Wasserstein-Aware Transfer: Class-Level Alignment for Robust Diffusion Model Adaptation

Zi-Xian Huang, Chuan-Xian Ren*

School of Mathematics, Sun Yat-Sen University, China
huangzx86@mail2.sysu.edu.cn, rchuanx@mail.sysu.edu.cn

Abstract

Diffusion models have achieved impressive generative performance across diverse domains such as image, video, and scientific data generation. However, fine-tuning these models for new tasks remains challenging due to their large scale, architectural diversity, and high sensitivity to hyperparameters—particularly learning rates. In this work, we propose Wasserstein-Aware Transfer (WAT), a principled and effective fine-tuning strategy grounded in diffusion trajectory analysis and optimal transport theory. Our key insight is that the distributional discrepancies between diffusion trajectories from different datasets decrease progressively over time and converge near the noise end. Based on this observation, we introduce a class-wise matching mechanism that minimizes the Wasserstein distance between class distributions of source and target datasets. This enables alignment at the class level without modifying the standard fine-tuning pipeline. To further enhance knowledge retention, we propose a novel sampling strategy that linearly combines class-conditional outputs from both pretrained and fine-tuned models. This method is simple yet effective, requiring negligible computational overhead while preserving domain-specific and generalizable knowledge. Extensive experiments across seven diverse benchmarks demonstrate that WAT reliably enhances generation quality under distribution shifts, outperforming competitive baselines. These results underscore its robustness and affirm the potential of optimal transport as a principled basis for knowledge transfer in diffusion models.

Introduction

Diffusion models have rapidly emerged as a powerful generative technology, demonstrating remarkable progress across a wide spectrum of tasks in recent years, including image generation (Ho, Jain, and Abbeel 2020; Rombach et al. 2022; Dhariwal and Nichol 2021), text-to-image synthesis (Saharia et al. 2022), and video generation (Ho et al. 2022b,a). Beyond these mainstream domains, diffusion models have also been effectively applied in specialized fields such as molecular generation (Hoogeboom et al. 2022; Xu et al. 2022) and molecular dynamics simulation (Corso et al. 2022) in biochemistry. Despite their impressive performance, diffusion models are typically large in scale, and their ability to produce high-quality results comes at the cost of computationally intensive and time-consuming training (Song et al. 2023; Hang et al. 2023). Consequently, research into fine-tuning techniques for diffusion models (Hu et al. 2022; Xie et al. 2023) has emerged as an increasingly important and active area.

For both large language models and diffusion models, recent fine-tuning approaches (Zaken, Ravfogel, and Goldberg 2021; Huang et al. 2025) focus on reducing the number of trainable parameters. These parameter-efficient fine-tuning (PEFT) methods significantly reduce GPU memory consumption, making training more accessible and scalable. However, these methods also bring new challenges—most notably, an increased sensitivity to learning rates, as highlighted by Zhong et al.. In our empirical study, we observe that fine-tuning the DiT model on a new dataset with a larger learning rate often leads to loss explosion and NaN errors. Based on this observation, we aim to improve generation performance by maintaining a moderate learning rate and employing the standard fine-tuning procedure, without additional hyperparameter tuning or architectural modifications. To do this, we analyze knowledge transfer in diffusion models and find that it is largely determined by the structure of the diffusion trajectory.

The feasibility of retaining pretrained knowledge during diffusion model transfer has been demonstrated in prior work. Specifically, Xie et al. achieved efficient fine-tuning by freezing the network's weight matrices, while Zhong et al. leveraged the unconditional branch of a pretrained model to improve generation quality. In addition, we conducted an experiment where only the label embedding and time embedding modules of the diffusion model were fine-tuned. Interestingly, the resulting model still achieved surprisingly strong generation performance. These observations collectively suggest that pretrained diffusion models contain rich, transferable general knowledge. Based on these observations, we leverage optimal transport (specifically the Wasserstein distance) (Li et al. 2020; Gu et al. 2023; Terpin et al. 2024) to quantify distributional differences and analyze the intrinsic discrepancies between diffusion models trained on different datasets. This approach facilitates effective transfer learning for diffusion-based models.

When training diffusion models on different data distributions, the primary differences lie in their diffusion trajecto-

*Corresponding Author

ries — the sequences of intermediate states generated during the diffusion process. To measure differences in trajectories at various time steps, we use the Wasserstein distance, a metric from optimal transport theory that quantifies the minimal effort required to transform one distribution into another. According to theoretical results, this distance between intermediate states is bounded by the discrepancy of the original data distributions. In our experiments, we observe that the Wasserstein distance between diffusion trajectories decreases over time, becoming negligible near the noise end. This suggests that in this late stage, diffusion models trained on different datasets converge to similar representations, enabling effective knowledge sharing across models.

