{u}_i, t) \quad \text{in general.} \tag{12}$$

where $\mathbf{u}_i \in \mathbb{R}^C$ denote the one-hot vector for class i, with $e(\mathbf{u}_i) = E_i^\top$. Our formulation is therefore an alternative parameterization of uncertainty-aware conditioning. Nevertheless, our empirical analysis indicates that training produces approximate agreement between mixed-label predictions and weighted class-wise predictions. This behavior emerges without an additional linearity regularizer, supporting mixed-label conditioning as an efficient approximation to TDSM's score aggregation.

3.2 DIFFUSION-MODEL-GUIDED SELECTIVE SAMPLING

Estimated label probabilities can remain inaccurate or ambiguous, and their usefulness for denoising depends on both the sampled noise level and the current model. We therefore evaluate whether the mixed condition improves reconstruction of the training image. Comparing conditional and unconditional prediction errors has also been explored in self-purifying flow matching (Kim et al., 2025). Here, we apply this principle to mixed-label diffusion predictions and retain examples whose conditional reconstruction is no worse than a zero-condition reference.

We express the selection criterion using denoised-image predictions. For the Gaussian perturbation kernel introduced in Section 2.2, we define:

$$D_\theta(x_t, v, t) = \frac{x_t + \sigma_t^2 s_\theta(x_t, v, t)}{\alpha_t}, \tag{13}$$

at training times with $\alpha_t > 0$ and $\sigma_t > 0$. For each sample, we evaluate the network twice using the same perturbed image and diffusion time to acquire a mixed-labeled conditional and unconditional output:

$$D^{\text{mix}} = D_\theta(x_t, v, t), \qquad D^0 = D_\theta(x_t, \mathbf{0}, t). \tag{14}$$

Here, $\mathbf{0} \in \mathbb{R}^C$ removes the class embedding because $e(\mathbf{0}) = \mathbf{0}$, while retaining the time conditioning. D^0 denotes the zero-condition reference and is not a class probability vector. These two

Table 1: Generation results under different label-noise settings. Bold indicates the best result among DSM, TDSM, and Ours within each noisy-label setting or average. Overall Avg. is the arithmetic mean over the seven reported noisy-label settings. CIFAR-100 results under 40% symmetric noise are omitted from the table and the average because of poor performance and are discussed in Section 4.2.

CIFAR-10

	Clean	Symmetric						Flip					
	0%	20%			40%			20%			40%		
Metric	DSM	DSM	TDSM	Ours	DSM	TDSM	Ours	DSM	TDSM	Ours	DSM	TDSM	Ours
FID ↓	7.54	8.52	**7.79**	8.93	9.10	**7.82**	9.52	7.93	**7.87**	8.16	8.38	**7.68**	8.45
IS ↑	8.66	8.50	8.67	**8.96**	8.48	8.66	**8.78**	8.65	8.70	**8.93**	8.66	8.70	**8.98**
Density ↑	79.74	77.45	78.78	**94.77**	75.12	78.79	**89.14**	78.10	78.38	**97.26**	77.73	79.93	**97.01**
Coverage ↑	80.29	78.46	**80.18**	80.08	77.29	**79.62**	78.34	79.75	80.70	**82.52**	79.59	81.08	**82.21**
CW-FID ↓	18.99	28.53	**22.47**	23.31	42.43	29.69	**24.32**	21.02	**20.09**	21.19	23.23	**20.94**	22.61
CW-Density ↑	72.95	62.98	69.13	**89.86**	52.80	62.38	**82.16**	69.89	71.24	**90.92**	67.65	71.49	**90.45**
CW-Coverage ↑	79.20	74.43	77.73	**78.03**	67.86	74.48	**75.90**	77.78	79.17	**80.51**	76.81	78.95	**79.82**

CIFAR-100 and overall averages

	Clean	Symmetric			Flip						Overall Avg.		
	0%	20%			20%			40%			7 noisy settings		
Metric	DSM	DSM	TDSM	Ours	DSM	TDSM	Ours	DSM	TDSM	Ours	DSM	TDSM	Ours
FID ↓	9.99	11.96	14.76	**9.33**	10.59	11.15	**9.99**	10.60	11.55	**9.92**	9.58	9.80	**9.19**
IS ↑	9.53	9.01	9.36	**10.64**	9.39	9.63	**10.44**	9.30	9.57	**10.41**	8.86	9.04	**9.59**
Density ↑	78.62	73.70	73.09	**88.08**	77.52	77.35	**84.25**	77.68	77.59	**85.30**	76.76	77.70	**90.83**
Coverage ↑	75.74	71.11	70.73	**77.84**	73.67	74.56	**76.64**	73.69	73.64	**76.18**	76.22	77.22	**79.12**
CW-FID ↓	84.63	101.35	96.22	**84.09**	92.86	90.25	**86.92**	102.60	102.33	**98.59**	58.86	54.57	**51.58**
CW-Density ↑	58.29	45.56	50.08	**73.39**	51.25	53.48	**69.52**	45.59	46.33	**56.26**	56.53	60.59	**78.94**
CW-Coverage ↑	71.46	63.71	65.98	**72.47**	66.86	68.41	**70.05**	61.99	62.09	**64.30**	69.92	72.40	**74.44**

outputs are generated separately through the same model. The corresponding mixed-conditioned and unconditional reconstruction losses are:

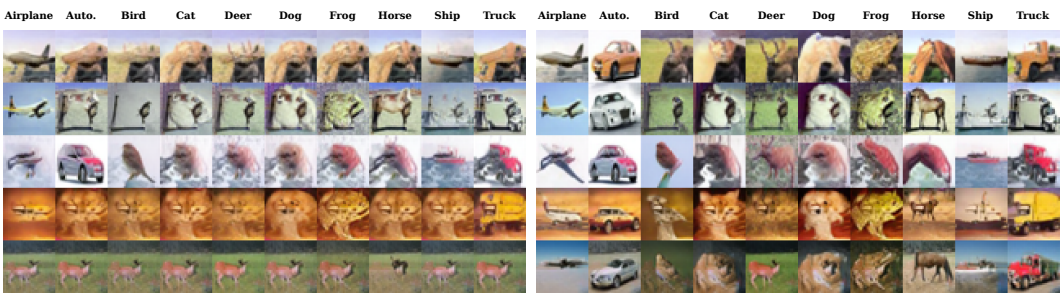
$$\ell^{\text{mix}} = \omega(t) \left\| D^{mix} - x \right\|_2^2, \qquad \ell^0 = \omega(t) \left\| D^0 - x \right\|_2^2, \tag{15}$$

where $\omega(t) > 0$ is the denoising loss weight. Because the two predictions use the same target, perturbation, and positive loss weight, their comparison measures the reconstruction benefit of supplying the mixed condition. We define the selection mask as $m = \mathbf{1}\left[\ell^{\text{mix}} \leq \ell^0\right]$. Examples satisfying this criterion contribute their mixed-label denoising losses to optimization, while others whose conditioning increases the reconstruction error are excluded from the current update. The decision is recomputed for each sampled perturbation rather than permanently removing an image from the training set. This criterion evaluates the current utility of the condition; it does not guarantee that a retained annotation is clean or that a rejected annotation is incorrect.

Unlike the conditioning-replacement strategy (Kim et al., 2025), this procedure does not replace v with either the observed label or a hard classifier prediction. All retained examples use their transition-aware mixed conditions. The zero-condition prediction is used only to determine the mask, and rejected examples do not receive an unconditional training loss.

3.3 Training and Generation

Mixed-label conditioning and selective sampling together define the diffusion training procedure. The former constructs each example's condition from the estimated TDSM weights, while the latter determines whether its mixed-label prediction contributes to the update. The classifier and transition matrix obtained in Section 2.3 remain frozen, so only the denoising network and class embeddings are optimized during this stage.



(a) DSM (b) Ours

Figure 2: Conditional sampling from perturbed CIFAR-10 images. Randomly selected images are corrupted and used to initialize sampling with DSM and our method trained on CIFAR-10 with 40% sym noise. Each row corresponds to an input image, and each column specifies the target class. Our method exhibits more distinct responses to the requested class conditions.

For a mini-batch $\mathcal{B} = \{(x_b, \widetilde{y}_b)\}_{b=1}^{B}$, we independently sample a diffusion time and a perturbed image for each example, before using the TDSM method and our mixed-label conditioning to calculate the estimated weights $\widehat{w}$ and the estimated output pair $\{D^{mix}, D^0\}$ sequentially. The selection mask m_b is subsequently acquired by the method in 3.2. The overall mini-batch objective is:

$$\widehat{\mathcal{L}}_{\text{ours}}(\theta; \mathcal{B}) = \frac{1}{B_{\text{sel}}} \sum_{b=1}^{B} m_b \, \ell_b^{\text{mix}}, \qquad B_{\text{sel}} = \sum_{b=1}^{B} m_b. \tag{16}$$

The retained mixed-label losses update the denoising network and embedding matrix, while the zero-condition branch serves only as a reference for selection. Examples excluded from the current update may be selected when revisited with a different perturbation or at a later training stage. Figure 1 illustrates our overall training procedure. At generation time, the requested class y is represented by the one-hot vector $\mathbf{u}_y$. Starting from the terminal prior, the model uses $s_\theta(x_t, \mathbf{u}_y, t)$ in the reverse process of equation 3, or its corresponding numerical sampler. Generation requires neither the weight estimator nor the sample selection procedure.

4 Experiments

4.1 Experimental Setup

We evaluate class-conditional image generation on CIFAR-10 and CIFAR-100 (Krizhevsky, 2009) under symmetric and asymmetric label noise with nominal corruption rates of 20% and 40%, following TDSM (Na et al., 2024). We use a class-conditional U-Net with EDM preconditioning (Karras et al., 2022), trained for 1,000 epochs with a batch size of 256 using Adam and a peak learning rate of 5×10^{-4}. The classifier and transition matrix are estimated before diffusion training and subsequently frozen. We evaluate overall generation quality using FID (Heusel et al., 2017), IS (Salimans et al., 2016), and density and coverage (Naeem et al., 2020), and assess class-conditional generation using CW-FID (Miyato & Koyama, 2018; Kaneko et al., 2019), CW-Density, and CW-Coverage (Chao et al., 2022), following TDSM (Na et al., 2024). Additional details of the datasets, noise construction, architecture, training, and evaluation are provided in Appendix A.2.

4.2 Main Results

The comparative results are presented in Table 1, where Overall Avg. summarizes performance over the seven reported noisy-label settings across CIFAR-10 and CIFAR-100. Our method achieves the best average performance on all seven metrics and consistently obtains the highest IS, density, CW-Density, and CW-Coverage in the included settings. Its advantages are broader on CIFAR-100, where it outperforms both DSM and TDSM on every reported metric. On CIFAR-10, the comparison is more mixed: TDSM generally achieves better FID, whereas our method provides higher density and class-wise coverage. Thus, the overall improvements reflect complementary strengths across evaluation criteria rather than uniform dominance in every setting.

