DELTA: ROBUSTLY TRAINING DIFFUSION MODELS WITH WEAK ANNOTATIONS

Dong-Dong Wu[1,2] **Jiacheng Cui**[3] **Wei Wang**[2,1] **Zhiqiang Shen**[3] **Masashi Sugiyama**[1,2]

[1]The University of Tokyo, Chiba, Japan
[2]RIKEN AIP, Tokyo, Japan
[3]Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
{dongdong.wu, wei.wang}@riken.jp
{jiacheng.cui, zhiqiang.shen}@mbzuai.ac.ae sugi@k.u-tokyo.ac.jp

ABSTRACT

Conditional diffusion models have achieved remarkable success in generative tasks, yet their standard training typically relies on clean labeled data. In contrast, real-world scenarios frequently involve weak annotations such as noisy, ambiguous, or incomplete supervision, which will lead to biased score estimation and significantly degrade generation performance. To address this challenge, we propose *DELTA*, a unified framework for robustly training Diffusion ModELs with weak annoTATions. To our knowledge, this is the first systematic study unifying these diverse weak annotations within diffusion models. Grounded in likelihood maximization, our framework decomposes the training objective into two synergistic components: a generative term that recovers clean conditional scores via posterior-weighted score matching, and a classification term that infers reliable class-posterior probabilities using the diffusion model itself. To improve computational efficiency, we further develop an optimized timestep sampling strategy for the diffusion classifier. Extensive experiments across multiple tasks demonstrate *DELTA*'s effectiveness in overcoming the limitations of weak annotations.

1 INTRODUCTION

Recently, diffusion models (DMs) (Ho et al., 2020; Song et al., 2020) have emerged as a premier generative framework, demonstrating unprecedented capabilities in synthesizing high-fidelity data (He et al., 2025; Ho et al., 2022). Building on this foundation, conditional diffusion models (CDMs) incorporate guidance mechanisms (Ho & Salimans, 2022) to enable controllable generation, allowing synthesis conditioned on specific attributes such as class labels. Driven by this flexibility, CDMs have achieved remarkable success in various applications, such as image inpainting (Cross-Zamirski et al., 2023), super-resolution (Okada et al., 2024), and semantic editing (Wang et al., 2025b).

Despite their success, CDMs require accurate conditioning signals, which are often unavailable in practice. Real-world data, typically sourced from the web or crowdsourcing, naturally suffer from various imperfections due to limited budget, privacy constraints, or human error. We refer to such scenarios collectively as *weak annotations* Sugiyama et al. (2022), where the provided annotations often deviate from the ground truth. Prevalent forms of weak annotations include noisy annotations (Li et al., 2017) that are corrupted, ambiguous annotations (Wang et al., 2025c;a) that offer only a candidate set, and semi-supervised data (He et al., 2023) where annotations are largely missing. Training CDMs directly with such unreliable signals will introduce incorrect inductive biases, leading to condition mismatch that severely degrades generation performance.

To address these challenges, recent studies have proposed adaptations of diffusion models, such as noise-robust (Na et al., 2024; Li et al., 2025; Dufour et al., 2024) and semi-supervised variants (You et al., 2023). However, these approaches typically target specific annotation types in isolation, lacking a unified perspective. Furthermore, they often rely on explicit external priors or auxiliary knowledge. For instance, prior works may require pre-estimating noise transition matrices (Na et al., 2024) or accessing risk confidence scores (Li et al., 2025). Such dependence on task-specific heuristics or

external supervision limits their generalizability and practical efficiency. Consequently, there remains a pressing need for a unified framework capable of robustly training CDMs under diverse forms of weak annotations, without relying on strong prior assumptions.

In line with prior research (Na et al., 2024; Chen et al., 2025), we treat *class-conditional generation* as the foundational paradigm of conditional synthesis and focus on it in this work. To robustly train a CDM in a unified manner, we first formulate the overall learning objective as maximizing the joint likelihood of data and weak annotations (Section 3.1). Then we decompose the learning objective via a variational lower bound into two synergistic components: a *Generative Term* that learns clean-conditional distributions via posterior-weighted score matching (Section 3.2), and a *Classification Term* that infers the posterior probabilities of true labels from weak annotations (Section 3.3). Our theoretical analysis reveals that the score conditioned on weak annotations can be expressed as a posterior-weighted expectation of clean-label scores. This mechanism allows the model to learn accurate class-conditional distributions by enforcing the posterior-weighted aggregation of predicted clean-label scores to match the observable weak annotation score. To mitigate the computational cost of iterative posterior inference, we further propose an efficient timestep sampling strategy (Section 3.4) that significantly reduces computational overhead without compromising accuracy. Extensive experiments on image generation and weakly supervised classification demonstrate the robustness of our framework against weak annotations. Our contributions are summarized as follows:

- We propose a unified diffusion framework for training CDMs under diverse forms of weak annotations, which is the first exploration in the diffusion model field.
- We develop a timestep sampling strategy for diffusion classifiers that greatly improves the computational efficiency of posterior estimation without compromising accuracy.
- Experiments across diverse benchmarks demonstrate our framework's robustness, yielding class-consistent generation and competitive classification performance.

2 BACKGROUND

Diffusion Models. Given a sample $\mathbf{x}_0 \in \mathcal{X} \subseteq \mathbb{R}^d$ from the real data distribution with density $q(\mathbf{x}_0)$, the forward diffusion process injects Gaussian noise to generate a sequence of latent variables $\{\mathbf{x}_t\}_{t=1}^T$. This transition is defined by the conditional Gaussian density $\mathcal{N}$:

$$q(\mathbf{x}_t \,|\, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \alpha_t \mathbf{x}_0, \sigma_t^2 \mathbf{I}), \tag{1}$$

where $\alpha_t > 0$ and $\sigma_t > 0$ are the signal scaling and noise standard deviation schedules, respectively, and $\mathbf{I}$ represents the identity matrix. The reverse generative process aims to invert this corruption by learning a denoising distribution $p_\theta(\mathbf{x}_{t-1} \,|\, \mathbf{x}_t)$ to recover $\mathbf{x}_0$ from a pure noise prior $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, where $\mathbf{0}$ denotes the zero vector. This is typically achieved by maximizing the evidence lower bound (ELBO) on the log-likelihood $\log p_\theta(\mathbf{x}_0)$. By parameterizing the reverse step with a score network $\mathbf{s}_\theta(\mathbf{x}_t, t)$, the ELBO simplifies to a weighted denoising objective:

$$\mathcal{L}(\theta) = \mathbb{E}_{t, \mathbf{x}_t}\left[w_t \left\| \mathbf{s}_\theta(\mathbf{x}_t, t) - \nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t) \right\|_2^2 \right], \tag{2}$$

where $\mathbb{E}_{t,\mathbf{x}_t}$ is the expectation over time t and noisy samples $\mathbf{x}_t$, w_t is a weighting function, and $\nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t)$ is the gradient of the log-density of the perturbed data density $q_t(\mathbf{x}_t)$. In practice, the intractable true score is replaced by the analytic conditional score $\nabla_{\mathbf{x}_t} \log q(\mathbf{x}_t \,|\, \mathbf{x}_0) = (\alpha_t \mathbf{x}_0 - \mathbf{x}_t)/\sigma_t^2$. By varying schedules α_t, σ_t, w_t, this framework unifies Denoising Diffusion Probabilistic Models (DDPM) (Ho et al., 2020), Score-based Stochastic Differential Equations (SDEs) (Song et al., 2020), and Elucidated Diffusion Models (EDM) (Karras et al., 2022) (see Appendix F.1). In this work, we adopt the EDM formulation as our default backbone for its well-optimized parameterization.

Weak Annotations. In real-world settings, the true label $Y \in \mathcal{Y} = \{1, \ldots, c\}$ is often latent. Instead, we observe a weak annotation Z, which conveys imperfect information about Y. We denote the dataset as $\mathcal{D} = \{(\mathbf{x}_i, z_i)\}_{i=1}^N$. We focus on three representative weak annotation scenarios:

- *Noisy Labels (Han et al., 2020):* The observed label $Z = \hat{Y}$ is a corrupted version of the true label Y, modeled by a noise transition distribution $p(\hat{Y} \,|\, X, Y)$.

- *Ambiguous Labels (Tian et al., 2023):* Each instance is annotated with a candidate set $Z = S \subseteq \mathcal{Y}$ that contains the true label, i.e., $Y \in S$.
- *Semi-supervised Data (Yang et al., 2022):* The annotation Z is either a precise label Y for labeled data, or an empty set $\emptyset$ indicating unlabeled data, i.e., $Z \in \mathcal{Y} \cup \{\emptyset\}$.

3 METHODOLOGY

In this section, we present a unified probabilistic framework for training class-conditional diffusion models using weak annotations. We derive the learning objective from a maximum likelihood perspective, and decompose it into generative and classification components.

3.1 UNIFIED PROBABILISTIC FRAMEWORK

Given a dataset with inputs X and weak annotations Z, we treat the unobserved true label Y as a latent variable. Our goal is to find the optimal model parameters θ that maximize the log-likelihood of the observed joint observation (X, Z):

$$\theta^* = \arg\max_\theta \log p_\theta(X, Z) = \arg\max_\theta \log \sum\nolimits_{Y \in \mathcal{Y}} p_\theta(X, Y, Z). \tag{3}$$

Since Eq. (3) involves a summation over latent variable Y inside the logarithm, direct optimization is intractable when the size of $\mathcal{Y}$ is large. To address this, we maximize a variational lower bound on the log-likelihood:

$$\begin{aligned} \hat{\theta} &= \arg\max_\theta \, \mathbb{E}_{p_\phi(Y|X,Z)} \big[\log p_\theta(X, Y, Z) \big] \\ &= \arg\max_\theta \Big\{ \underbrace{\mathbb{E}_{p_\phi(Y|X,Z)} \big[\log p_\theta(X \,|\, Z) \big]}_{\text{Generative Term}} + \underbrace{\mathbb{E}_{p_\phi(Y|X,Z)} \big[\log p_\theta(Y \,|\, X, Z) \big]}_{\text{Classification Term}} \Big\}, \end{aligned} \tag{4}$$

where $\hat{\theta}$ denotes the optimized parameters, and $p_\phi(Y \mid X, Z)$ serves as a proxy for the true class-posterior probability, controlled by parameters ϕ. Detailed derivation of Eq. (4) is provided in Appendix F.6. In practice, we set ϕ as the exponential moving average (EMA) of θ to stabilize training, leveraging the predictions from the EMA model as soft targets. We adopt the standard class-conditional assumption (Liu et al., 2023), where the generation of the weak annotation Z only depends on the true label Y. Eq. (4) effectively decomposes the learning objective into two synergistic parts: a *Generative Term* that trains the diffusion model to capture the data distribution conditioned on latent labels, and a *Classification Term* that refines the class-posterior probability.

3.2 GENERATIVE LEARNING VIA POSTERIOR-WEIGHTED SCORE MATCHING

In this subsection, we focus on optimizing the *Generative Term* in Eq. (4). Our ultimate goal is to learn a class-conditional score network, denoted as $\mathbf{s}_\theta(\mathbf{x}_t, y, t)$, which approximates the clean-conditional score $\nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t \,|\, y)$, leveraging only weakly annotated pairs $(\mathbf{x}, z)$.

Mathematically, since the generative term $\log p_\theta(X \mid Z)$ is independent of the latent variable Y, the expectation $\mathbb{E}_{p_\phi}$ in Eq. (4) could theoretically be omitted. This would simplify the objective to directly maximizing the weak-conditional likelihood $\log p_\theta(X \,|\, Z)$. However, directly optimizing this term corresponds to learning the weak-label score $\nabla \log q_t(\mathbf{x}_t \,|\, z)$, which, as we analyze later in the mixture bias of naive training, represents a mixture density that leads to class mismatch and homogenization. To resolve this, we need to disentangle the clean signal from this mixture. Inspired by the relationships between noisy and clean distributions explored in noise-robust diffusion training (Na et al., 2024), we formally establish the link between the observable *weak-conditional score* $\nabla \log q_t(\mathbf{x}_t \,|\, z)$ and the latent *clean-conditional score* $\nabla \log q_t(\mathbf{x}_t \,|\, y)$.