The key insight is that reducing the distributional gap between datasets allows us to better leverage the knowledge encoded in pretrained latent-space generative models. However, directly minimizing the discrepancy between data distributions in the latent space often requires retraining critical components—like the encoder and decoder in VAEs—which leads to significant computational costs. To address this, we propose achieving distribution alignment at the class level within a conditional latent-space generation framework, while maintaining the standard fine-tuning pipeline and avoiding expensive model retraining.

Specifically, we precompute a class-wise matching between the source and target datasets by minimizing the Wasserstein distance between class distributions—using an optimal transport-based algorithm. As a result, each class in the target dataset is paired with the most distributionally similar class from the source dataset. After class-level matching and standard fine-tuning, we further reduce distribution discrepancies by linearly combining the class-conditional branches from the pretrained and fine-tuned models, thereby leveraging their complementary strengths to enhance generation quality. Our method, termed Wasserstein-Aware Transfer (WAT), has been experimentally validated to achieve the desired results. Our contributions can be summarized as follows.

- We introduce a trajectory-based framework for diffusion model transfer that leverages optimal transport theory to quantify distributional discrepancies between models trained on different datasets. This framework is theoretically grounded and empirically validated. It reveals how pretrained diffusion models can improve generative performance under distribution shifts and enables a Wasserstein-aware transfer strategy.
- We propose a category-wise matching strategy using optimal transport to explicitly reduce class-conditional distributional gaps. Unlike classifier-free guidance, our principled linear combination of pretrained and fine-tuned conditional outputs facilitates effective class-specific knowledge transfer without altering the fine-tuning pipeline.
- We demonstrate consistent performance gains over strong baselines across multiple benchmarks, with ablation studies supporting the robustness and theoretical soundness of our approach.

WAT: Theory and Method

Preliminaries

Diffusion Models. Denoising Diffusion Probabilistic Models (DDPMs) (Ho, Jain, and Abbeel 2020) are a class of generative models that have recently achieved remarkable success in image synthesis and related tasks. The forward process progressively adds Gaussian noise to data over T discrete time steps, transforming a data sample $\mathbf{x}_0$ into pure noise $\mathbf{x}_T$. Specifically, the forward process is defined as:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}),$$

where $\{\beta_t\}_{t=1}^T$ is a variance schedule. By applying the reparameterization trick, we can express $\mathbf{x}_t$ as:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}), \tag{1}$$

where $\alpha_t = 1 - \beta_t$, and $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$ is the cumulative product of α_t over time. The generative process learns a neural network to approximate the reverse conditional distributions $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$, which are also modeled as Gaussians:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)).$$

The model is trained by minimizing a variational bound on the negative log-likelihood, which simplifies to a weighted mean squared error loss when predicting the noise added at each diffusion step. Recent advances, such as the Diffusion Transformer (DiT) (Peebles and Xie 2023), replace the traditional U-Net backbone with a Transformer that operates in the latent space induced by a pretrained VAE, enabling more scalable and semantically meaningful generation. In this work, we build on DiT and perform domain transfer and class-level alignment directly within its latent space.

Classifier-free Guidance. Classifier-free guidance (CFG) (Ho and Salimans 2022) is a widely used technique to enhance conditional generation in diffusion models without relying on an external classifier. During training, the model is jointly optimized on both conditional and unconditional objectives by randomly dropping the conditioning label with a fixed probability p_{drop}. Let $\epsilon_\theta(\mathbf{x}_t, t, y)$ denote the model's prediction of the added noise given class label y, and $\epsilon_\theta(\mathbf{x}_t, t, \emptyset)$ the prediction without any condition. The training loss is then defined as:

$$\mathcal{L}_{\text{CFG}} = \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon},y'}\left[\|\boldsymbol{\epsilon} - \epsilon_\theta(\mathbf{x}_t, t, y')\|^2\right], \tag{2}$$