Table 2: Generation results on CIFAR-10 and CIFAR-100 with clean labels (0% noise). Bold indicates the best result for each metric and dataset.

	CIFAR-10			CIFAR-100		
Metric	DSM	TDSM	Ours	DSM	TDSM	Ours
FID ↓	7.54	**7.51**	7.95	9.99	10.21	**8.71**
IS ↑	8.66	8.65	**8.93**	9.53	9.90	**10.80**
Density ↑	79.74	79.22	**98.71**	78.62	79.19	**86.27**
Coverage ↑	80.29	81.51	**82.84**	75.74	76.39	**78.85**
CW-FID ↓	18.99	**18.95**	20.66	84.63	83.51	**81.72**
CW-Density ↑	72.95	73.08	**93.16**	58.29	60.46	**72.84**
CW-Coverage ↑	79.20	80.42	**81.30**	71.46	73.01	**73.70**

Table 3: Result for mixed-label conditioning without sample selection (Mixed) on CIFAR-10 with 20% label noise.

	Symmetric 20%		Flip 20%	
Metric	TDSM	Mixed	TDSM	Mixed
FID ↓	**7.79**	8.23	7.87	**7.67**
IS ↑	**8.67**	8.55	**8.70**	8.69
Density ↑	**78.78**	76.90	78.38	**79.02**
Coverage ↑	**80.18**	78.94	**80.70**	80.26
CW-FID ↓	**22.47**	23.37	20.09	**19.95**
CW-Density ↑	**69.13**	67.05	71.24	**71.39**
CW-Coverage ↑	**77.73**	77.61	**79.17**	78.79

Beyond the average results, comparing aggregate and class-wise metrics shows that label noise can impair conditional alignment even when overall image quality remains relatively stable. For DSM, increasing symmetric noise degrades CW-FID more strongly than aggregate FID. Both TDSM and our method alleviate this deterioration on CIFAR-10, with our method achieving better CW-FID under 40% symmetric noise despite a higher aggregate FID. Our consistently higher CW-Density and CW-Coverage suggest improvements in within-class fidelity and diversity, while the reported CIFAR-100 settings also show gains in both aggregate and class-wise FID. Figure 2 further compares generation initialized from Gaussian-perturbed CIFAR-10 images: DSM tends to preserve similar semantic content across target conditions, whereas our method produces more distinct class-dependent changes. Additional examples are provided in Appendix in Figure 3 and Figure 4.

CIFAR-100 with 40% symmetric noise remains a challenging exception and is excluded from the main table and the reported averages, in which both TDSM and our method deteriorate substantially This limitation suggests that the benefits observed under moderate corruption do not extend automatically to severe symmetric noise in a many-class setting. One possible explanation is that unreliable conditioning weights become more difficult to compensate for, although verifying this mechanism requires further analysis of the weight estimator and selection behavior.

4.3 Additional Studies

Clean-label evaluation. We compare DSM, TDSM, and our method under clean training labels to assess whether the proposed framework remains effective in the absence of annotation corruption. As shown in Table 2, DSM and TDSM perform similarly, whereas our method achieves higher IS, density, coverage, CW-Density, and CW-Coverage on both datasets. The concurrent improvements in density and coverage suggest that the framework benefits sample fidelity while preserving broader coverage of the real-data distribution, including within individual classes. Our method also achieves the best FID and CW-FID on CIFAR-100, although these two metrics are slightly worse on CIFAR-10, indicating that its advantages are not uniform across all evaluation criteria. Overall, the favorable results under clean supervision show that the framework's effectiveness is not restricted to learning from corrupted annotations. Together with the noisy-label experiments, these findings support its adaptability to training data with varying levels of label reliability.

Training efficiency. We measure training time per epoch to evaluate the computational cost of mixed-label conditioning and the additional overhead introduced by sample selection. As shown in Table 4, DSM provides the fastest reference, whereas TDSM incurs substantially higher training costs, particularly on CIFAR-100. This pattern is consistent with the additional class-wise denoiser evaluations required by explicit score aggregation. In contrast, mixed-label conditioning combines class information in the embedding layer, substantially reducing this overhead. Mixed achieves speedups of $1.72\times$ and $3.84\times$ over TDSM on CIFAR-10 and CIFAR-100, respectively.

Adding sample selection increases training time relative to Mixed because the selection criterion requires an additional zero-condition reference prediction. Nevertheless, the complete framework remains $1.43\times$ faster than TDSM on CIFAR-10 and $3.27\times$ faster on CIFAR-100, corresponding

Table 4: Training time per epoch and speedup relative to TDSM. Mixed represents models using mixed-label conditioning alone, while Ours uses mixed-label + sample selection. The mixed-only method shows a significant advantage in efficiency, while sample selection produces moderate costs.

	CIFAR-10		CIFAR-100	
Method	Time (s)↓	Speedup	Time (s)↓	Speedup
DSM	**28.61**	2.00×	**28.56**	4.97×
TDSM	57.23	1.00×	142.03	1.00×
Mixed	33.29	1.72×	36.97	3.84×
Ours	39.98	1.43×	43.50	3.27×

Table 5: Prediction agreement under 20% symmetric label noise. Original uses TDSM weights; Random uses random probability weights. R is the discrepancy-to-reconstruction-error ratio, expressed as a percentage.