Theorem 1 (Score Decomposition under Weak Annotations)**.** *Under the class-conditional independence assumption ($X \perp Z \mid Y$), the score function of the perturbed data density conditioned on a weak annotation z decomposes as the expectation of clean-label scores, weighted by the class-posterior probability:*

$$\nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t \,|\, z) = \sum\nolimits_{y=1}^{c} \, p_\phi(y \,|\, \mathbf{x}_t, z) \, \nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t \,|\, y). \tag{5}$$

The proof is provided in Appendix F.2. Theorem 1 reveals that the observable weak gradient is simply a posterior-weighted combination of the underlying clean gradients. This decomposition implies that by correctly estimating the weights $p_\phi(y \mid \mathbf{x}_t, z)$, we can disentangle and learn the clean-conditional densities from weak annotations.

Posterior-Weighted Score Matching. Guided by Theorem 1, we construct our generative objective. Instead of fitting the weak-label score directly, we parameterize the model to predict the clean-label score $\mathbf{s}_\theta(\mathbf{x}_t, y, t)$, and enforce its posterior-weighted aggregation to match the target signal derived from the data. We substitute the decomposition in Eq. (5) into the standard score matching objective, yielding the posterior-weighted score matching loss Eq. (6). In Eq. (6), $\nabla_{\mathbf{x}_t} \log q_{t|0}(\mathbf{x}_t \mid \mathbf{x}_0, z)$ is the analytic conditional score of the forward process, which serves as the training target. This formulation allows the model to disentangle the clean-label score from the weak annotation. For instance, if the posterior $p(y \mid \mathbf{x}_t, z)$ is sharp (confident), the gradient primarily updates the corresponding specific class; if it is flat (uncertain), the gradient is distributed, preventing the model from over-committing to a potentially wrong label.

$$\mathcal{L}_{\text{Gen}}(\theta) = \mathbb{E}_{t,\mathbf{x}_t,z}\left[w_t \left\| \sum\nolimits_{y=1}^{c} p_\phi(y \mid \mathbf{x}_t, z)\, \mathbf{s}_\theta(\mathbf{x}_t, y, t) - \nabla_{\mathbf{x}_t} \log q_{t|0}(\mathbf{x}_t \mid \mathbf{x}_0, z) \right\|_2^2\right]. \tag{6}$$

Analysis: The Mixture Bias of Naive Training. To illustrate the necessity of Eq. (6), we consider a naive approach that trains a CDMs directly on pairs $(\mathbf{x}, z)$. Such an objective would minimize $\mathbb{E}_{t,\mathbf{x}_t,z}[w_t \|\mathbf{s}_\theta(\mathbf{x}_t, z, t) - \nabla \log q_{t|0}(\mathbf{x}_t \mid \mathbf{x}_0, z)\|_2^2]$. As discussed in Remark 1 in Appendix E.1, the optimal solution would converge to the mixture score $\nabla \log q_t(\mathbf{x}_t \mid z)$. Generating from this mixture forces the model to average over different signals, typically resulting in class mismatch or severe homogenization, failing to faithfully recover the distinct features of the true class y. In contrast, our generation objective guarantees the recovery of the true underlying densities:

Proposition 1. *The global minimizer θ^* of the objective in Eq. (6) satisfies*

$$\mathbf{s}_{\theta^*}(\mathbf{x}_t, y, t) = \nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t \mid y), \quad \forall y \in \mathcal{Y}. \tag{7}$$

Proposition 1 (proved in Appendix F.3) confirms that our method statistically identifies the true clean-label scores, fundamentally resolving the issue of weak annotations.

3.3 Classification via Diffusion-Based Posterior Inference

Effective generative learning in Eq. (6) and classification rely on accurate class-posterior probabilities: $p_\phi(y \mid \mathbf{x}_t, z)$ for weighting and $p_\theta(y \mid \mathbf{x}, z)$ for prediction. To infer these without introducing an external classifier, we employ the diffusion model itself as a zero-shot classifier.

Diffusion Classifier[1]. By applying Bayes' rule, the class-posterior probability $p_\theta(y \mid \mathbf{x}_t)$ can be computed by normalizing the joint density over all classes. It is equivalent to a softmax operation:

$$p_\theta(y \mid \mathbf{x}_t) = \frac{\exp\{\log p_\theta(\mathbf{x}_t \mid y) + \log p(y)\}}{\sum_{y'} \exp\left(\log p_\theta(\mathbf{x}_t \mid y') + \log p(y')\right)}. \tag{8}$$

To evaluate the likelihood term $\log p_\theta(\mathbf{x}_t \mid y)$, following Chen et al. (2024b), we employ the conditional ELBO approximation given by

$$\log p_\theta(\mathbf{x}_t \mid y) \approx -\int_{u(t+1)}^{u(T-1)} w_\tau \hbar(\tau, y)\, \mathrm{d}\tau, \qquad \hbar(\tau, y) = \mathbb{E}_{\mathbf{x}_\tau}\left[\left\|\mathbf{h}_\theta(\mathbf{x}_\tau, y, \tau) - \mathbf{x}_0\right\|_2^2\right], \tag{9}$$

where the expectation $\mathbb{E}_{\mathbf{x}_\tau}$ is taken over $\mathbf{x}_\tau \sim q(\mathbf{x}_\tau \mid \mathbf{h}_\theta(\mathbf{x}_t, y, t))$, $\mathbf{h}_\theta(\cdot)$ denotes a denoised data estimator, and w_τ follows a log-normal distribution.[2] Although we adopt a uniform prior in our main experiments, the formulation above explicitly demonstrates that non-uniform priors simply

[1] For brevity, we detail the derivation using θ, though the same applies to the EMA parameters ϕ.

[2] Specifically, $w_\tau = \exp\left\{-(\log \sigma_\tau - P_{\text{mean}})^2/(2P_{\text{std}}^2)\right\} \cdot (\sigma_\tau P_{\text{std}}\sqrt{2\pi})^{-1}$, and $u(t) = ({\sigma_{\max}}^{1/\rho} + t(T-1)^{-1}({\sigma_{\min}}^{1/\rho} - {\sigma_{\max}}^{1/\rho}))^\rho$. Following EDM (Karras et al., 2022), we set $P_{\text{mean}} = -1.2$, $P_{\text{std}} = 1.2$, $\sigma_{\min} = 0.002$, $\sigma_{\max} = 80$, $\rho = 7$, and $\sigma_\tau = \tau$.

function as additive biases to the logits. For scenarios requiring non-uniform priors, $p(y)$ can be readily estimated from the training set (Tarvainen & Valpola, 2017; Liu et al., 2020).

We denote the resulting normalized posterior probabilities as $f_\theta(\mathbf{x}_t) \in \Delta^{c-1}$, where the y-th entry is $f_\theta(\mathbf{x}_t)_y = p_\theta(y \mid \mathbf{x}_t)$. This formulation allows us to update the class-posterior probabilities dynamically as the model trains.

Classification Objectives under Weak Annotations. To optimize the *Classification Term* in Eq. (4), we recall the factorization from Sec. 3.1 that $p_\theta(y \mid \mathbf{x}, z) \propto p(z \mid y) p_\theta(y \mid \mathbf{x})$. Crucially, since the transition probability $p(z \mid y)$ is independent of the model parameters θ, the optimization reduces to

$$\mathcal{L}_{\text{Cls}}(\theta) = -\sum\nolimits_{y \in \mathcal{Y}} p_\phi(y \mid \mathbf{x}, z) \log p_\theta(y \mid \mathbf{x}). \tag{10}$$

This objective minimizes the cross-entropy weighted by the posterior $p_\phi(y \mid \mathbf{x}, z)$, creating a self-training loop where the model refines its own supervision using signals derived from the EMA model. We instantiate this objective for three specific forms of weak annotations z as follows:

1) Noisy Labels ($Z = \hat{Y}$). In this setting, deriving the exact posterior $p_\phi(y \mid \mathbf{x}, \hat{y})$ is intractable as the noise transition distribution is unknown. Instead of explicitly estimating the transition distribution, we employ the early learning regularization (ELR) principle (Liu et al., 2020), which exploits the observation that neural networks fit clean patterns before memorizing noise. We propose a robust proxy for $p_\phi(y \mid \mathbf{x}, \hat{y})$ that amplifies contributions from clean samples while suppressing noisy ones. Specifically, we instantiate it using a dynamically rectified target formulated as Eq. (11), where $\hat{\mathbf{y}}$ is the one-hot vector of $\hat{y}$, $\delta = \langle f_\theta(\mathbf{x}), f_\phi(\mathbf{x}) \rangle$ is the inner product, $\odot$ is the Hadamard product, and $\langle \cdot, \cdot \rangle$ denotes the inner product. The operator $\text{sg}(\cdot)$ is the stop-gradient operation. This formulation inherently prevents label noise memorization, which is analyzed in Appendix E.2.

$$\mathcal{L}_{\text{Cls}}^{\text{NL}}(\mathbf{x}) = -\sum_{y \in \mathcal{Y}} p_\phi(y \mid \mathbf{x}, \hat{y}) \log f_\theta(\mathbf{x})_y, \quad p_\phi(\cdot \mid \mathbf{x}, \hat{y}) = \text{sg}\left(\hat{\mathbf{y}} - \frac{f_\theta(\mathbf{x}) \odot \left(\delta \mathbf{1} - f_\phi(\mathbf{x})\right)}{1 - \delta}\right), \tag{11}$$

2) Ambiguous Labels ($Z = S$). Here, the ground-truth label is restricted to a candidate set $S \subseteq \mathcal{Y}$. For a specific instance $\mathbf{x}$ with an observed candidate set s, since the true label is guaranteed to reside within s, any label outside this set has strictly zero probability. Accordingly, we instantiate the posterior weight $p_\phi(y \mid \mathbf{x}, s)$ by masking out invalid classes and re-normalizing over valid candidates as Eq. (12), where $\mathbb{I}(\cdot)$ denotes the indicator function. This acts as an EMA-stabilized progressive identification (Lv et al., 2020), enabling true label disambiguation from s.

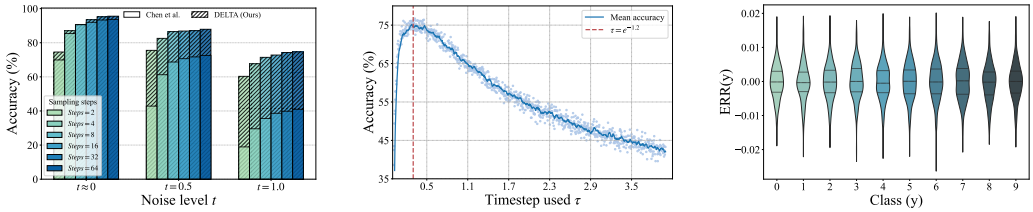
$$\mathcal{L}_{\text{Cls}}^{\text{AL}}(\mathbf{x}) = -\sum_{y \in \mathcal{Y}} p_\phi(y \mid \mathbf{x}, s) \log f_\theta(\mathbf{x})_y, \quad p_\phi(y \mid \mathbf{x}, s) = \mathbb{I}(y \in s) \Big(\sum_{y' \in s} f_\phi(\mathbf{x})_{y'} \Big)^{-1} f_\phi(\mathbf{x})_y, \tag{12}$$

3) Semi-Supervised Data ($Z \in \mathcal{Y} \cup \{\emptyset\}$). In this setting, the annotation z is either a precise label or missing. We instantiate the posterior weight $p_\phi(y \mid \mathbf{x}, z)$ using a composite indicator formulation that adaptively switches between supervision sources as Eq. (13), where $\gamma = \mathbb{I}(z \in \mathcal{Y})$ indicates labeled data. This derivation effectively recovers the self-training objective (Tarvainen & Valpola, 2017), where the model leverages EMA-guided soft-labels on unlabeled data.

$$\mathcal{L}_{\text{Cls}}^{\text{SS}}(\mathbf{x}) = -\sum_{y \in \mathcal{Y}} p_\phi(y \mid \mathbf{x}, z) \log f_\theta(\mathbf{x})_y, \quad p_\phi(y \mid \mathbf{x}, z) = \gamma \, \mathbb{I}(y = z) + (1 - \gamma) \, f_\phi(\mathbf{x})_y, \tag{13}$$

3.4 Efficient Posterior Estimation via Optimized Timestep Sampling

The diffusion classifier Eq. (8) requires repeated calculation of the conditional ELBO across all classes and all sampled timesteps, leading to substantial computational costs. To address this, Chen et al. (2024b) adopted a shared $\mathbf{x}_\tau$ strategy across classes and sampled timesteps uniformly from $[u(t+1), u(T-1)]$. However, Figure 1(a) indicates that uniform sampling is empirically suboptimal. To understand why, we evaluate the classification accuracy using only a single timestep in Figure 1(b). The results reveal that the model's discriminative power is highly non-uniform: early steps induce negligible noise yielding trivial reconstructions with low label sensitivity, whereas later



(a) Results across different t. (b) Results with a single timestep. (c) Class-wise ERR distribution.