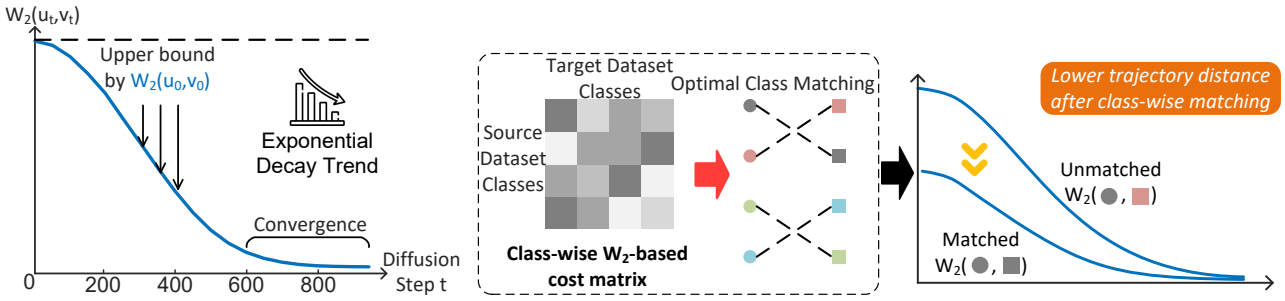
where $y' = y$ with probability $1 - p_{\text{drop}}$ and $y' = \emptyset$ with probability p_{drop}. This formulation allows the model to learn both conditional and unconditional denoising behaviors.

At inference time, the model predictions are interpolated to guide sampling towards the desired class label. Specifically, the noise prediction used during sampling is given by:

$$\epsilon_{\text{CFG}}(\mathbf{x}_t, t, y) = (1+w)\cdot\epsilon_\theta(\mathbf{x}_t, t, y) - w\cdot\epsilon_\theta(\mathbf{x}_t, t, \emptyset), \tag{3}$$

where $w \geq 0$ is the guidance scale controlling the trade-off between diversity and fidelity. CFG has become a standard component in class-conditional diffusion pipelines due to its simplicity and effectiveness.

Wasserstein Distance. The Wasserstein distance (Villani et al. 2008; Cuturi 2013), also known as the Earth Mover's



(a) Distributional Decay Along Diffusion. (b) Class-Wise Matching Reduces Cross-Domain Trajectory Discrepancies.

Figure 1: Wasserstein trajectory analysis and class-wise matching strategy. (a): Diffusion trajectories from different distributions converge over time. (b): Class-wise matching further reduces trajectory-level discrepancies.

Distance (EMD), provides a geometrically meaningful way to measure the discrepancy between probability distributions. Given two distributions μ and ν defined over a metric space $\mathcal{X}$ with cost function $c(x, y)$ (typically $c(x, y) = \|x - y\|^p$), the p-Wasserstein distance is defined as:

$$W_p(\mu, \nu) = \left(\inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} c(x, y)\, d\gamma(x, y) \right)^{1/p}, \quad (4)$$

where $\Pi(\mu, \nu)$ denotes the set of all joint distributions (couplings) with marginals μ and ν. Intuitively, it quantifies the minimum cost required to transform one distribution into another by moving "mass" between points in the space.

Wasserstein Distance in Diffusion Model

For diffusion model transfer tasks—or more broadly, for fine-tuning diffusion models on new datasets, Eq. (1) clearly shows that the difference in diffusion trajectories between the source and target datasets mainly stems from the discrepancy in the data distribution of $\mathbf{x}_0$, since both models share the same noise schedule β_t. Therefore, to better understand the essence of fine-tuning a pretrained diffusion model on a new dataset, we first investigate how the diffusion trajectories differ when the data distributions vary but the noise schedule β_t remains the same. To analyze the evolution of stochastic trajectories, we rely on the following theorem applicable to general stochastic processes.

Theorem 1 *(Bouvrie and Slotine 2019) Consider the Itô stochastic differential equation:*

$$d\mathbf{X}_t = f(\mathbf{X}_t)\, dt + \sigma(\mathbf{X}_t, t)\, d\mathbf{B}_t, \quad t \in [0, T], \quad \mathbf{X}_0 \sim \mu_0,$$

where $\mathbb{E}\|\mathbf{X}_0\|^2 < \infty$, $\mathbf{B}_t \in \mathbb{R}^d$ is a standard Brownian motion, and $\mathbf{X}_t \in \mathbb{R}^d$. Assume the following conditions hold:

1. *The drift function f is contracting in some uniformly positive definite metric $\mathbf{M}(t)$ with contraction rate $\beta > 0$, where $\mathbf{M}(t)$ satisfies $\mathbf{x}^\top \mathbf{M}(t)\mathbf{x} \geq \alpha\|\mathbf{x}\|^2$, for some $\alpha > 0$.*
2. *The diffusion satisfies $\operatorname{tr}\left(\sigma(\mathbf{x}, t)^\top \mathbf{M}(t)\sigma(\mathbf{x}, t)\right) \leq C_\sigma$, for all $\mathbf{x}$ and $t \in [0, T]$.*
3. *The diffusion matrix satisfies the coercivity condition:*
$$\langle \sigma(\mathbf{x}, t)\sigma(\mathbf{x}, t)^\top \mathbf{y}, \mathbf{y} \rangle \geq c\|\mathbf{y}\|^2$$
for some constant $c > 0$ and all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, $t \in [0, T]$.
4. *There exist constants $K_1, K_2 > 0$ such that for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and $t \in [0, T]$,*
$$\|f(\mathbf{x})\| + \|\sigma(\mathbf{x}, t)\|_F \leq K_1(1 + \|\mathbf{x}\|),$$
$$\|f(\mathbf{x}) - f(\mathbf{y})\| + \|\sigma(\mathbf{x}, t) - \sigma(\mathbf{y}, t)\|_F \leq K_2\|\mathbf{x} - \mathbf{y}\|.$$

Then, for any two solutions $\mathbf{X}_t$ and $\mathbf{Y}_t$ to the SDE with respective distributions μ_t and ν_t, the following holds:

$$W_2(\mu_t, \nu_t) \leq \alpha^{-1/2} e^{-\beta t} W_2(\mu_0, \nu_0) + \sqrt{C_\sigma/\beta}, \quad \forall t \in [0, T].$$

For the diffusion process in DDPM, its equivalent SDE formulation is given as follows (Song et al. 2020):

$$d\mathbf{x}_t = -\frac{1}{2}\beta(t)\mathbf{x}_t dt + \sqrt{\beta(t)} d\mathbf{B}_t.$$

This equivalent SDE satisfies all the conditions of the aforementioned theorem, as detailed in the appendix. Therefore, the diffusion trajectories of DDPM also exhibit the stochastic contraction property in Wasserstein distance. From this result, we observe that in the context of fine-tuning diffusion models on a new dataset, the distributional discrepancy between datasets provides an upper bound on the difference between the two diffusion processes. Moreover, this upper bound decreases as time t increases.

To further illustrate the above phenomenon, we visualize the difference between diffusion paths (as defined in Eq. (1)) for two different datasets in the latent space provided by a VAE, as shown in Fig. 1(a). It can be observed that, the divergence between diffusion trajectories decreases almost monotonically over time t, which can be characterized by:

$$W_2(\mu_t, \nu_t) \searrow \quad \text{as} \quad t \to T.$$

Specifically, experimental observations show that near the noisy end, the distribution difference between the two paths becomes minimal. This indicates that within a certain temporal range, the pretrained model exhibits a degree of knowledge generalization. This finding will also be validated in our analytical experiments. In fact, in the latent space of a VAE, the data distribution approximately follows a Gaussian distribution. Under this Gaussian assumption, we can derive the following monotonicity property.

Theorem 2 *Let $\mathcal{N}(\mu, \Sigma)$ and $\mathcal{N}(\mu', \Sigma')$ be two Gaussian distributions. Define their forward diffusion paths at time $t \in [0, 1]$ under a linear variance schedule $\alpha(t)$ as:*

$$x_t^{(1)} \sim \mathcal{N}(\sqrt{\alpha(t)}\mu,\ (1-\alpha(t))I + \alpha(t)\Sigma),$$

$$x_t^{(2)} \sim \mathcal{N}(\sqrt{\alpha(t)}\mu',\ (1-\alpha(t))I + \alpha(t)\Sigma'),$$

where $\alpha(t)$ is a strictly decreasing function on $[0, 1]$ with $\alpha(0) = 1$, $\alpha(1) = 0$. Let $W_2^2(t)$ denote the squared 2-Wasserstein distance between the marginal distributions $x_t^{(1)}$ and $x_t^{(2)}$. Then the function $W_2^2(t)$ is monotonically decreasing in $t \in [0, 1]$.

In summary, both the theoretical upper bound and the monotonicity observed along the diffusion paths indicate that the overall divergence in diffusion trajectories is primarily driven by differences in data distributions. Motivated by this, we propose to design fine-tuning strategies for diffusion model transfer from the perspective of minimizing the distribution shift, thereby enabling better reuse of knowledge learned by the pretrained model.

Wasserstein-Aware Transfer

In the context of diffusion model transfer, prior studies have shown that preserving the knowledge of a pretrained model can significantly improve generation quality after fine-tuning (Xie et al. 2023; Zhong et al. 2024, 2025). Moreover, the preceding theoretical analysis further suggests that pretrained models retain a certain degree of generalizable and reusable knowledge. To ensure that our method is compatible with standard fine-tuning pipelines and applicable to arbitrary conditional generation models—while also aiming to reduce distributional discrepancies, we begin with a class-wise matching strategy. An overview of the proposed approach is illustrated in Fig. 1(b).