	Prediction agreement		
Dataset	Weights	R (%) ↓	Cosine ↑
CIFAR-10	Original	0.0749	0.99995
	Random	0.7664	0.99965
CIFAR-100	Original	0.1864	0.99989
	Random	1.1419	0.99954

to reductions in per-epoch training time of 30.1% and 69.4%. The larger advantage on CIFAR-100 supports the computational benefit of replacing class-wise score aggregation with embedding mixing when more classes are involved. These results also distinguish the roles of the two components: mixed-label conditioning provides the main computational savings, while sample selection introduces a modest additional cost to guide optimization.

Performance of mixed-label conditioning To show that our mixed-label conditioning does not produce major training interference while maintaining output linearity, we evaluate mixed-label conditioning without sample selection to examine its generation performance and its agreement with explicit class-wise output aggregation. Table 3 shows that Mixed achieves numerically close performance to TDSM on CIFAR-10 under 20% label noise. TDSM performs better under symmetric noise, whereas Mixed achieves slightly better FID, CW-FID, density, and CW-Density under flip noise. Together with the lower training cost reported in Table 4, these results support embedding mixing as a computationally efficient alternative under moderate label noise. To further investigate this behavior, we compare $D_{\text{mix}} = D_\theta(x_t, w, t)$ with $D_{\text{agg}} = \sum_{i=1}^{C} w_i D_\theta(x_t, i, t)$ using the same network, perturbed images, and noise levels. Table 5 reports cosine similarity and the discrepancy ratio $R = \mathbb{E}[\|D_{\text{mix}} - D_{\text{agg}}\|_2^2]/\mathbb{E}[\|D_{\text{mix}} - x\|_2^2]$. With TDSM-estimated weights, the discrepancy remains below 0.2% of the reconstruction error on both datasets, accompanied by cosine similarities close to one. Random probability weights increase the discrepancy, but it remains small relative to the reconstruction error. These observations provide empirical evidence that embedding mixing can approximate output aggregation, without establishing exact network linearity or equivalence to an independently trained TDSM model. The evaluation of the mixed-only condition on CIFAR-10 with 40% can be found in Table 6 in the appendix.

5 Conclusion

In this paper, we presented a framework for learning class-conditional diffusion models from noisy annotations through mixed-label conditioning and diffusion-based sample selection. By combining class embeddings according to estimated label probabilities, mixed-label conditioning represents label uncertainty within the conditioning space and replaces explicit class-wise score aggregation with a single mixed-condition denoiser evaluation, reducing the repeated computation associated with evaluating multiple class conditions. Sample selection further guides optimization by comparing mixed-condition and unconditional denoising losses, retaining examples for which conditioning improves reconstruction. Experiments on CIFAR-10 and CIFAR-100 demonstrate competitive generation performance, with consistent improvements in several distributional metrics across the reported settings and lower per-epoch training costs than TDSM. Additional experiments further support the applicability of the framework across different levels of annotation reliability, as well as supporting embedding mixing as a practical alternative to explicit score aggregation. Nevertheless, performance trade-offs remain across metrics. Future work will investigate more reliable conditioning-weight estimation and selection strategies for stronger label corruption.

AI USE STATEMENT

In this work, we used generative AI tools to assist with implementing methods. We have not used generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for mathematical proofs, write proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, translate research material, clean or reformat datasets, support qualitative or thematic data analysis, or interpret results. Additionally, we used generative AI tools to edit the manuscript for grammar, clarity, and readability. We reviewed all AI-assisted work, checking the implementation against the intended methodology and ensuring that textual revisions preserved the intended technical meaning. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the assistance of generative AI.

REFERENCES

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 *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2203.14206.

Dat Nguyen Cong, Hieu Tran Bao, and Tung Hoang-Thanh. Guiding noisy label conditional diffusion models with score-based discriminator correction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18531–18541, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Cong_Guiding_Noisy_Label_Conditional_Diffusion_Models_with_Score-based_Discriminator_Correction_ICCV_2025_paper.html.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794, 2021.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/a19744e268754fb0148b017647355b7b-Abstract.html.

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 *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. URL https://arxiv.org/abs/2207.12598.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005. URL https://www.jmlr.org/papers/v6/hyvarinen05a.html.

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*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2304–2313. PMLR, 2018. URL https://proceedings.mlr.press/v80/jiang18c.html.

Takuhiro Kaneko, Yoshitaka Ushiku, and Tatsuya Harada. Label-noise robust generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2467–2476, 2019.

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*, 2022. URL https://arxiv.org/abs/2206.00364.

Hyeongju Kim, Yechan Yu, June Young Yi, and Juheon Lee. Training flow matching models with reliable labels via self-purification, 2025. URL https://arxiv.org/abs/2509.19091.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Junnan Li, Richard Socher, and Steven C. H. Hoi. DivideMix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HJgExaVtwr.

Shikun Li, Xiaobo Xia, Shiming Ge, and Tongliang Liu. Selective-supervised contrastive learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 316–325, 2022.

Xuefeng Li, Tongliang Liu, Bo Han, Gang Niu, and Masashi Sugiyama. Provably end-to-end label-noise learning without anchor points. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6403–6413. PMLR, 2021. URL https://proceedings.mlr.press/v139/li21l.html.

Yifan Li, Hu Han, Shiguang Shan, and Xilin Chen. DISC: Learning from noisy labels via dynamic instance-specific selection and correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24070–24079, 2023.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

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*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/ea89621bee7c88b2c5be6681c8ef4906-Abstract.html.