Figure 1: (a): Test accuracy comparison on CIFAR-10 dataset under the time complexity reduction technique from Chen et al. (2024c) and ours. (b): Test accuracy on CIFAR-10 dataset evaluated with only a single timestep per class. (c): Class-wise distribution of the optimality residual ERR (Theorem 3). Presented as a violin plot where wider regions indicate higher density

steps are dominated by noise rendering predictions unreliable. Consequently, uniform sampling wastes significant computation on these uninformative regions. This motivates us to concentrate evaluation on the most effective timestep interval.

Optimal Timestep Sampling. We aim to identify a compact subinterval $[l, r]$, which minimizes the discrepancy between the expected reconstruction error on the subinterval $[l, r]$ and the original interval $[u(t+1), u(T-1)]$:

$$\left\| \frac{1}{W_{[l,r]}} \int_l^r w_\tau \hbar(\tau, y)\, \mathrm{d}\tau - \frac{1}{W_{\mathcal{T}}} \int_{\mathcal{T}} w_\tau \hbar(\tau, y)\, \mathrm{d}\tau \right\|_2^2, \tag{14}$$

where $\mathcal{T} = [u(t+1), u(T-1)]$ and $W_{[a,b]} = \int_a^b w_\tau d\tau$ is the total weight mass within an interval $[a, b]$. Since the loss landscape $\hbar(\tau, y)$ is data-dependent and evolves during training, a robust strategy is to maximize the covered weight mass $W_{[l,r]}$ given a fixed length Δ, ensuring evaluation focuses on the most informative region. For the log-normal weight density used in EDM, this maximization is equivalent to centering the interval around the median of the weight density, as formalized below.

Theorem 2 (Median-Centered Interval via Weight Mass)**.** *Consider the EDM log-normal weighting function $w(\tau)$ with cumulative weight function $\mathcal{F}_w(b) = \int_0^b w(\tau) d\tau$. Given a target interval length Δ, the optimal range $[l, r]$ that maximizes the covered weight mass is determined by*

$$\mathcal{F}_w(l) + \mathcal{F}_w(l + \Delta) = 1, \qquad \textit{with} \quad r = l + \Delta. \tag{15}$$

The optimal l can be computed via numerical root-finding (Brent, 2013). The proof of Theorem 2 as well as a similar conclusion for DDPM (Ho et al., 2020) are provided in Appendix F.4. Notably, as the interval shrinks to a point, the optimal choice converges to the weight density median $e^{-1.2} \approx 0.3$. This justifies our finding in Figure 1(b) that median timesteps yield peak performance. Furthermore, we derive a necessary condition for any theoretically optimal subinterval (l^*, r^*) minimizing the objective Eq. (14):

Theorem 3 (Necessary Optimality Condition)**.** *A necessary condition for the optimal subinterval is that the expected error within the interval equals the average error at the boundaries:*

$$\int_{l^*}^{r^*} \frac{w_\tau \hbar(\tau, y)}{W_{[l^*, r^*]}}\, \mathrm{d}\tau = \frac{\hbar(l^*, y) + \hbar(r^*, y)}{2}. \tag{16}$$

The proof is provided in Appendix F.5. To empirically validate our timestep sampling strategy given a fixed interval length Δ, we analyze the class-wise residual error $\mathrm{ERR}(y)$ (defined as the difference between the LHS and RHS of Theorem 3). As shown in Figure 1(c), the class-wise distribution of $\mathrm{ERR}(y)$ is naturally concentrated around zero, conforming to the theoretical optimality condition. In practice, we combine this median-centered timestep sampling strategy with the noise reuse technique (Chen et al., 2024c) to maximize the computational efficiency.

4 EXPERIMENTS

We evaluate our framework's versatility across image generation and weakly supervised classification. To comprehensively assess its robustness, we conduct experiments on two categories of

Table 1: Generative results under various weak annotations. The gray-shaded 'Clean' column represents the best performance that is trained on the fully clean dataset. 'uncond'/'cond' denote unconditional/conditional metrics. **Bold** numbers indicate better performance.

		Metric		Noisy label				Ambiguous label				Semi-supervised data				Clean
				Sym-40%		Asym-40%		Random		Class-Dept.		1% Labels		10% Labels		
				Vanilla	*DELTA*	*Vanilla*	*DELTA*	*Vanilla*	*DELTA*	*Vanilla*	*DELTA*	*Vanilla*	*DELTA*	*Vanilla*	*DELTA*	
CIFAR-10	uncond	FID	(↓)	**3.33**	3.47	3.23	**3.10**	7.76	**2.26**	11.75	**2.77**	3.16	**3.12**	2.93	**2.89**	2.05
		IS	(↑)	9.56	**9.68**	9.02	**9.73**	9.09	**9.80**	9.62	**9.68**	10.03	**10.57**	9.80	**9.83**	10.61
		Density	(↑)	101.39	**109.75**	100.06	**109.69**	103.21	**106.49**	108.76	**109.06**	97.19	**108.18**	99.96	**108.87**	112.59
		Coverage	(↑)	81.12	**81.21**	80.71	**81.30**	68.45	**82.69**	64.90	**81.52**	78.44	**81.00**	81.85	**82.00**	83.27
	cond	CW-FID	(↓)	29.84	**13.85**	14.70	**13.24**	27.18	**10.65**	32.44	**11.56**	16.25	**16.12**	11.84	**11.77**	9.83
		CW-Density	(↑)	72.98	**107.23**	90.85	**107.07**	102.04	**105.75**	102.43	**108.66**	89.99	**100.73**	96.29	**107.94**	111.70
		CW-Coverage	(↑)	73.39	**80.11**	79.63	**79.65**	65.45	**82.09**	61.45	**81.24**	75.03	**76.84**	80.80	**81.12**	83.91
ImageNette	uncond	FID	(↓)	14.11	**13.44**	13.93	**13.91**	79.13	**72.62**	91.28	**79.12**	23.88	**19.26**	14.32	**12.84**	11.52
		IS	(↑)	12.69	**13.21**	12.51	**13.73**	9.19	**9.40**	**9.27**	9.11	12.23	**13.72**	12.80	**13.16**	13.81
		Density	(↑)	109.31	**112.52**	**111.66**	106.78	95.33	**99.83**	94.29	**102.58**	115.94	**125.68**	105.27	**109.23**	117.23
		Coverage	(↑)	76.62	**76.81**	78.32	**79.81**	21.44	**32.48**	16.69	**22.30**	53.53	**55.39**	73.79	**75.55**	80.12
	cond	CW-FID	(↓)	80.31	**60.12**	62.26	**58.20**	157.76	**63.58**	163.45	**67.92**	71.66	**70.27**	49.22	**44.31**	40.20
		CW-Density	(↑)	73.99	**81.12**	93.53	**94.58**	93.38	**95.83**	91.50	**95.21**	115.90	**118.69**	103.41	**115.67**	120.35
		CW-Coverage	(↑)	67.89	**71.94**	74.18	**75.82**	19.76	**24.35**	15.88	**18.93**	51.73	**52.15**	72.61	**74.85**	78.48

datasets: (1) **Synthetic benchmarks**, where we systematically apply controlled corruptions to Fashion-MNIST (Xiao et al., 2017), CIFAR-10 (Krizhevsky et al., 2009), and ImageNette (Deng et al., 2009); (2) **Real-world benchmarks** with naturally weak annotations, including noisy-label datasets (CIFAR10-N, CIFAR100-N (Wei et al., 2022)) and an ambiguous-label dataset (PLCIFAR10 (Wang et al., 2025c)). For ablation, we construct a *Vanilla* baseline trained with the identical joint objective structure but replacing our posterior-weighted generative term in Eq. (6) with the naive formulation $\mathbb{E}_{t,\mathbf{x}_t,z}[w_t \|\mathbf{s}_\theta(\mathbf{x}_t, z, t) - \nabla \log q_{t|0}(\mathbf{x}_t \,|\, \mathbf{x}_0, z)\|_2^2]$. Our training configurations follow the EDM framework. Detailed implementation specifics and additional experimental results are provided in Appendix C and Appendix D, respectively.

Weak Annotation Protocols for Synthetic Benchmarks. We corrupt benchmark datasets as follows: (1) **Noisy Label**: labels are flipped with 40% probability via symmetric noise (*'Sym-40%'*, flipping uniformly to any incorrect class) or asymmetric noise (*'Asym-40%'*, flipping to semantically similar classes); (2) **Ambiguous Label**: candidate sets are generated via a random policy (*'Random'*, including non-target labels with 50% probability) or a class-dependent policy (*'Class-Dept.'*, selecting semantically similar labels with specific probabilities); (3) **Semi-Supervised Data**: we randomly retain a specific fraction of labeled samples per class (denoted as *'1% Labels'* and *'10% Labels'*).

4.1 Task 1: Image Generation

Setup. Evaluation relies on four unconditional metrics (Chao et al., 2022): Fréchet Inception Distance (FID), Inception Score (IS), Density, and Coverage. Additionally, to assess class-consistency, we employ three Class-Wise (CW) metrics (Kaneko et al., 2019): CW-FID, CW-Density, and CW-Coverage, computed per class and averaged. Detailed definitions of these metrics are in Appendix D.1.

Results. Table 1 shows *DELTA* consistently outperforms *Vanilla* across all metrics, with gains most pronounced under ambiguous labels, highlighting its ability to disambiguate latent labels. Qualitatively, Figure 5 provides a visual comparison of the generated samples. While *Vanilla* frequently generates samples with class mismatch due to annotation corruption, *DELTA* produces high-fidelity images that are semantically aligned with the target classes. This validates that our posterior-weighted score matching effectively resolves the mixture bias.

Table 2: Comparison against noise-robust diffusion baselines.

Method		*DELTA*	*CAD*	*TDSM*	*SBDC*	*RTDC*
Sym-40%	FID (↓)	**3.47±0.12**	4.10±0.15	3.85±0.10	4.25±0.20	4.00±0.18
	IS (↑)	**9.68±0.05**	9.08±0.07	9.40±0.06	8.95±0.08	9.20±0.09
Asym-40%	FID (↓)	**3.10±0.10**	3.87±0.12	3.96±0.11	4.02±0.15	3.85±0.13
	IS (↑)	9.73±0.06	9.16±0.07	**10.12±0.07**	9.05±0.08	9.30±0.09

To further validate our approach, Table 2 presents a comparison on CIFAR-10 with 40% symmetric and asymmetric noise between *DELTA* and several noise-robust diffusion baselines, including CAD (Dufour et al., 2024), TDSM (Na et al., 2024), SBDC (Cong et al., 2025), and RTDC (Chen et al., 2025). Importantly, these baseline methods typically assume access to auxiliary information (e.g., oracle confidence), whereas *DELTA* operates without such priors. Despite this, *DELTA* achieves competitive or superior performance, suggesting our method is effective while operating

Table 3: Classification results (test accuracy, %) under various weak annotations (♠: noisy label, ♡: ambiguous label, ♣: semi-supervised data). **Bold** numbers indicate the best performances.