Class matching. Based on the analysis above, our objective is to reduce the distributional gap between the source and target datasets. However, within the DiT framework, this cannot be achieved directly at the data level unless we retrain the entire VAE module. Such retraining would incur significant computational cost and risk erasing the valuable knowledge embedded in the pretrained model.

To address this, we shift our focus from minimizing overall distributional differences to aligning the categories between the two datasets. In typical fine-tuning scenarios, the diffusion model is adapted from a large-scale dataset to a smaller one, which involves a reduction in both the number of samples and the number of categories. Notably, only a subset of conditional branches in the pretrained model are usually utilized during fine-tuning on the target dataset.

Motivated by this observation, we first perform class-level matching between the source and target datasets before fine-tuning. This allows each conditional branch of the pretrained model to be matched with a target class that has a relatively similar data distribution, thereby ensuring smaller distributional shifts from a class alignment perspective and facilitating more effective transfer.

Let C_{ij} be the W_2 distance between the i-th source class and the j-th target class. To determine the optimal class-wise alignment, we minimize the following cost:

$$\min_{\sigma \in \mathrm{Perm}(n)} \frac{1}{n} \sum_{i=1}^{n} C_{i\sigma(i)}, \tag{5}$$

where $C_{ij} = W_2\left(\mu_i^{\text{src}}, \mu_j^{\text{tgt}}\right)$, and μ_i^{src} and μ_j^{tgt} denote the empirical distributions of the i-th source class and the j-th target class, respectively. The optimal permutation problem described above can be solved using the transport plan of the discrete W_1 distance, under the guarantee as follows.

Proposition 1 *(Peyré, Cuturi et al. 2019) Consider the transportation polytope:*

$$U(a, b) \coloneqq \left\{P \in \mathbb{R}_+^{n \times n} : P\mathbf{1}_n = a \text{ and } P^\top \mathbf{1}_n = b\right\},$$

where $\mathbf{1}_n$ denotes the n-dimensional vector of ones. The optimal transport problem is given by:

$$\mathcal{L}(a, b) \coloneqq \min_{P \in U(a,b)} \langle C, P\rangle = \min_{P \in U(a,b)} \sum_{i,j=1}^{n} C_{ij} P_{ij}. \tag{6}$$

When $m = n$ and $a = b = \mathbf{1}_n/n$, the following holds:

- *There exists an optimal solution P^* to Problem Eq. (6) that is a permutation matrix.*
- *This permutation matrix corresponds to an optimal permutation $\sigma^* \in \mathrm{Perm}(n)$ solving Eq. (5).*

In practice, we first sample from the pretrained model to obtain representative samples from the source dataset. To reduce computational cost, it is typically sufficient to sample fewer than 10 examples per class. Moreover, since the number of classes in the target dataset is usually small, it is unnecessary to sample from all classes in the source dataset. Based on the collected samples, we obtain the cost matrix using the class-wise W_2 distances, and then solve a discrete W_1 optimal transport problem to get the best class-wise matching. Since each class has been matched to a counterpart with a relatively small distributional discrepancy, we only need to remap the class labels according to the matching matrix before standard fine-tuning, while keeping the rest of the fine-tuning pipeline unchanged.

Class-level retention. After fine-tuning, we perform sampling using a variant of CFG designed to retain knowledge from the pretrained model. Since class-wise alignment has already been completed in the previous stage, it is theoretically justified to generate samples using a linear combination of the conditional outputs from both the pretrained and fine-tuned models, as detailed below.

$$\epsilon(\mathbf{x}_t, t, y) = (1+w)\cdot\epsilon_{\text{adapt}}(\mathbf{x}_t, t, y) - w\cdot\epsilon_{\text{pre}}(\mathbf{x}_t, t, y), \tag{7}$$

where ϵ_{adapt} denotes the fine-tuned model and ϵ_{pre} represents the pre-trained model, and the scalar weight w in Eq. (7) is referred to as the *class-level retention (CLR) scale*. For the conditional generation framework described above, we choose to discard the unconditional branch of the pretrained model (as adopted in methods like DoG (Zhong et al. 2025)) in favor of our proposed approach. This decision is primarily based on the following considerations:

1. Through an optimization-based process, accurate cross-class matching has been achieved, effectively minimizing

Algorithm 1: Class Permutation Matching

Require: Pretrained model ϵ_{pre}, target dataset $\mathcal{X}^t$
Parameter: Number of classes k, samples per class m
Ensure: Class permutation $\sigma : [k] \to [k]$

for $i = 1$ to k **do**
  Generate m synthetic samples $\mathcal{X}_i^s \leftarrow \epsilon_{\text{pre}}(\text{class} = i)$
  for $j = 1$ to k **do**
    Compute $C_{ij} \leftarrow W_2(\mathcal{X}_i^s, \mathcal{X}_j^t)$
  end for
end for
Solve optimal transport problem:

$$P \leftarrow \underset{P \in \mathbb{R}^{k \times k}}{\text{argmin}} \sum_{i,j} P_{ij} C_{ij}$$

subject to: $P\mathbf{1} = \frac{1}{k}\mathbf{1}, P^\top \mathbf{1} = \frac{1}{k}\mathbf{1}$
Compute permutation mapping:

$$\sigma(i) \leftarrow \underset{j \in [k]}{\text{argmax}}\, P_{ij}, \quad \forall i \in [k]$$

return σ

Algorithm 2: Wasserstein-Aware Transfer

Require: Pretrained model ϵ_{pre}; Target dataset $\mathcal{X}^t$
Parameter: Total fine-tuning steps N; Class-level retention scale w
Ensure: Adapted model ϵ_{adapt}; Generated samples $\{\hat{x}_i\}_{i=1}^M$

Stage 1: Class Alignment
Run Alg. 1 to obtain optimal class permutation σ
Stage 2: Label Remapping
for each target sample $(x, y) \in \mathcal{X}^t$ **do**
  Remap label: $\tilde{y} \leftarrow \sigma^{-1}(y)$
end for
Stage 3: Model Fine-tuning
for step $= 1$ to N **do**
  Sample a batch $\{(x_i, \tilde{y}_i)\}_{i=1}^B \sim \mathcal{X}^t$
  Compute loss: $\mathcal{L} = \|\epsilon_{\text{pre}}(x_i, \tilde{y}_i) - \epsilon_{\text{clean}}\|_2^2$
  Update ϵ_{pre} using gradient descent
end for
Set $\epsilon_{\text{adapt}} \leftarrow \epsilon_{\text{pre}}$
Stage 4: Sampling
Generate $\{\hat{x}_i\}_{i=1}^M$ using Eq. (7) with CLR scale w

the initial distribution discrepancy between the classes associated with each conditional branch.

2. Introducing the unconditional branch during conditional generation injects implicit identifiers of unconditional categories. This not only disrupts the alignment of distributions across matched conditional classes but also amplifies the original distribution discrepancy, i.e.,

$$W_2\left(\begin{bmatrix}\mu_i^{src}\\ \phi\end{bmatrix}, \begin{bmatrix}\mu_{\sigma(i)}^{tgt}\\ y\end{bmatrix}\right) \geq W_2\left(\begin{bmatrix}\mu_i^{src}\\ y\end{bmatrix}, \begin{bmatrix}\mu_{\sigma(i)}^{tgt}\\ y\end{bmatrix}\right).$$

Based on the above analysis, it is evident that CLR-based generation alone can effectively mitigate inherent distributional discrepancies. When combined with class-wise alignment, these discrepancies are further reduced, allowing for more efficient utilization of the pretrained model's knowledge. The detailed procedure of our method is illustrated in Algs. 1 and 2. Regarding the computation of the cost matrix, we use only a small number of samples per class (fewer than 10 in our experiments), making the calculation of Wasserstein-2 distances highly efficient. In practice, the entire process typically completes within one to two minutes. Therefore, the overall computational complexity of our method is on par with that of conventional fine-tuning.

Experiments

In this section, we evaluate our method across multiple datasets, focusing primarily on comparisons with the standard CFG technique and the state-of-the-art fine-tuning-based sampling method DoG (Zhong et al. 2025), as both rely on the conventional fine-tuning pipeline. Furthermore, we conduct comprehensive ablation studies to demonstrate the robustness of our approach and validate the theoretical foundations underpinning it. Additional experimental analyses and detailed discussions are provided in the Appendix.