Takeru Miyato and Masanori Koyama. cGANs with projection discriminator. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1802.05637.

Byeonghu Na, Yeongmin Kim, HeeSun Bae, Jung Hyun Lee, Se Jung Kwon, Wanmo Kang, and Il-Chul Moon. Label-noise robust diffusion models. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2402.17517.

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*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7176–7185. PMLR, 2020. URL https://proceedings.mlr.press/v119/naeem20a.html.

Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep K. Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Patrini_Making_Deep_Neural_CVPR_2017_paper.html.

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*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4334–4343, 2018. URL https://proceedings.mlr.press/v80/ren18a.html.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/8a3363abe792db2d8761d6403605aeb7-Abstract.html.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265, 2015. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2011.13456.

Kiran K. Thekumparampil, Ashish Khetan, Zinan Lin, and Sewoong Oh. Robustness of conditional GANs to noisy labels. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/565e8a413d0562de9ee4378402d2b481-Abstract.html.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011. doi: 10.1162/NECO_a_00142.

Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 322–330, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Wang_Symmetric_Cross_Entropy_for_Robust_Learning_With_Noisy_Labels_ICCV_2019_paper.html.

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*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/f2925f97bc13ad2852a7a551802feea0-Abstract.html.

A Appendix

A.1 Related Work

A.1.1 Diffusion and Conditional Generative Models

Diffusion models learn generative processes by reversing progressive noise corruption, an approach initially developed through nonequilibrium thermodynamics (Sohl-Dickstein et al., 2015). Denoising diffusion probabilistic models (DDPMs) (Ho et al., 2020) established an effective denoising objective for image synthesis, while score-based generative modeling (Song et al., 2021) unified diffusion processes through stochastic differential equations. Elucidated Diffusion Models (EDM) (Karras et al., 2022) further improved preconditioning, noise-level sampling, and numerical sampling procedures, providing the training formulation adopted in our implementation. Related to diffusion, flow matching (Lipman et al., 2023) learns time-dependent vector fields along prescribed probability paths, providing another framework for transporting noise into data.

Conditional generation incorporates semantic information into the generative process to control the resulting samples. Classifier guidance (Dhariwal & Nichol, 2021) uses gradients from a noise-conditioned classifier to guide diffusion sampling. Classifier-free guidance (Ho & Salimans, 2022) instead combines conditional and unconditional predictions without requiring an external classifier at inference time. Latent diffusion models (Rombach et al., 2022) reduce the cost of high-resolution synthesis by operating in a learned latent space and support flexible conditioning through cross-attention. These developments improve generation quality, controllability, and efficiency, but do not by themselves resolve corrupted image–condition associations. Our work focuses on training with uncertain class annotations, using mixed-label embeddings to avoid explicit aggregation of multiple class-conditioned denoiser predictions.

A.1.2 Learning with Noisy Labels

Discriminative learning. Traditional learning with noisy labels encompasses several complementary directions. Sample selection and weighting methods identify or emphasize reliable supervision (Jiang et al., 2018; Han et al., 2018), while robust loss functions reduce sensitivity to corrupted annotations (Zhang & Sabuncu, 2018; Wang et al., 2019). Transition-based loss correction explicitly models class-dependent corruption (Patrini et al., 2017), with subsequent work addressing transition estimation without anchor points (Li et al., 2021). Other approaches include meta-learning of example weights using clean validation data (Ren et al., 2018), early-learning regularization (Liu et al., 2020), and semi-supervised learning with estimated clean and noisy subsets (Li et al., 2020). Later methods further explore robust representation learning through selective contrastive supervision (Li et al., 2022) and dynamic instance-specific selection and label correction (Li et al., 2023). These strategies primarily target discriminative prediction. Conditional generation additionally requires preserving the image distribution of each clean class, while diffusion training makes the available class evidence dependent on the noise level.

Conditional generation with noisy labels. Label-noise robustness has also been investigated in conditional generative adversarial networks. Robust Conditional GANs (Thekumparampil et al., 2018) corrupt generated conditioning labels before adversarial comparison, allowing the generator to model clean-label distributions while matching noisy observations. Label-noise robust GANs (Kaneko et al., 2019) similarly incorporate a noise transition model into conditional adversarial learning. For diffusion models, TDSM (Na et al., 2024) combines class-conditioned score predictions using instance- and time-dependent transition-aware weights. This formulation connects noisy-label supervision to clean conditional scores, but explicit aggregation introduces additional class-wise network evaluations. Score-based Discriminator Correction (SBDC) (Cong et al., 2025) takes a different approach: it trains an auxiliary discriminator to provide sampling-time correction for a pretrained noisy-label conditional diffusion model, without retraining the denoiser. Training Flow Matching Models with Reliable Labels via Self-Purification (Kim et al., 2025) introduces Self-Purifying Flow Matching (SPFM), which compares conditional and unconditional flow-matching losses to identify unreliable conditions during training. Samples with suspicious conditions are retained for unconditional training, allowing their data content to remain useful. This work demonstrates that a generative model's own predictions can provide a signal for evaluating conditioning

reliability. Our sample selection shares this motivation, while our framework focuses on mixed-label diffusion training: it represents label uncertainty through weighted embeddings and evaluates the utility of the resulting mixed condition for denoiser optimization. Thus, our emphasis is on combining efficient uncertainty-aware conditioning with model-guided selective training.