Dataset	Type	*CE*	*Mixup*	*Coteaching*	*ELR*	*PENCIL*	*Vanilla*	*DELTA*
Fashion-MNIST♠	Sym-40%	76.18±0.26	92.21±0.03	92.17±0.34	93.13±0.13	90.85±0.58	90.11±1.24	**93.40±0.40**
	Asym-40%	82.01±0.06	92.01±1.02	92.78±0.25	92.82±0.09	91.77±0.69	85.41±0.96	**93.20±0.30**
CIFAR-10♠	Sym-40%	67.22±0.26	84.26±0.64	86.54±0.57	85.68±0.13	85.91±0.26	80.22±0.10	**88.63±0.12**
	Asym-40%	76.98±0.42	83.21±0.85	79.38±0.39	81.32±0.31	84.89±0.49	86.31±0.10	**88.83±0.33**
	Type	*PRODEN*	*IDGP*	*PiCO*	*CRDPLL*	*DIRK*	*Vanilla*	*DELTA*
Fashion-MNIST♡	Random	93.31±0.07	92.26±0.25	93.32±0.12	94.03±0.14	94.11±0.22	80.20±1.29	**94.27±0.55**
	Class-Dept.	93.44±0.21	93.07±0.16	93.32±0.33	93.80±0.23	93.99±0.24	66.03±1.43	**94.20±0.15**
CIFAR-10♡	Random	90.02±0.22	89.65±0.53	86.40±0.89	92.74±0.26	93.48±0.14	60.25±0.17	**94.70±0.49**
	Class-Dept.	90.44±0.44	90.83±0.34	87.51±0.66	92.89±0.27	93.22±0.37	56.34±0.50	**93.53±0.12**
	Type	*Dash*	*CoMatch*	*FlexMatch*	*SimMatch*	*SoftMatch*	*Vanilla*	*DELTA*
Fashion-MNIST♣	1% Labels	84.73±0.09	85.31±0.29	84.43±0.30	84.69±0.17	84.72±0.23	78.37±0.72	**85.92±0.13**
	10% Labels	91.16±0.20	90.52±0.12	90.69±0.03	91.18±0.13	91.22±0.11	90.50±1.00	**92.97±0.21**
CIFAR-10♣	1% Labels	70.14±0.69	61.45±1.46	70.72±0.93	73.33±1.02	73.74±0.82	53.49±0.15	**76.40±0.54**
	10% Labels	81.50±0.68	77.79±0.53	81.35±0.48	82.90±0.43	88.66±0.60	85.13±0.12	**92.47±0.39**

under strictly weaker assumptions about available supervision. Additionally, Table 4 validates *DELTA*'s effectiveness on real-world datasets with competitive FID and IS scores.

4.2 Task 2: Weakly Supervised Classification

Setup. We evaluate our method under three weakly supervised scenarios. For learning from noisy-label data, we compare with *Coteaching* (Han et al., 2018), *ELR* (Liu et al., 2020), *PENCIL* (Yi & Wu, 2019), standard cross-entropy (CE), and *Mixup* (Zhang et al., 2018). For learning from ambiguous-label data, we compare against *PRODEN* (Lv et al., 2020), *IDGP* (Qiao et al., 2023), *PiCO* (Wang et al., 2023), *CRDPLL* (Wu et al., 2022), and *DIRK* (Wu et al., 2024). For semi-supervised learning, we adopt *Dash* (Xu et al., 2021), *CoMatch* (Li et al., 2021), *FlexMatch* (Zhang et al., 2021), *SimMatch* (Zheng et al., 2022), and *SoftMatch* (Chen et al., 2023). To ensure fairness, the discriminative classifier is Wide-ResNet-40-10 with 55.84M parameters, while our generative model has 55.73M parameters. All models are trained from scratch without pre-training.

Table 4: Performance under real-world weak-annotated datasets.

Dataset	FID (↓)	IS (↑)	Acc (%)
CIFAR10-N	3.22	9.66	91.21
PLCIFAR10	2.95	9.82	93.65
CIFAR100-N	4.85	10.42	70.68

Table 5: Inference latency comparison during testing phase.

Metric	*Chen et al.*	*DELTA*
Time (min)	18.43	**0.67**
Acc (%)	87.13	**88.63**
Speedup	1.0×	**27.5**×

Results. Table 3 reports classification accuracies across three weak annotations. Our method *DELTA*, evaluated via a diffusion-based classifier, consistently achieves the best performance, demonstrating the stronger generalization capability of diffusion models over prior discriminative approaches. Furthermore, as shown in Table 4, *DELTA* maintains high accuracy on real-world datasets, verifying its robustness beyond synthetic benchmarks.

4.3 Inference Efficiency Comparison

To evaluate inference efficiency, we benchmark inference latency on CIFAR-10 using a single NVIDIA A100 GPU. As shown in Table 5, our optimized sampling strategy achieves a 27.5× speedup over Chen et al. (2025) without compromising accuracy. This efficiency advantage is further corroborated by Figure 1(a), where *DELTA* yields higher accuracy at identical sampling budgets, with the performance gap distinctively widening as the noise level increases.

5 Conclusion

In this paper, we proposed *DELTA*, a unified framework for robustly training CDMs under weak annotations via joint likelihood maximization. By synergizing posterior-weighted score matching with an efficient diffusion classifier, *DELTA* dynamically infers reliable class-posterior probabilities to disentangle clean generation from weak supervision. Extensive experiments across image generation and weakly supervised classification demonstrate its robustness and versatility in handling diverse weak annotations.

REFERENCES

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, pp. 5050–5060, 2019.

Richard P Brent. *Algorithms for minimization without derivatives*. Courier Corporation, 2013.

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.

Hao Chen, Ran Tao, Yue Fan, Yidong Wang, Jindong Wang, Bernt Schiele, Xing Xie, Bhiksha Raj, and Marios Savvides. Softmatch: Addressing the quantity-quality tradeoff in semi-supervised learning. In *International Conference on Learning Representations*, 2023.

Hao Chen, Ankit Shah, Jindong Wang, Ran Tao, Yidong Wang, Xiang Li, Xing Xie, Masashi Sugiyama, Rita Singh, and Bhiksha Raj. Imprecise label learning: A unified framework for learning with various imprecise label configurations. In *Advances in Neural Information Processing Systems*, pp. 59621–59654, 2024a.

Huanran Chen, Yinpeng Dong, Shitong Shao, Zhongkai Hao, Xiao Yang, Hang Su, and Jun Zhu. Diffusion models are certifiably robust classifiers. In *Advances in Neural Information Processing Systems*, pp. 50062–50097, 2024b.

Huanran Chen, Yinpeng Dong, Zhengyi Wang, Xiao Yang, Chengqi Duan, Hang Su, and Jun Zhu. Robust classification via a single diffusion model. In *Proceedings of the International Conference on Machine Learning*, pp. 6643–6665, 2024c.

Xin Chen, Gillian Dobbie, Xinyu Wang, Feng Liu, Di Wang, and Jingfeng Zhang. Robust learning of diffusion models with extremely noisy conditions. *arXiv preprint arXiv:2510.10149*, 2025.

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.

Jan Oscar Cross-Zamirski, Praveen Anand, Guy Williams, Elizabeth Mouchet, Yinhai Wang, and Carola-Bibiane Schönlieb. Class-guided image-to-image diffusion: Cell painting from brightfield images with class labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3800–3809, 2023.

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

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*, pp. 6264–6273, 2024.

Lei Feng, Jiaqi Lv, Bo Han, Miao Xu, Gang Niu, Xin Geng, Bo An, and Masashi Sugiyama. Provably consistent partial-label learning. In *Advances in Neural Information Processing Systems*, pp. 10948–10960, 2020.

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*, pp. 8536–8546, 2018.

Bo Han, Quanming Yao, Tongliang Liu, Gang Niu, Ivor W Tsang, James T Kwok, and Masashi Sugiyama. A survey of label-noise representation learning: Past, present and future. *arXiv preprint arXiv:2011.04406*, 2020.

Chunming He, Chengyu Fang, Yulun Zhang, Longxiang Tang, Jinfa Huang, Kai Li, Zhenhua Guo, Xiu Li, and Sina Farsiu. Reti-diff: Illumination degradation image restoration with retinex-based latent diffusion model. In *International Conference on Learning Representations*, 2025.

Wei He, Kai Han, Ying Nie, Chengcheng Wang, and Yunhe Wang. Species196: A one-million semi-supervised dataset for fine-grained species recognition. In *Advances in Neural Information Processing Systems*, pp. 44957–44975, 2023.

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*, pp. 6626–6637, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pp. 6840–6851, 2020.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*, pp. 8633–8646, 2022.

Takuhiro Kaneko, Yoshitaka Ushiku, and Tatsuya Harada. Label-noise robust generative adversarial networks. In *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*, pp. 26565–26577, 2022.

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

Alexander C Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2206–2217, 2023.

Junnan Li, Caiming Xiong, and Steven CH Hoi. Comatch: Semi-supervised learning with contrastive graph regularization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9475–9484, 2021.

Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool. Webvision database: Visual learning and understanding from web data. *arXiv preprint arXiv:1708.02862*, 2017.

Yangming Li, Max Ruiz Luyten, and Mihaela van der Schaar. Risk-sensitive diffusion: Robustly optimizing diffusion models with noisy samples. In *International Conference on Learning Representations*, 2025.

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*, pp. 20331–20342, 2020.

Yang Liu, Hao Cheng, and Kun Zhang. Identifiability of label noise transition matrix. In *Proceedings of the International Conference on Machine Learning*, pp. 21475–21496, 2023.

Jiaqi Lv, Miao Xu, Lei Feng, Gang Niu, Xin Geng, and Masashi Sugiyama. Progressive identification of true labels for partial-label learning. In *Proceedings of the International Conference on Machine Learning*, pp. 6500–6510, 2020.

Takeru Miyato and Masanori Koyama. cgans with projection discriminator. In *International Conference on Learning Representations*, 2018.

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.

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 International Conference on Machine Learning*, pp. 7176–7185, 2020.

Shuntaro Okada, Ryota Yoshihashi, Hirokatsu Kataoka, et al. Real-srgd: Enhancing real-world image super-resolution with classifier-free guided diffusion. In *Proceedings of the IEEE/CVF Asian Conference on Computer Vision*, pp. 3739–3755, 2024.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, 2019.

Congyu Qiao, Ning Xu, and Xin Geng. Decompositional generation process for instance-dependent partial label learning. In *International Conference on Learning Representations*, 2023.

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*, pp. 2226–2234, 2016.

Vinay Shukla, Zhe Zeng, Kareem Ahmed, and Guy Van den Broeck. A unified approach to count-based weakly supervised learning. In *Advances in Neural Information Processing Systems*, pp. 38709–38722, 2023.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, pp. 11895–11907, 2019.

Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. In *Advances in Neural Information Processing Systems*, pp. 12438–12448, 2020.

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*, 2020.

Masashi Sugiyama, Han Bao, Takashi Ishida, Nan Lu, and Tomoya Sakai. *Machine learning from weak supervision: An empirical risk minimization approach*. MIT Press, 2022.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016.

Hiroshi Takahashi, Tomoharu Iwata, Atsutoshi Kumagai, Yuuki Yamanaka, and Tomoya Yamashita. Positive-unlabeled diffusion models for preventing sensitive data generation. In *International Conference on Learning Representations*, 2025.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, pp. 1195–1204, 2017.

Yingjie Tian, Xiaotong Yu, and Saiji Fu. Partial label learning: Taxonomy, analysis and outlook. *Neural Networks*, 161:708–734, 2023.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

Haobo Wang, Ruixuan Xiao, Yixuan Li, Lei Feng, Gang Niu, Gang Chen, and Junbo Zhao. Pico+: Contrastive label disambiguation for robust partial label learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 3183–3198, 2023.

Hsiu-Hsuan Wang, Tan-Ha Mai, Nai-Xuan Ye, Wei-I Lin, and Hsuan-Tien Lin. Climage: Human-annotated datasets for complementary-label learning. *Transactions on Machine Learning Research*, 2025, 2025a.