Implementation Details

In line with prevailing practices in diffusion model transfer learning, we adopt the publicly available `DiT-XL/2` model, which has been pre-trained on the ImageNet dataset at a resolution of 256×256 for 7 million steps. This model achieves a Fréchet Inception Distance (FID) (Heusel et al. 2017) of 2.27, reflecting strong generative capabilities. We evaluate our method across seven fine-grained downstream datasets: `Food-101` (Bossard, Guillaumin, and Van Gool 2014), `SUN397` (Xiao et al. 2010), `DF20-Mini` (Picek et al. 2022), `Caltech-101` (Griffin et al. 2007), `CUB-200-2011` (Wah et al. 2011), `ArtBench-10` (Liao et al. 2022), `Stanford Cars` (Krause et al. 2013). Among these, five datasets are commonly used benchmarks in computer vision. Notably, `DF20-Mini` comprises fungal species images with no class overlap with ImageNet, while `ArtBench-10` features artistic images with markedly different distributions from natural images. This diverse selection enables a rigorous evaluation under varying degrees of domain shift relative to the pre-training data. All fine-tuning experiments are conducted using the following training settings: batch size $= 32$, learning rate $= 1 \times 10^{-4}$, image resolution $= 256 \times 256$, training steps $= 24{,}000$. We follow the standard CFG strategy during fine-tuning, employing a label dropout probability of 10%. Each experiment is carried out on a single NVIDIA A100 80GB GPU, requiring approximately 4 hours to complete.

Following established evaluation protocols (Xie et al. 2023), we generate 10,000 samples per dataset. Sampling is conducted using 50 denoising steps and a fixed CLR scale of 0.5, consistent with CFG and DoG. These settings ensure fair and consistent comparison across all models and datasets. Following DoG, we compute evaluation metrics between the generated images and the test set, reporting both the widely used FID and the more recent $\text{FD}_{\text{DINOv2}}$ (Stein et al. 2023) to enable a more comprehensive assessment.

Method	Food	SUN	Caltech	CUB Bird	Stanford Car	DF-20M	ArtBench	Average FID
Fine-tuning (w/o guidance)	16.04	21.41	31.34	9.81	11.29	17.92	22.76	18.65
+ Classifier-free guidance	10.93	14.13	23.84	5.37	6.32	15.29	19.94	13.69
Domain guidance	9.25	11.69	**23.05**	**3.52**	**4.38**	12.22	16.76	11.55
Relative promotion	15.36%	17.27%	3.31%	34.45%	30.70%	20.08%	15.95%	19.59%
Wasserstein-Aware Transfer	**9.01**	**9.76**	23.40	3.55	4.43	**11.55**	**13.43**	**10.73**
Relative promotion	17.57%	30.93%	1.85%	33.89%	29.91%	24.46%	32.65%	**21.62%**

Table 1: Comparisons on downstream tasks with pre-trained DiT-XL-2-256x256, where *Relative Promotion* denotes the percentage improvement over CFG. FID ↓

Method	Food	SUN	Caltech	CUB Bird	Stanford Car	DF-20M	ArtBench	Average $\text{FD}_{\text{DINOv2}}$
Fine-tuning (w/o guidance)	626.90	796.77	551.69	421.29	351.97	594.50	337.87	501.48
+ Classifier-free guidance	423.90	653.19	416.78	198.12	219.25	326.77	291.23	363.58
Domain guidance	351.93	620.58	392.92	140.00	134.15	151.39	257.39	292.62
Relative promotion	20.0%	5.0%	5.7%	29.3%	38.8%	53.7%	11.62%	23.4%
Wasserstein-Aware Transfer	**307.85**	**510.12**	**381.40**	**128.71**	**114.94**	**138.48**	**222.48**	**257.71**
Relative promotion	27.38%	21.90%	8.49%	35.03%	47.58%	57.62%	23.61%	**29.12%**

Table 2: Comparisons on downstream tasks with pre-trained DiT-XL-2-256x256. $\text{FD}_{\text{DINOv2}}$ ↓

Results

Experimental results across seven datasets are presented in Tables 1 and 2, where *Relative Promotion* denotes the percentage improvement over CFG. It is worth noting that we do not include results from methods such as DiffFit (Xie et al. 2023) and Diff-Tuning (Zhong et al. 2024) in our comparisons, as these approaches do not adhere to the standard fine-tuning protocol. In general, these methods either modify the training objective or adjust the optimization settings, making a direct and fair comparison difficult. Nevertheless, our method is orthogonal to and fully compatible with such approaches. It introduces no interference with their training dynamics and can be seamlessly incorporated as a modular component to further enhance their performance.