A.2 Additional Experimental Details

Datasets and label corruption. CIFAR-10 and CIFAR-100 (Krizhevsky, 2009) each contain 50,000 training images and 10,000 test images at a resolution of 32×32. CIFAR-10 contains 10 classes, with 5,000 training images per class, whereas CIFAR-100 contains 100 classes, with 500 training images per class, organized into 20 superclasses. We apply synthetic corruption to training labels while retaining the original images. Following TDSM (Na et al., 2024), we consider symmetric and asymmetric class-conditional noise at nominal rates of 20% and 40%, together with clean-label experiments. Symmetric noise randomly replaces annotations across classes. For asymmetric noise on CIFAR-10, the prescribed transitions are truck $\rightarrow$ automobile, bird $\rightarrow$ airplane, deer $\rightarrow$ horse, and cat $\leftrightarrow$ dog; the remaining classes retain their labels. For CIFAR-100, asymmetric corruption cyclically changes labels among the five classes within each superclass. The nominal asymmetric noise rate applies to eligible transitions and need not equal the fraction of corrupted labels across the entire dataset. The main comparison and its seven-setting average exclude CIFAR-100 with 40% symmetric noise, for which substantial performance degradation is discussed separately.

Denoising architecture. We use a class-conditional U-Net with EDM preconditioning (Karras et al., 2022). The network has a base channel width of 64, channel multipliers $(1, 2, 2, 2)$, and two residual blocks per encoder resolution level. Single-head self-attention is applied at the 16×16 resolution and in the bottleneck. The dropout rate is 0.13. Class embeddings and the transformed noise-level embedding have dimension 256. Mixed-label conditioning forms a weighted sum of learnable class embeddings using the estimated TDSM weights and adds the resulting vector to the noise-level embedding. We use linear embedding mixing without additional embedding-norm rescaling. For the unconditional reference prediction used in sample selection, the class-conditioning contribution is set to zero.

Diffusion training. The default configuration uses 1,000 epochs and a batch size of 256. Training samples are shuffled, and the incomplete final batch is discarded, yielding 195 updates per epoch and 195,000 updates in total. We initialize Adam with $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$, and zero weight decay. A linear OneCycle learning-rate schedule increases the learning rate from 5×10^{-8} to 5×10^{-4} during the first 10% of training and then decreases it to 5×10^{-8}. The scheduler retains its default momentum-cycling setting. We maintain an exponential moving average of the denoiser parameters with decay 0.999 and use these averaged parameters for generation. The default random seed is 42, and random horizontal flipping is disabled.

Following the EDM training parameterization, we sample the Gaussian noise scale according to:

$$\log \sigma \sim \mathcal{N}(-1.2, 1.2^2), \qquad x_\sigma = x + \sigma\epsilon, \qquad \epsilon \sim \mathcal{N}(0, I). \tag{17}$$

We set $\sigma_{\mathrm{data}} = 0.5$ and weight the squared reconstruction error by:

$$\lambda(\sigma) = \frac{\sigma^2 + \sigma_{\mathrm{data}}^2}{(\sigma\sigma_{\mathrm{data}})^2}. \tag{18}$$

For sample selection, mixed-label and unconditional predictions are evaluated using the same perturbed image, noise scale, and clean reconstruction target.

Weight-estimator pretraining. We estimate the classifier and label transition matrix using a VolMinNet-based objective (Li et al., 2021) with noise-level conditioning as in TDSM (Na et al., 2024). The classifier uses a ResNet-18 backbone for CIFAR-10 and a ResNet-34 backbone for CIFAR-100, with positional noise-level embeddings. Pretraining lasts 150 epochs with a batch size of 128. Images are perturbed using the same noise-scale distribution as diffusion training. The classifier is optimized using SGD with an initial learning rate of 10^{-2}, momentum 0.9, and weight decay of 10^{-3} for CIFAR-10 and 10^{-4} for CIFAR-100. Its learning rate is reduced by a factor of 0.1

Table 6: Comparison between TDSM and mixed-label conditioning without sample selection (Mixed) on CIFAR-10 with 40% label noise. Bold indicates the better result for each metric and noise type.

	Symmetric 40%		Flip 40%	
Metric	TDSM	Mixed	TDSM	Mixed
FID ↓	**7.82**	8.86	**7.68**	7.98
IS ↑	**8.66**	8.53	**8.70**	8.66
Density ↑	**78.79**	75.46	**79.93**	79.22
Coverage ↑	**79.62**	77.74	**81.08**	80.00
CW-FID ↓	**29.69**	32.14	**20.94**	21.41
CW-Density ↑	**62.38**	59.25	**71.49**	70.73
CW-Coverage ↑	**74.48**	72.06	**78.95**	77.75

at epochs 30 and 60. The transition-matrix parameters use a separate Adam optimizer with learning rate 10^{-3} and zero weight decay. The objective combines the negative log-likelihood of the observed labels with a log-determinant volume regularizer weighted by 10^{-4}. The transition matrix is updated during the first 25 pretraining epochs, whereas classifier optimization continues for all 150 epochs. After pretraining, both components are frozen throughout diffusion training and provide the weights used for mixed-label conditioning.

Generation and evaluation protocol. We generate 50,000 images per evaluated model in 50 batches of 1,000 using the exponential-moving-average denoiser parameters. We employ the deterministic EDM sampler (Karras et al., 2022) with 18 sampling steps, using second-order Heun updates and a final Euler step, requiring 35 denoiser evaluations per image. The noise schedule uses $\sigma_{\min} = 0.002$, $\sigma_{\max} = 80$, and $\rho = 7$, with stochastic churn disabled ($S_{\text{churn}} = 0$). Sampling starts from Gaussian noise, with target classes drawn uniformly at random. We use the training images as the real-image reference set and their original clean labels for class-wise evaluation.