Qimin Wang, Xinda Liu, and Guohua Geng. Guidpaint: Class-guided image inpainting with diffusion models. *arXiv preprint arXiv:2507.21627*, 2025b.

Wei Wang, Dong-Dong Wu, Jindong Wang, Gang Niu, Min-Ling Zhang, and Masashi Sugiyama. Realistic evaluation of deep partial-label learning algorithms. In *International Conference on Learning Representations*, 2025c.

Yanghao Wang and Long Chen. Noise matters: Optimizing matching noise for diffusion classifiers. In *Advances in Neural Information Processing Systems*, 2025.

Yidong Wang, Hao Chen, Yue Fan, Wang Sun, Ran Tao, Wenxin Hou, Renjie Wang, Linyi Yang, Zhi Zhou, Lan-Zhe Guo, et al. Usb: A unified semi-supervised learning benchmark for classification. In *Advances in Neural Information Processing Systems*, pp. 3938–3961, 2022.

Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. In *International Conference on Learning Representations*, 2021.

Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. In *International Conference on Learning Representations*, 2022.

Zixi Wei, Lei Feng, Bo Han, Tongliang Liu, Gang Niu, Xiaofeng Zhu, and Heng Tao Shen. A universal unbiased method for classification from aggregate observations. In *Proceedings of the International Conference on Machine Learning*, pp. 36804–36820, 2023.

Dong-Dong Wu, Deng-Bao Wang, and Min-Ling Zhang. Revisiting consistency regularization for deep partial label learning. In *Proceedings of the International Conference on Machine Learning*, pp. 24212–24225, 2022.

Dong-Dong Wu, Deng-Bao Wang, and Min-Ling Zhang. Distilling reliable knowledge for instance-dependent partial label learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 15888–15896, 2024.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Zheng Xie, Yu Liu, Hao-Yuan He, Ming Li, and Zhi-Hua Zhou. Weakly supervised auc optimization: A unified partial auc approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7):4780–4795, 2024.

Yi Xu, Lei Shang, Jinxing Ye, Qi Qian, Yu-Feng Li, Baigui Sun, Hao Li, and Rong Jin. Dash: Semi-supervised learning with dynamic thresholding. In *Proceedings of the International Conference on Machine Learning*, pp. 11525–11536, 2021.

Xiangli Yang, Zixing Song, Irwin King, and Zenglin Xu. A survey on deep semi-supervised learning. *IEEE transactions on knowledge and data engineering*, 35(9):8934–8954, 2022.

Kun Yi and Jianxin Wu. Probabilistic end-to-end noise correction for learning with noisy labels. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7017–7025, 2019.

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 *Advances in Neural Information Processing Systems*, pp. 43479–43495, 2023.

Weichen Yu, Ziyan Yang, Shanchuan Lin, Qi Zhao, Jianyi Wang, Liangke Gui, Matt Fredrikson, and Lu Jiang. Is your text-to-image model robust to caption noise? *arXiv preprint arXiv:2412.19531*, 2024.

Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. In *Advances in Neural Information Processing Systems*, pp. 18408–18419, 2021.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

Yivan Zhang, Nontawat Charoenphakdee, Zhenguo Wu, and Masashi Sugiyama. Learning from aggregate observations. In *Advances in Neural Information Processing Systems*, pp. 7993–8005, 2020.

Mingkai Zheng, Shan You, Lang Huang, Fei Wang, Chen Qian, and Chang Xu. Simmatch: Semi-supervised learning with similarity matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14471–14481, 2022.

Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National science review*, 5(1): 44–53, 2018.

CONTENTS

A Notation and Definitions

We present the notation table for each symbol used in this paper in Table 6.

Table 6: List of common mathematical symbols used in this paper.

Symbol	Definition
$\mathbf{x}$	A sample of training data
y	Clean (ground-truth, or true) class label
z	Weak annotation associated with a sample
$\hat{y}$	Noisy class label
s	Ambiguous label set ($s \subseteq \mathcal{Y}$)
c	Total number of classes
$\mathcal{X}$	Input space ($\mathbf{x} \in \mathcal{X}$)
$\mathcal{Y}$	Label space ($y \in \mathcal{Y}$)
$\mathcal{D}$	Training dataset
X	Random variable for data instances
Y	Random variable for clean labels
Z	Random variable for weak annotations
$\hat{Y}$	Random variable for noisy labels
S	Random variable for ambiguous label sets ($S \subseteq \mathcal{Y}$)
θ	Learnable parameters of the diffusion model
ϕ	Exponential moving average of θ
$\mathbf{I}$	Identity matrix
$\mathbf{x}_t$	Latent state (noisy data) at timestep t
t	Discrete timestep index ($t \in \{0, \dots, T\}$)
τ	Continuous time variable
α_t, σ_t	Noise schedule parameters at timestep t
Δ	Length of a subsampled timestep interval
$q(\mathbf{x})$	Data distribution
$q(y \mid \mathbf{x})$	True class posterior distribution
$p_\theta(\mathbf{x})$	Model distribution parameterized by θ
$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$
$\mathbf{s}_\theta(\mathbf{x}_t, t)$	Score prediction network
$f_\theta(\mathbf{x})$	Diffusion classifier function with parameters θ
$f_\phi(\mathbf{x})$	Diffusion classifier function with parameters ϕ
$\mathcal{L}_{\text{Gen}}(\theta)$	Generative objective (Posterior-Weighted Score Matching)
$\mathcal{L}_{\text{Cls}}(\theta)$	Classification objective
$\mathcal{L}_{\text{Cls}}^{\text{NL}}(\mathbf{x})$	Classification loss for noisy-label data
$\mathcal{L}_{\text{Cls}}^{\text{AL}}(\mathbf{x})$	Classification loss for ambiguous-label data
$\mathcal{L}_{\text{Cls}}^{\text{SS}}(\mathbf{x})$	Classification loss for semi-supervised data

B Related Work

B.1 Noise-robust Diffusion Models

Standard conditional diffusion models (Ho et al., 2020; Ho & Salimans, 2022) typically rely on high-quality supervision. However, real-world data frequently involves *weak annotations* (Zhou, 2018), such as noisy, ambiguous labels, or semi-supervised data, which can bias the learned conditional score and degrade generation controllability.

To address label noise, a representative line of work explicitly models the corruption process. For instance, Na et al. (2024) utilizes a noise transition matrix to treat the noisy-conditional score as a mixture of clean-class scores, optimizing a reweighted denoising objective. Other approaches focus on optimization stability; for example, Risk-Sensitive Diffusion (Li et al., 2025) formulates a risk-parameterized SDE to identify stability periods where noisy samples can safely optimize the score model. Instead of modeling the noise distribution, another category of methods aims to repair or filter the conditioning signals. Chen et al. (2025) propose replacing noisy conditions with trainable pseudo-conditions, refined via temporal ensembling and early stopping to stabilize training. Alternatively, to avoid trusting all conditions uniformly, Coherence-Aware Diffusion (Dufour et al., 2024) introduces an auxiliary coherence signal to down-weight unreliable conditioning. Similar reweighting strategies appear in text-to-image generation, where Yu et al. (2024) treat caption hallucinations as token-level noise, using confidence scores to modulate the contribution of unreliable tokens. For inference-time solutions, Score-based Discriminator Correction (SBDC) (Cong et al., 2025) trains a separate lightweight discriminator to distinguish correct conditions and injects a correction term into the sampling trajectory without retraining the backbone.

Research has also extended robustness to other forms of weak annotations. In semi-supervised settings, You et al. (2023) leverages large unlabeled datasets via consistency regularization to improve performance when labels are scarce. In safety-oriented positive-unlabeled settings, Takahashi et al. (2025) propose suppressing sensitive concepts using only unlabeled data and a small positive set, avoiding the need for exhaustive negative labels.

DELTA is complementary to these approaches. Unlike methods that require explicit noise transition matrices, external discriminators for post-hoc correction, or heuristic condition filtering, *DELTA* treats clean labels as *latent variables*. By decomposing the clean-conditional score into a posterior-weighted combination of observable scores, our method provides a unified framework specifically designed for diverse weak annotations, including noisy, ambiguous, and semi-supervised data, while remaining fully compatible with standard diffusion training pipelines.

B.2 Learning with Weak Annotations

Learning with *weak annotations* addresses scenarios where supervision is corrupted, ambiguous, or incomplete relative to clean ground-truth labels. This domain has been extensively studied in discriminative classification. Canonical settings include: *Noisy-Label Learning (NLL)*, which handles corrupted or flipped labels (Han et al., 2018; Wei et al., 2021; Han et al., 2020); *Ambiguous-Label Learning (ALL)*, also known as *Partial-Label Learning (PLL)*, where each instance is annotated with a candidate set containing the true label (Feng et al., 2020; Wu et al., 2022; Tian et al., 2023; Lv et al., 2020; Wang et al., 2025c); and *Semi-Supervised Learning (SSL)*, which utilizes a small labeled set alongside a large unlabeled corpus (Berthelot et al., 2019; Zhang et al., 2021; Yang et al., 2022; Wang et al., 2022). Beyond these individual settings, recent work has explored *mixture weak supervision* (Chen et al., 2024a; Zhang et al., 2020; Wei et al., 2023; Shukla et al., 2023; Xie et al., 2024), combining multiple forms of imperfection within a single framework.

While prior literature predominantly focuses on improving predictive accuracy for discriminative models, our work lifts these foundational concepts to the generative domain. We provide a unified view of conditional score modeling under weak annotations, enabling robust generation without requiring clean validation data or task-specific architectural modifications.

C Implementation Details

Our implementation is based on PyTorch 1.12 (Paszke et al., 2019), and all experiments were conducted on NVIDIA Tesla A100 GPUs with CUDA 12.4.

Weak annotations construction. For all class-dependent ambiguous-label datasets, we construct a 10×10 circulant transition matrix T where each row maps a true label to a candidate set of labels with varying probabilities. We set $q = 0.5$ and define the matrix as:

$$T = \begin{bmatrix} 1 & q{+}0.2 & q & \cdots & q{+}0.2 & q & q{-}0.2 \\ q{-}0.2 & 1 & q{+}0.2 & \cdots & q & q{-}0.2 & q{+}0.2 \\ q & q{-}0.2 & 1 & \cdots & q{-}0.2 & q{+}0.2 & q \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ q{+}0.2 & q & q{-}0.2 & \cdots & q & q{-}0.2 & 1 \end{bmatrix}.$$

For noisy-label datasets with asymmetric noise (40% flip probability), we adopt the following mappings: *Fashion-MNIST:* 'Pullover'→'Sneaker', 'Dress'→'Bag', 'Sandal'→'Shirt', 'Shirt'→'Sandal'. *CIFAR-10:* 'Truck'→'Automobile', 'Bird'→'Airplane', 'Deer'→'Horse', 'Cat'→'Dog', 'Dog'→'Cat'. *ImageNette:* 'Tench'→'English springer', 'Cassette player'→'Garbage truck', 'Chain saw'→'Church', 'Golf ball'→'Parachute', 'Parachute'→'Golf ball'.

Model setup. We build our framework upon the EDM formulation (Karras et al., 2022), utilizing the DDPM++ network architecture with a U-Net backbone (Song & Ermon, 2020) for all experiments. Optimization is performed using the Adam optimizer with a learning rate of 1×10^{-3}, $(\beta_1, \beta_2) = (0.9, 0.999)$, and $\epsilon = 1 \times 10^{-8}$. We apply an Exponential Moving Average (EMA) decay of 0.5. The batch sizes are set to 128 for Fashion-MNIST, 64 for CIFAR-10, and 16 for ImageNette. Regarding the proposed diffusion classifier, the timestep interval length Δ is set to 6.4. All models are trained from scratch for a total of 200k iterations.

D Additional Experimental Results

D.1 Evaluation Metrics

We evaluate the trained CDMs using four unconditional metrics, including Fréchet Inception Distance (FID) (Heusel et al., 2017), Inception Score (IS) (Salimans et al., 2016), Density, and Coverage (Naeem et al., 2020), and three conditional metrics, namely CW-FID, CW-Density, CW-Coverage (Chao et al., 2022). All metrics are computed using the official implementation of DLSM (Chao et al., 2022). Although these metrics have been introduced in related work (Na et al., 2024), we briefly recap them here for completeness and clarity.

Unconditional metrics. Unconditional metrics evaluate generated samples without reference to class labels. In our experiments, images are first generated conditionally per class but then pooled without labels when computing the metrics. This evaluation protocol is consistent with prior studies (Kaneko et al., 2019; Chao et al., 2022).

- FID measures the distance between real and generated image distributions in the pre-trained feature space (Szegedy et al., 2016), indicating the fidelity and diversity of generated images.
- IS evaluates whether generated images belong to distinct classes and whether each image is class-consistent, reflecting the realism and class separability of generated images.
- Density and Coverage are reliable versions of Precision and Recall (Naeem et al., 2020), respectively. Density measures how well generated samples cover real data distribution, while Coverage assesses how well real samples are represented by generated ones.

Conditional metrics. To measure conditional consistency, we adopt class-wise (CW) variants of the above metrics, which compute each metric separately within each class and then average across classes. Notably, CW-FID (also called intra-FID) is widely used in conditional generative modeling (Miyato & Koyama, 2018; Kaneko et al., 2019), and has been highlighted as a key measure of conditional distribution quality.

Remark: It should be noted that the Fashion-MNIST dataset is not suitable for evaluation using these metrics, so we do not perform evaluation on the Fashion-MNIST dataset.

E DISCUSSION

E.1 ANALYSIS AND EMPIRICAL EVIDENCE OF MIXTURE BIAS IN NAIVE TRAINING

In Section 3, we argued that naively training a conditional diffusion model on weakly annotated pairs $(\mathbf{x}, z)$ leads to mixture bias. Here, we provide a formal derivation of this phenomenon and present empirical evidence across different weak annotations.

Remark 1 (The Mixture Bias)**.** Consider the naive objective function:

$$\mathcal{L}_{\text{naive}}(\theta) = \mathbb{E}_{t,z,\mathbf{x}_t}\Big[w_t \big\|\mathbf{s}_\theta(\mathbf{x}_t, z, t) - \nabla_{\mathbf{x}_t} \log q_{t|0}(\mathbf{x}_t \,|\, \mathbf{x}_0, z)\big\|_2^2\Big]. \tag{17}$$

The global optimum $\mathbf{s}_{\theta^*}$ of this objective does not recover the clean-conditional score corresponding to the true label y. Instead, it converges to the gradient of the mixture distribution conditioned on the weak label z:

$$\mathbf{s}_{\theta^*}(\mathbf{x}_t, z, t) = \nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t \,|\, z) = \nabla_{\mathbf{x}_t} \log \left(\sum_{y \in \mathcal{Y}} p(y \,|\, z)\, q_t(\mathbf{x}_t \,|\, y) \right).$$

Generating from this mixture forces the model to compromise between diverse class attributes. Instead of recovering a specific class, the generation follows an averaged path, causing class mismatch or severe homogenization.

DERIVATION

Although this result aligns with fundamental properties of score matching (Vincent, 2011; Song & Ermon, 2019), we provide the specific derivation for the weak annotations here. Let $\mathcal{L}_{\text{DSM}}(\theta)$ and $\mathcal{L}_{\text{ESM}}(\theta)$ denote the denoising score matching (DSM) objective used in training and the explicit score matching (ESM) objective target, respectively:

$$\mathcal{L}_{\text{DSM}}(\theta) := \mathbb{E}_{t,z}\Big[w_t\, \mathbb{E}_{\mathbf{x}_t,\mathbf{x}_0 \sim q(\mathbf{x}_t,\mathbf{x}_0|z)} \big\|\mathbf{s}_\theta(\mathbf{x}_t, z, t) - \nabla_{\mathbf{x}_t} \log q_{t|0}(\mathbf{x}_t \,|\, \mathbf{x}_0, z)\big\|_2^2\Big],$$

$$\mathcal{L}_{\text{ESM}}(\theta) := \mathbb{E}_{t,z}\Big[w_t\, \mathbb{E}_{\mathbf{x}_t \sim q(\mathbf{x}_t|z)} \big\|\mathbf{s}_\theta(\mathbf{x}_t, z, t) - \nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t \,|\, z)\big\|_2^2\Big].$$

By expanding the squared norms, it can be shown that these two objectives differ only by a constant C that is independent of θ:

$$\mathcal{L}_{\text{ESM}}(\theta) = \mathcal{L}_{\text{DSM}}(\theta) + C.$$

Therefore, both objectives share the same minimizer θ^*. By inspecting $\mathcal{L}_{\text{ESM}}$, it is trivial to see that the quadratic term is minimized when the model prediction matches the target exactly:

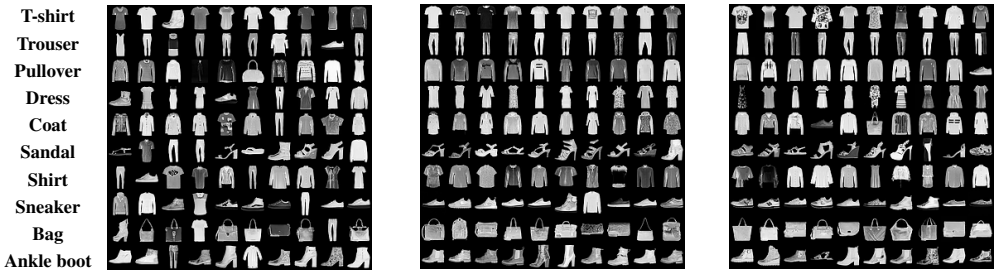
$$\mathbf{s}_{\theta^*}(\mathbf{x}_t, z, t) = \nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t \,|\, z).$$

This confirms that naive training learns the score of the mixture distribution $q_t(\mathbf{x}_t \,|\, z)$.

EMPIRICAL OBSERVATION

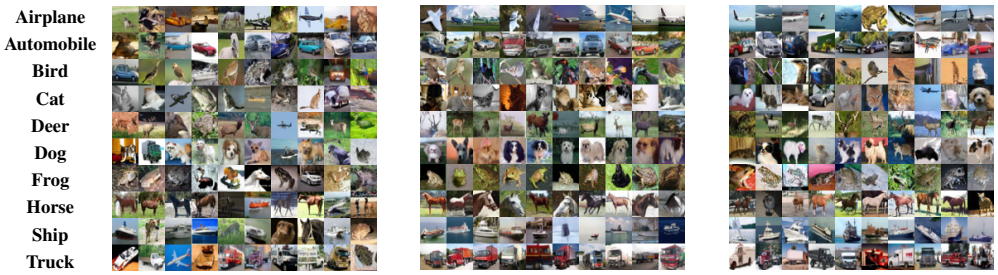
To visualize the practical consequences of this bias, we trained baseline conditional diffusion models using the naive objective in Eq. (17) under three types of weak annotations. The generated samples are shown in Figures 2, 3, and 4. The distinct failure modes align with our theoretical analysis:

- **Noisy-label annotation (class mismatch):** As shown in sub-figures (a), the model often suffers from semantic mismatch. The strong false signals in the mixture distribution $q_t(\mathbf{x}_t \,|\, z)$ cause the model to hallucinate incorrect classes (e.g., generating a dog when conditioned on a cat label), failing to disentangle the clean signal from the noise.
- **Ambiguous-label annotation (homogenization):** As shown in sub-figures (b), the generated images often lack diversity. This is particularly pronounced on ImageNette, where samples within the same class appear highly similar. While the generated categories generally align with the ground-truth supersets, the mixture gradient pulls trajectories toward a safe "mean" mode, preventing the model from capturing diverse intra-class variations.
- **Semi-supervised data (low diversity):** As shown in sub-figures (c), this setting combines the challenges of limited labeled data and abundant unlabeled data. Due to the scarcity of conditional training samples, the model struggles to capture the full data distribution, resulting in images with significantly reduced diversity.



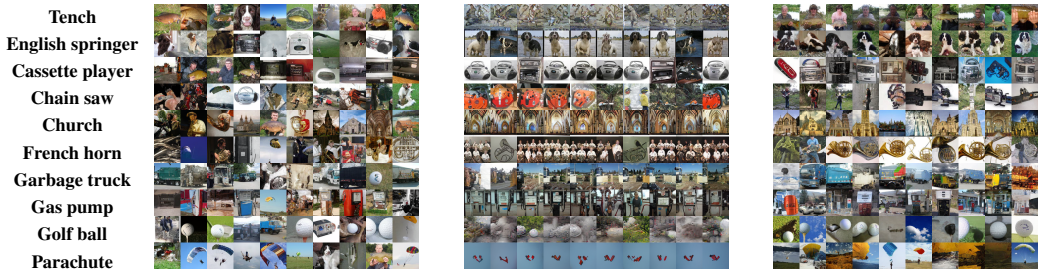
(a) Noisy Labels (b) Ambiguous Labels (c) Semi-Supervised Data

Figure 2: Examples of randomly generated Fashion-MNIST images from naive models trained under different types of weak annotations.



(a) Noisy Labels (b) Ambiguous Labels (c) Semi-Supervised Data

Figure 3: Examples of randomly generated CIFAR-10 images from naive models trained under different types of weak annotations.



(a) Noisy Labels (b) Ambiguous Labels (c) Semi-Supervised Data

Figure 4: Examples of randomly generated ImageNette images from naive models trained under different types of weak annotations.

E.2 Analysis of Early-Learning Regularization in Eq. (11)

The effectiveness of Eq. (11) can be better understood by examining the form of its gradient. For clarity, we restate the loss with the following notation: given a noisy-labeled input $(\mathbf{x}, \hat{y})$, we denote the model's output probabilities as $f_\theta(\mathbf{x})$ and the corresponding EMA target as $f_\phi(\mathbf{x})$.

Let $\hat{\mathbf{y}} \in \mathbb{R}^c$ be the one-hot vector corresponding to the noisy label $\hat{y}$. Then the loss over the whole dataset $\mathcal{D} = \{(\mathbf{x}^{[i]}, \hat{\mathbf{y}}^{[i]})\}_{i=1}^n$ can be computed according to Eq. (11) as

$$\mathcal{L}_{\text{Cls}}^{\text{NL}}(\mathcal{D}) = -\frac{1}{n}\sum_{i=1}^{n} \left\langle \text{sg}(\mathbf{r}^{[i]}),\, \log f_\theta(\mathbf{x}^{[i]}) \right\rangle, \quad \mathbf{r}^{[i]} = \hat{\mathbf{y}}^{[i]} - \lambda \frac{f_\theta(\mathbf{x}^{[i]}) \odot \left(\delta^{[i]}\mathbf{1} - f_\phi(\mathbf{x}^{[i]})\right)}{1 - \delta^{[i]}}, \quad (18)$$

where $\delta^{[i]} = \langle f_\theta(\mathbf{x}^{[i]}), f_\phi(\mathbf{x}^{[i]}) \rangle$, $\text{sg}(\cdot)$ denotes the stop-gradient operator, and $\odot$ is the Hadamard product. By construction $\mathbf{r}^{[i]}$ is treated as a *constant* w.r.t. θ due to the stop-gradient.