Overall generation metrics. We report Fréchet Inception Distance (FID) (Heusel et al., 2017), Inception Score (IS) (Salimans et al., 2016), and density and coverage (Naeem et al., 2020). FID compares the means and covariances of real and generated image features, with lower values indicating closer distributional agreement. IS rewards confident class predictions for individual generated images and diversity in the marginal predicted class distribution; it does not directly measure agreement with the requested conditioning labels. Density measures the normalized number of real-sample neighborhoods containing generated samples, whereas coverage measures the fraction of real-sample neighborhoods containing at least one generated sample. These metrics provide complementary assessments of sample fidelity and distributional coverage. Higher IS, density, and coverage are preferred.

Class-wise generation metrics. To evaluate conditional generation, we report CW-FID, also referred to as intra-FID (Miyato & Koyama, 2018; Kaneko et al., 2019), together with CW-Density and CW-Coverage (Chao et al., 2022). Following TDSM (Na et al., 2024), each metric is computed between real images from a clean class and generated images conditioned on that class, then averaged uniformly across classes:

$$\text{CW-}M = \frac{1}{C}\sum_{c=1}^{C} M\left(\mathcal{X}_c^{\text{real}}, \mathcal{X}_c^{\text{gen}}\right), \tag{19}$$

where M denotes FID, density, or coverage. Unlike pooled metrics, these comparisons assess whether the generated distribution matches the requested class. Lower CW-FID and higher CW-Density and CW-Coverage are preferred.

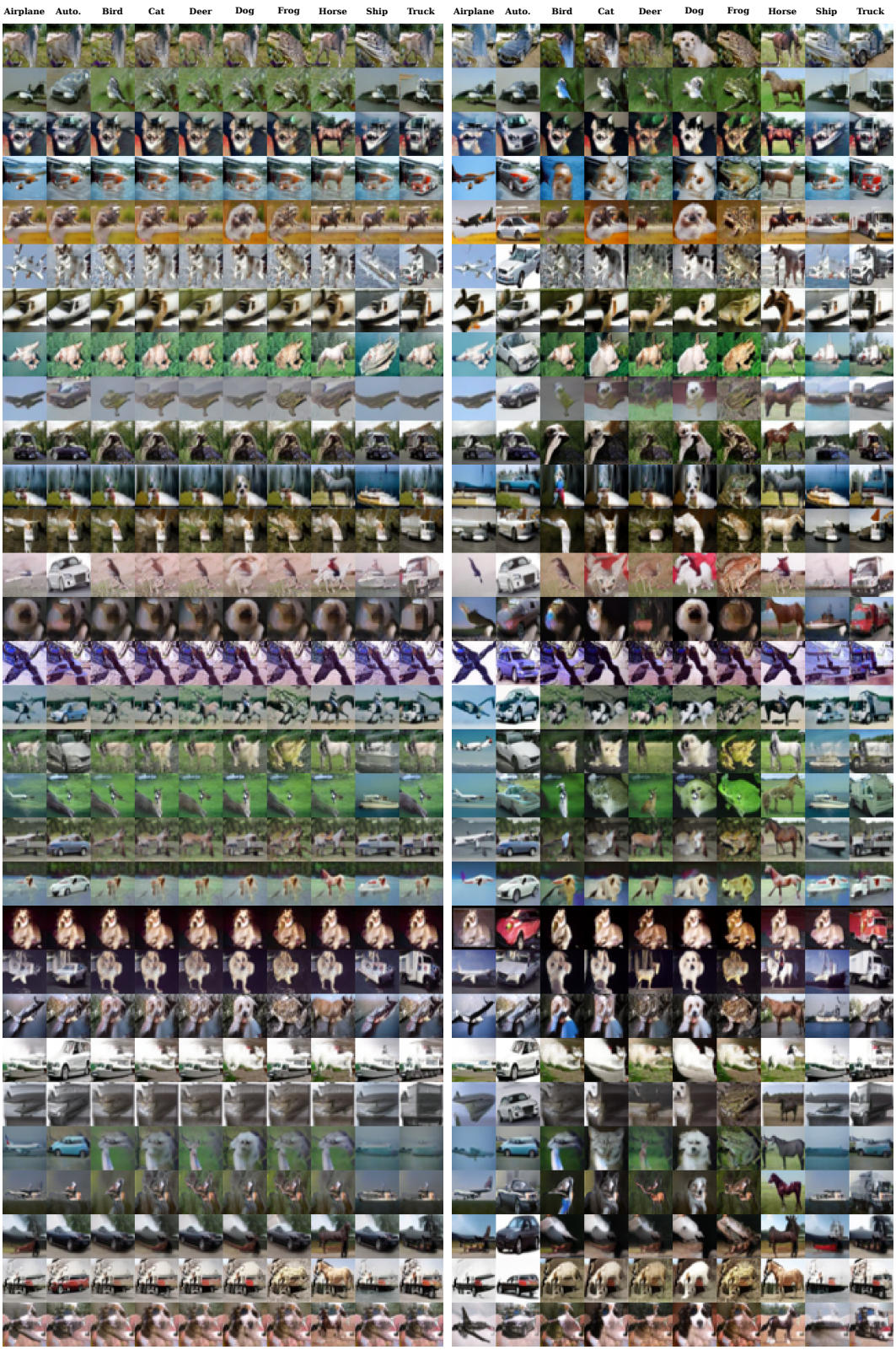
A.3 Additional Results for Mixed-Label Conditioning

A.3.1 Performance of mixed-label conditioning under stronger label noise.