Lemma 1. *Let $\psi_\theta(\mathbf{x})$ denote the pre-softmax logits such that $f_\theta(\mathbf{x}) = \mathrm{softmax}(\psi_\theta(\mathbf{x}))$. For the loss in Eq. (11), the gradients are*

$$\frac{\partial \mathcal{L}_{\mathrm{Cls}}^{\mathrm{NL}}(\mathbf{x}^{[i]})}{\partial \psi_\theta(\mathbf{x}^{[i]})} = f_\theta(\mathbf{x}^{[i]}) - \mathrm{sg}\left(\mathbf{r}^{[i]}\right), \quad \textit{for each } i = 1, \ldots, n, \tag{19}$$

and, by the chain rule,

$$\nabla_\theta \mathcal{L}_{\mathrm{Cls}}^{\mathrm{NL}}(\mathcal{D}) = \frac{1}{n} \sum_{i=1}^{n} J_{\mathbf{z}_\theta}(\mathbf{x}^{[i]})^\top \left[f_\theta(\mathbf{x}^{[i]}) - \mathrm{sg}\big(\mathbf{r}^{[i]}\big) \right], \tag{20}$$

where $J_{\mathbf{z}_\theta}(\mathbf{x}) = \partial \mathbf{z}_\theta(\mathbf{x}) / \partial \theta$ is the Jacobian of the logits w.r.t. the parameters. Equivalently, expanding $\mathbf{r}^{[i]}$ gives

$$\nabla_\theta \mathcal{L}_{\mathrm{Cls}}^{\mathrm{NL}}(\mathcal{D}) = \frac{1}{n} \sum_{i=1}^{n} J_{\mathbf{z}_\theta}(\mathbf{x}^{[i]})^\top \left[f_\theta(\mathbf{x}^{[i]}) - \hat{\mathbf{y}}^{[i]} + \lambda \, \mathrm{sg}\Big(\frac{f_\theta(\mathbf{x}^{[i]}) \odot \big(\delta^{[i]} \mathbf{1} - f_\phi(\mathbf{x}^{[i]})\big)}{1 - \delta^{[i]}} \Big) \right]. \tag{21}$$

Proof. For any $i \in \{1, \ldots, n\}$, let us first verify that $\mathbf{r}^{[i]}$ sums to 1. With

$$\mathbf{r}^{[i]} = \hat{\mathbf{y}}^{[i]} - \lambda \, \frac{f_\theta(\mathbf{x}^{[i]}) \odot \big(\delta^{[i]} \mathbf{1} - f_\phi(\mathbf{x}^{[i]})\big)}{1 - \delta^{[i]}},$$

we sum over classes and using $\langle f_\theta(\mathbf{x}^{[i]}), \mathbf{1} \rangle = 1$ yields

$$\mathbf{1}^\top \mathbf{r}^{[i]} = 1 - \lambda \, \frac{\delta^{[i]} - \langle f_\theta(\mathbf{x}^{[i]}), f_\phi(\mathbf{x}^{[i]}) \rangle}{1 - \delta^{[i]}} = 1,$$

so $\mathbf{r}^{[i]}$ lies on the simplex (hence Eq. (18) is an ordinary cross-entropy with a fixed target). Let $\psi^{[i]} = \psi_\theta(\mathbf{x}^{[i]})$ be the logits and recall $\frac{\partial \log \mathrm{softmax}(\psi)}{\partial \psi} = I - \mathrm{softmax}(\psi) \mathbf{1}^\top$. For the per-sample loss $\ell^{[i]} = -\langle \mathrm{sg}(\mathbf{r}^{[i]}), \log \mathrm{softmax}(\psi^{[i]}) \rangle$, the derivative w.r.t. logits is

$$\frac{\partial \ell^{[i]}}{\partial \psi^{[i]}} = \mathrm{softmax}(\psi^{[i]}) - \mathrm{sg}(\mathbf{r}^{[i]}) = f_\theta(\mathbf{x}^{[i]}) - \mathrm{sg}(\mathbf{r}^{[i]}),$$

which is Eq. (19). Applying the chain rule and averaging over i gives Eq. (20). Replacing $\mathrm{sg}(\mathbf{r}^{[i]})$ by its explicit form produces Eq. (21). □

Remark. Eq. (21) shows that $\mathcal{L}_{\mathrm{Cls}}^{\mathrm{NL}}$ behaves like the standard cross-entropy gradient plus an ELR-like corrective term. This term amplifies gradients on clean samples and counteracts gradients on noisy samples. Specifically, we expand this ELR-like corrective term into:

$$\mathbf{g}_y^{[i]} := \frac{f_\theta(\mathbf{x}^{[i]})}{1 - \langle f_\theta(\mathbf{x}^{[i]}), f_\phi(\mathbf{x}^{[i]}) \rangle} \sum_{k=1}^{c} (f_\phi(\mathbf{x}^{[i]})_k - f_\phi(\mathbf{x}^{[i]})_y) f_\theta(\mathbf{x}^{[i]})_k. \tag{22}$$

If y^* is the true class, then the y^*th entry of $f_\phi(\mathbf{x}^{[i]})$ tends to be dominant during early-learning. In that case, the y^*th entry of $\mathbf{g}^{[i]}$ is negative. This is useful both for examples with clean labels and for examples with noisy labels. For examples with clean labels, the cross-entropy term $f_\theta(\mathbf{x}^{[i]}) - \hat{\mathbf{y}}^{[i]}$ tends to vanish after the early-learning stage because $f_\theta(\mathbf{x}^{[i]})$ is very close to $\hat{\mathbf{y}}^{[i]}$, allowing examples with wrong labels to dominant the gradient. Adding $\mathbf{g}^{[i]}$ counteracts this effect by ensuring that the magnitudes of the coefficients on examples with clean labels remain large. Thus, $\mathbf{g}^{[i]}$ fulfils the two desired properties that boosting the gradient of examples with clean labels, and neutralizing the gradient of the examples with false labels.

F PROOF

F.1 CONNECTIONS AMONG DIFFERENT DIFFUSION MODELS.

The diffusion model we define in this paper allows for a flexible parametrization that aligns with common diffusion frameworks, such as DDPM (Ho et al., 2020), SMLD (Song & Ermon, 2019), VE-SDE (Song et al., 2020) and VP-SDE (Song et al., 2020). This demonstrates that our formulation is compatible with diverse diffusion paradigms while facilitating unified theoretical analysis via specific choices of schedules α_t and σ_t.

DDPM. DDPM defines a sequence $\{\beta_t\}_{t=0}^T$ and $\mathbf{x}_t = \sqrt{\prod_{i=0}^t (1-\beta_i)}\mathbf{x}_0 + \sqrt{1-\prod_{i=0}^t (1-\beta_i)}\boldsymbol{\epsilon}$, which can be seen as a special case of Eq. (1) where we can set $\alpha_t = \sqrt{\prod_{i=0}^t (1-\beta_i)}$ and $\sigma_t = \sqrt{1-\prod_{i=0}^t (1-\beta_i)}$.

SMLD. SMLD defines a noise schedule $\sigma(t)_{t=0}^T$ and $\mathbf{x}_t = \mathbf{x}_0 + \sigma(t)\boldsymbol{\epsilon}$, with $\sigma(1) < \sigma(2) < \cdots < \sigma(T)$. In this scenario, Eq. (1) reduces to $\alpha_t = 1$, $\sigma_t = \sigma(t)$.

VP-SDE. VP-SDE is the continuous case of DDPM, which defines a stochastic differential equation (SDE) as

$$dX_t = -\frac{1}{2}\beta(t)X_t\,\mathrm{d}t + \sqrt{\beta(t)}\,\mathrm{d}W_t, \quad t \in [0,1],$$

where $\beta(t) = \beta_{t\cdot T}\cdot T$. In this setup, $\alpha_t = \sqrt{\exp\left(-\int_0^t \beta(s)\,\mathrm{d}s\right)}$, and $\sigma_t = \sqrt{1-\exp\left(-\int_0^t \beta(s)\,\mathrm{d}s\right)}$.

VE-SDE. VE-SDE is the continuous case of SMLD, whose forward process of VE-SDE is defined as

$$dX_t = \sqrt{\frac{\mathrm{d}\sigma(t)^2}{\mathrm{d}t}}\,\mathrm{d}W_t.$$

In this setup, $\alpha_t = 1$ and $\sigma_t = \sqrt{\sigma^2(t) - \sigma^2(0)}$.

Regarding the weighting function w_t. Different frameworks typically adopt specific prediction targets (e.g., ϵ-prediction in DDPM or score-prediction in SMLD), which implicitly corresponds to the usage of specific weighting functions w_t in the score matching objective. For instance, minimizing the error on $\boldsymbol{\epsilon}$ is equivalent to setting $w_t \propto \sigma_t^2$, while minimizing the error on $\mathbf{x}_0$ corresponds to a different weighting. While the models above each define their own specific frameworks, EDM (Karras et al., 2022) proposes a unified structure and optimizes these parameter choices, including preconditioning the network inputs and outputs, to ensure training stability. Therefore, for our implementation, we adopt EDM as the foundational framework. We provide the conversion logic in Algorithm 1 to demonstrate how our score network $\mathbf{s}_\theta$ adapts to different prediction targets ($\mathbf{x}_0$, $\boldsymbol{\epsilon}$, or $\boldsymbol{v}$), thereby effectively recovering the distinct loss weightings used in prior works.

F.2 PROOF OF THEOREM 1

The derivation here is analogous to that of Theorem 1 in Na et al. (2024), and we provide the full proof below for completeness. First, under the weak annotation setting, the weak data distribution conditioned on z forms a mixture model:

$$q_t(\mathbf{x}_t|z) = \sum_{y=1}^{c} p(y|z)q_t(\mathbf{x}_t|y) \quad \forall \mathbf{x}_t \in \mathcal{X}, z \subset \mathcal{Y}.$$

This implies that the transition from weak annotations to clean annotations is independent of the timesteps. Consequently, Eq. (5) can be derived as follows,

$$
\begin{aligned}
&\nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t|z) \\
=&\frac{\nabla_{\mathbf{x}_t} q_t(\mathbf{x}_t|z)}{q_t(\mathbf{x}_t|z)} \\
=&\frac{\sum_{y=1}^{c} p(y|z)\nabla_{\mathbf{x}_t} q_t(\mathbf{x}_t|y)}{q_t(\mathbf{x}_t|z)} \\
=&\sum_{y=1}^{c} \frac{p(y|z)q_t(\mathbf{x}_t|y)}{q_t(\mathbf{x}_t|z)} \cdot \frac{\nabla_{\mathbf{x}_t} q_t(\mathbf{x}_t|y)}{q_t(\mathbf{x}_t|y)} \\
=&\sum_{y=1}^{c} \frac{p(y|z)q_t(\mathbf{x}_t|y)}{q_t(\mathbf{x}_t|z)} \cdot \nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t|y) \\
=&\sum_{y=1}^{c} p(y|z) \cdot \frac{p(z)}{p(y)} \cdot \frac{p(y|\mathbf{x}_t)}{p(z|\mathbf{x}_t)} \cdot \frac{q_t(\mathbf{x}_t)}{q_t(\mathbf{x}_t)} \cdot \nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t|y) \\
=&\sum_{y=1}^{c} p(z|y) \cdot \frac{p(y|\mathbf{x}_t)}{p(z|\mathbf{x}_t)} \cdot \nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t|y) \\
=&\sum_{y=1}^{c} p(z|y,\mathbf{x}_t) \cdot \frac{p(y|\mathbf{x}_t)}{p(z|\mathbf{x}_t)} \cdot \nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t|y) \quad \text{(Conditional indep. of } z \text{ and } \mathbf{x}_t \text{ given } y.\text{)} \\
=&\sum_{y=1}^{c} \frac{p(z|y,\mathbf{x}_t)p(y|\mathbf{x}_t)}{p(z|\mathbf{x}_t)} \cdot \nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t|y) \\
=&\sum_{y=1}^{c} p(y|\mathbf{x}_t,z)\nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t|y)
\end{aligned}
$$

Remark 2. In practice, the true posterior weight $p(y|\mathbf{x}_t, z)$ is unknown and intractable to compute. Following the framework of variational inference (analogous to the introduction of $p_\phi(Y \mid X, Z)$ in Appendix F.6), we introduce a learnable approximation $p_\phi(y|\mathbf{x}_t, z)$, parameterized by an EMA diffusion-based classifier network ϕ, to estimate this posterior.

Algorithm 1 Formulation Conversion (Our Model to EDM)

Require: A score network $\mathbf{s}_\theta$, a noisy input $\mathbf{x}_t$, noise level t, linear schedule $\{\alpha_i\}_{i=1}^T$ and $\{\sigma_i\}_{i=1}^T$.
1: Calculate the denoised image $\mathbf{x}_0$ using $\mathbf{s}_\theta$: $\mathbf{x}_0 = (\mathbf{x}_t + \sigma_t^2 \mathbf{s}_\theta(\mathbf{x}_t/\alpha_t, \sigma_t/\alpha_t))/\alpha_t$
2: **if** performing $\mathbf{x}_0$-prediction **then**
3: Return $\mathbf{x}_0$.
4: **end if**
5: Calculate the noise component $\boldsymbol{\epsilon}$: $\boldsymbol{\epsilon} = \frac{\mathbf{x}_t - \alpha_t \mathbf{x}_0}{\sigma_t}$
6: **if** performing $\boldsymbol{\epsilon}$-prediction **then**
7: Return $\boldsymbol{\epsilon}$.
8: **end if**
9: Calculate the velocity component $\boldsymbol{v}$: $\boldsymbol{v} = \alpha_t \boldsymbol{\epsilon} - \sigma_t \mathbf{x}_0$
10: **if** performing $\boldsymbol{v}$-prediction **then**
11: Return $\boldsymbol{v}$.
12: **end if**

F.3 Proof of Proposition 1

From Remark 1, the optimal score network $\mathbf{s}_{\theta^*}$ targets the gradient of the log-mixture density. Applying the decomposition from Theorem 1, we align this target with our generative formulation:

$$\mathbf{s}_{\theta^*_{\text{Gen}}}(\mathbf{x}_t, z, t) = \nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t \,|\, z) = \sum_{y=1}^{c} p(y \,|\, \mathbf{x}_t, z)\, \nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t \,|\, y).$$

$\mathcal{L}_{\text{Gen}}(\theta)$ minimizes the error between the model's weighted output and this target score:

$$\mathcal{L}_{\text{Gen}}(\theta) = \mathbb{E}_{t,\mathbf{x}_t,z}\left[w_t \left\| \sum_{y=1}^{c} p(y \,|\, \mathbf{x}_t, z) \Big(\mathbf{s}_\theta(\mathbf{x}_t, y, t) - \nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t \,|\, y) \Big) \right\|_2^2 \right].$$

Since the objective is strictly convex, the global minimum is achieved when the integrand vanishes pointwise. Given that $w_t > 0$, for any candidate class y with non-zero posterior support (i.e., $p(y \,|\, \mathbf{x}_t, z) > 0$), the optimality condition necessitates:

$$\mathbf{s}_{\theta^*_{\text{Gen}}}(\mathbf{x}_t, y, t) = \nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t \,|\, y).$$

Thus, the model recovers the true conditional score for all valid classes.

F.4 Proof of Theorem 2

We define optimality by centering the subinterval $[l, r]$ around the median τ_{mid} of the weight distribution $w(\tau)$, thereby capturing the most informative region.

Case 1: Log-Normal Distribution (EDM Framework). In the EDM formulation (Karras et al., 2022), the weighting function $w(\tau)$ is designed to peak at intermediate noise levels, effectively inducing a distribution over timesteps. Specifically, the weight density follows a log-normal shape:

$$w(\tau) \propto \frac{1}{\tau} \exp\left(-\frac{(\ln \tau - P_{\text{mean}})^2}{2P_{\text{std}}^2} \right),$$

with $P_{\text{mean}} = -1.2$ and $P_{\text{std}} = 1.2$. Normalizing this weight function yields a probability density function $p_w(\tau) = w(\tau)/Z$, which corresponds to the PDF of a log-normal distribution for τ. The cumulative distribution function (CDF) of this weight mass is given by

$$\mathcal{F}_w(\tau) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{\ln \tau - P_{\text{mean}}}{P_{\text{std}}\sqrt{2}} \right) \right],$$

where $\text{erf}(x)$ denotes the error function.

Let τ_{mid} be the median of this weight distribution, satisfying $\mathcal{F}_w(\tau_{\text{mid}}) = 0.5$. To center the interval of length Δ around this median weight mass, we impose that the cumulative weight mass excluded on the left equals that excluded on the right:

$$\mathcal{F}_w(l) = 1 - \mathcal{F}_w(r) = 1 - \mathcal{F}_w(l + \Delta).$$

Rearranging terms yields the implicit equation for the optimal start time l:

$$\mathcal{F}_w(l) + \mathcal{F}_w(l + \Delta) = 1. \tag{23}$$

Given the monotonicity of the CDF, this equation has a unique solution for l, which we compute numerically (e.g., using Brent's method Brent (2013)). The optimal interval is then $[l, l + \Delta]$.

Case 2: Uniform Weighting (DDPM Framework). In standard DDPM (Ho et al., 2020), discrete timesteps are often weighted equally in simplified objectives, implying a constant weighting function $w(\tau) = \text{const}$. The corresponding cumulative weight function is linear: $\mathcal{F}_w(\tau) = \tau$. Applying the same symmetry condition from Eq. (23):

$$\begin{aligned} l + (l + \Delta) &= 1 \\ 2l &= 1 - \Delta \\ l &= \frac{1 - \Delta}{2}. \end{aligned}$$

Consequently, the interval becomes $[\frac{1-\Delta}{2}, \frac{1+\Delta}{2}]$, which is perfectly symmetric around the midpoint 0.5.

Conclusion. In both cases, this derivation guarantees that the selected interval concentrates on the region with the highest weight density. As $\Delta \to 0$, the interval converges to the median of the weight distribution, effectively identifying the single timesteps that contribute most significantly to the weighted objective, consistent with findings in prior diffusion classifier literature Li et al. (2023); Wang & Chen (2025).

F.5 Proof of Theorem 3

For notational conciseness, we omit the class dependency y in the reconstruction error, denoting it simply as $\hbar(\tau)$, and define $w(\tau)$ as the weighting function (referring to w_τ in the main text).

1. Objective Formulation. Define the weight mass $W_{[l,r]}$ and weighted error $H_{[l,r]}$ on $[l, r]$ as:

$$W_{[l,r]} := \int_l^r w(\tau)\,\mathrm{d}\tau, \quad H_{[l,r]} := \int_l^r w(\tau)\hbar(\tau)\,\mathrm{d}\tau.$$

Let $\mu_{[l,r]} = H_{[l,r]}/W_{[l,r]}$ and $\mu_{\mathcal{T}}$ denote the expected errors on $[l, r]$ and the effective range $\mathcal{T} = [u(t+1), u(T-1)]$, respectively. We minimize their squared difference subject to $W_{[l,r]} = C$:

$$\min_{l,r} \quad \mathcal{J}(l, r) := (\mu_{[l,r]} - \mu_{\mathcal{T}})^2, \quad \text{s.t.} \quad W_{[l,r]} - C = 0. \tag{24}$$

2. Optimization via Lagrange Multipliers. We construct the Lagrangian $\mathcal{L}(l, r, \lambda)$:

$$\mathcal{L}(l, r, \lambda) = (\mu_{[l,r]} - \mu_{\mathcal{T}})^2 + \lambda(W_{[l,r]} - C).$$

Applying the Leibniz integral rule, the partial derivatives w.r.t. the interval boundaries are

$$\frac{\partial H}{\partial l} = -w(l)\hbar(l), \quad \frac{\partial H}{\partial r} = w(r)\hbar(r), \quad \frac{\partial W}{\partial l} = -w(l), \quad \frac{\partial W}{\partial r} = w(r).$$

Differentiating $\mu_{[l,r]} = H_{[l,r]}/W_{[l,r]}$ yields

$$\frac{\partial \mu_{[l,r]}}{\partial l} = \frac{w(l)}{W_{[l,r]}}(\mu_{[l,r]} - \hbar(l)), \quad \frac{\partial \mu_{[l,r]}}{\partial r} = \frac{w(r)}{W_{[l,r]}}(\hbar(r) - \mu_{[l,r]}).$$

Setting the gradients of the Lagrangian $\nabla\mathcal{L}$ to zero yields the first-order necessary conditions for the optimal interval $[l^*, r^*]$:

$$\left.\frac{\partial \mathcal{L}}{\partial l}\right|_{l^*} = 2(\mu_{[l^*,r^*]} - \mu_{\mathcal{T}}) \cdot \frac{w(l^*)}{W}(\mu_{[l^*,r^*]} - \hbar(l^*)) - \lambda w(l^*) = 0,$$
$$\left.\frac{\partial \mathcal{L}}{\partial r}\right|_{r^*} = 2(\mu_{[l^*,r^*]} - \mu_{\mathcal{T}}) \cdot \frac{w(r^*)}{W}(\hbar(r^*) - \mu_{[l^*,r^*]}) + \lambda w(r^*) = 0$$

Because $w(l^*), w(r^*) > 0$, we can divide by $w(\cdot)$ and equate the terms involving λ:

$$\frac{2}{W}(\mu_{[l^*,r^*]} - \mu_{\mathcal{T}})(\mu_{[l^*,r^*]} - \hbar(l^*)) = \lambda = \frac{2}{W}(\mu_{[l^*,r^*]} - \mu_{\mathcal{T}})(\hbar(r^*) - \mu_{[l^*,r^*]}).$$

We then divide by the common factor $\frac{2}{W}(\mu_{[l^*,r^*]} - \mu_{\mathcal{T}})$ to obtain:

$$\mu_{[l^*,r^*]} - \hbar(l^*) = \hbar(r^*) - \mu_{[l^*,r^*]} \quad \Longrightarrow \quad \mu_{[l^*,r^*]} = \frac{\hbar(l^*) + \hbar(r^*)}{2}.$$

Specifically, this recovers the necessary optimal condition stated in Theorem 3.

F.6 Derivation of Eq. (4)

To maximize the log-likelihood $\log p_\theta(X, Z)$, we introduce an approximate posterior distribution $p_\phi(Y \mid X, Z)$ over the latent variable Y. Applying Jensen's inequality, we derive the ELBO as

follows:

$$\begin{aligned}
\log p_\theta(X, Z) &= \log \sum_{Y \in \mathcal{Y}} p_\theta(X, Y, Z) \\
&= \log \sum_{Y \in \mathcal{Y}} p_\phi(Y \mid X, Z) \frac{p_\theta(X, Y, Z)}{p_\phi(Y \mid X, Z)} \\
&\geq \sum_{Y \in \mathcal{Y}} p_\phi(Y \mid X, Z) \log \frac{p_\theta(X, Y, Z)}{p_\phi(Y \mid X, Z)} \\
&= \mathbb{E}_{p_\phi(Y \mid X, Z)} \Big[\log p_\theta(X, Y, Z) - \log p_\phi(Y \mid X, Z) \Big].
\end{aligned}$$

Since the entropy term $-\mathbb{E}_{p_\phi}[\log p_\phi]$ does not depend on θ, maximizing the ELBO w.r.t. θ simplifies to

$$\hat{\theta} = \arg\max_\theta \; \mathbb{E}_{p_\phi(Y \mid X, Z)} \big[\log p_\theta(X, Y, Z) \big].$$

Using the chain rule of probability, we can decompose the log-likelihood as

$$\log p_\theta(X, Y, Z) = \log p_\theta(X \mid Z) + \log p_\theta(Y \mid X, Z) + \log p_\theta(Z).$$

Substituting this back into the expectation:

$$\begin{aligned}
\hat{\theta} &= \arg\max_\theta \; \mathbb{E}_{p_\phi(Y \mid X, Z)} \Big[\log p_\theta(X \mid Z) + \log p_\theta(Y \mid X, Z) + \log p_\theta(Z) \Big] \\
&= \arg\max_\theta \; \Big\{ \mathbb{E}_{p_\phi(Y \mid X, Z)} \big[\log p_\theta(X \mid Z) \big] + \mathbb{E}_{p_\phi(Y \mid X, Z)} \big[\log p_\theta(Y \mid X, Z) \big] \Big\}.
\end{aligned}$$

The term $\log p_\theta(Z)$ is constant w.r.t. the optimization and thus omitted, yielding the objective in Eq. (4).

(a) Fashion-MNIST (b) CIFAR-10 (c) ImageNette

Figure 5: Class-conditional generation samples: *Vanilla* (top) vs. *DELTA* (bottom).