To further evaluate mixed-label conditioning, we extend the comparison to 40% label noise. As shown in Table 6, TDSM outperforms Mixed on all reported metrics under both corruption types. Although the performance gap generally widens compared with the 20% noise setting, particularly under symmetric noise, the differences remain modest overall, with closer agreement under flip noise. These results support the effectiveness of mixed-label conditioning as a computationally efficient alternative to explicit class-wise score aggregation, while showing a moderate performance trade-off under stronger label noise.

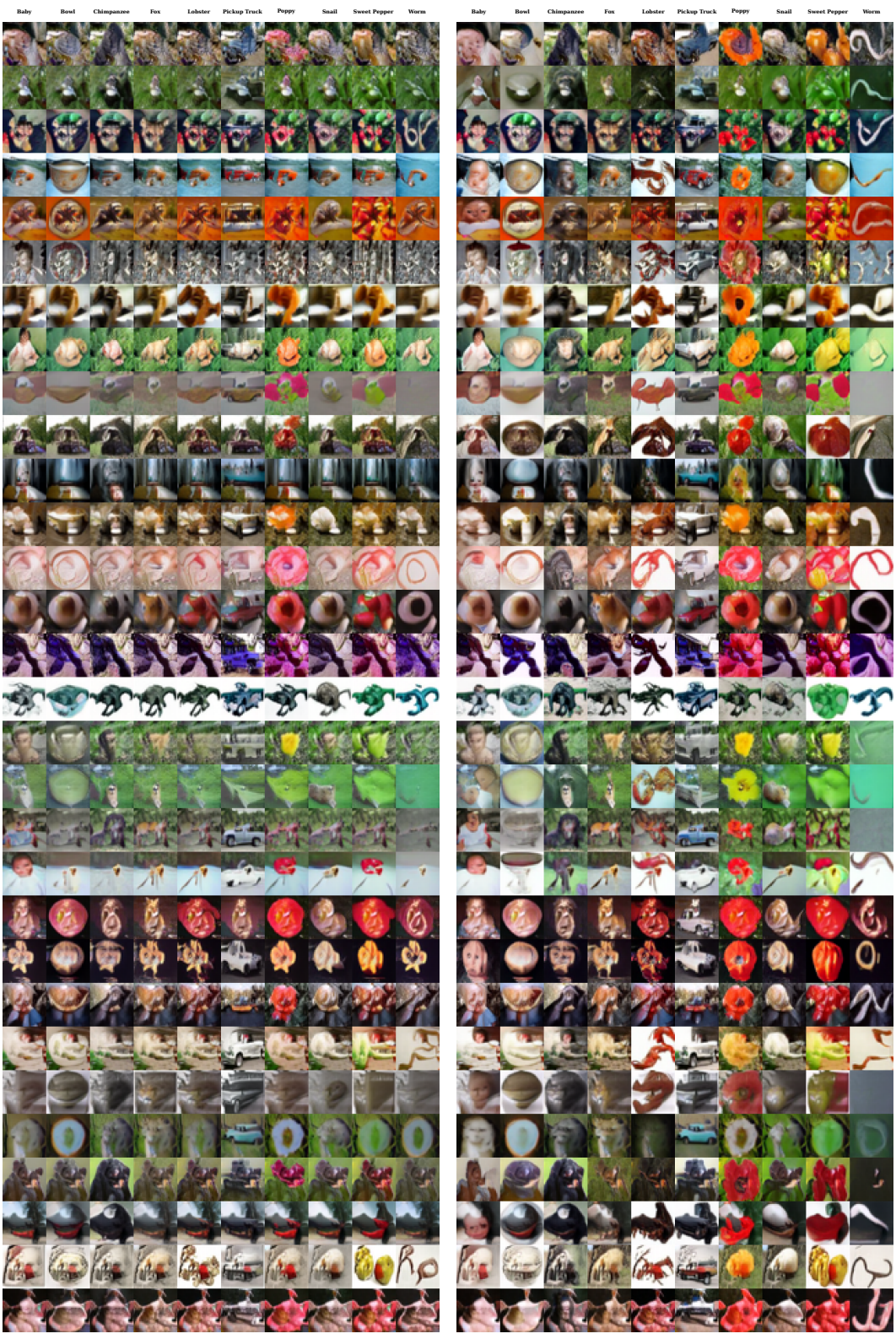
A.3.2 Additional Sampling results on CIFAR-10 and CIFAR100 datasets.

Figure 3 and Figure 4 demonstrate more generation results of DSM and our method trained on CIFAR-10 with 40% symmetric noise and CIFAR100 with 20% symmetric noise. The additional examples show a similar pattern to the main qualitative comparison. DSM frequently preserves the input's dominant semantic appearance across different target classes, whereas our method produces more noticeable class-dependent changes, particularly between vehicle and animal conditions. Many outputs also retain aspects of the input color palette and background while modifying the foreground structure.



(a) DSM (b) Ours

Figure 3: Additional sampling from perturbed CIFAR-10 images by DSM and our method trained on CIFAR-10 dataset with 40% sym noise.



(a) DSM (b) Ours

Figure 4: Qualitative comparison on CIFAR-100 under 20% symmetric label noise. Columns correspond to ten randomly selected target classes, and rows show generation from different Gaussian-perturbed input images. Our method exhibits more pronounced class-dependent changes than DSM, although some outputs retain substantial input characteristics.