As shown in Table 1, our method consistently outperforms CFG in terms of FID across all datasets, with particularly notable improvements on most of them, although the gain on Caltech is less pronounced. Compared to the state-of-the-art method DoG, our approach achieves the best FID scores on four out of seven tasks, with substantial gains observed especially on SUN and ArtBench. This indicates that WAT can enhance generation quality even when the target dataset (e.g., ArtBench) exhibits a large distribution shift from the source. On average, our method improves the FID score by 0.82 over DoG (from 11.55 to 10.73). For the more recent metric $\text{FD}_{\text{DINOv2}}$, our method achieves clear and consistent improvements over both CFG and DoG. Specifically, we obtain a relative improvement of 29.12% over CFG and an absolute improvement of 34.91 over DoG (reducing from 292.62 to 257.71) in the average $\text{FD}_{\text{DINOv2}}$ score.

Analytical Experiments

Stochastic Contraction in W_2. As discussed in our methodology, the diminishing distributional differences along diffusion trajectories, along with the contraction behavior of the associated stochastic processes, suggest that pretrained models retain a degree of generalized knowledge near the noise end. This highlights how knowledge transfer works in diffusion models. To validate this hypothesis, we conduct two sets of analytical experiments.

1) Training Perspective: Motivated by the non-uniform time-step sampling strategies proposed in (Wang et al. 2025; Kim et al. 2025) and the empirical observations illustrated in Fig. 1(a), we observe that distributional differences between diffusion trajectories tend to converge within the time interval [600, 1000]. Based on this insight, we revise the fine-tuning strategy of the pretrained model. Specifically, during training, we restrict time-step sampling to the [0, 600] range, while maintaining CFG for sample generation. The experimental results are presented in Tab. 3, where a positive Δ indicates that restricting training to the [0, 0.6] interval yields better generative performance than standard full-range fine-tuning under the same CFG sampling strategy. Remarkably, the model achieves competitive performance even when fine-tuned on only a partial segment of the diffusion trajectory. Notably, for most datasets, focusing training on the sample-proximal end leads to improved generative quality. These results validate the soundness of both our methodological motivation and theoretical analysis.

2) Sampling Perspective: Unlike the truncated training strategy, we can also verify the reusability of pretrained models near the noise end directly from the sampling pro-

Method	Food	SUN	Caltech	CUB Bird
$[0, 600]$	11.24	13.35	24.99	5.01
Δ	-2.84%	+5.52%	-4.82%	+6.70%

Method	Standford Car	DF-20M	ArtBench
$[0, 600]$	5.98	13.22	19.71
Δ	+5.38%	+13.54%	+1.15%

Table 3: Results of training with time steps sampled only from [0, 600]. Δ indicates the difference in generative performance compared to full-time-step fine-tuning. FID↓

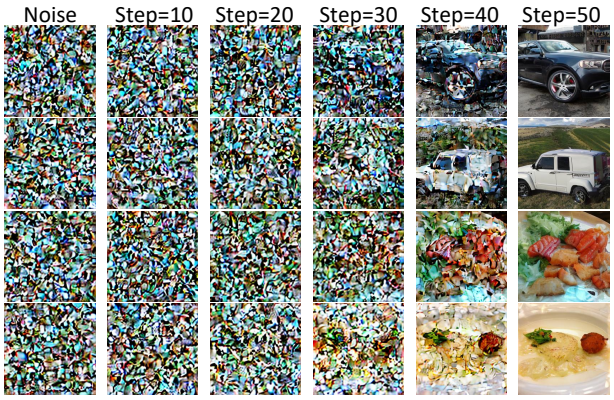


Figure 2: Generation results using a hybrid model: steps [0, 20] use the pretrained model, while steps [20, 50] are performed using the fine-tuned model.

cess. Specifically, we combine the fine-tuned and pretrained models: the fine-tuned model is applied during the denoising interval [0,600], while the pretrained model is used for the remaining steps in [600,1000]. The total number of sampling steps is 50. We present generation results on both the Stanford Cars dataset and food categories that are not present in ImageNet, with the latter sourced from the Food101 dataset. The overall image generation process is illustrated in Fig. 2. These results indicate that even when the pretrained model is employed during the early stages of sampling, it can still generate images aligned with the target dataset categories. Remarkably, this hybrid sampling strategy remains effective even when the target classes bear no relation to those seen during pretraining. This suggests that the pretrained model demonstrates strong generalization capability near the noisy end of the diffusion trajectory.

Sampling steps. Following the standard diffusion generation setting, we also evaluate the performance of our method under different sampling steps. As shown in Tab. 4, our approach consistently outperforms the conventional CFG sampling method, aligning with the effectiveness observed in DoG. Notably, on the SUN dataset, our method achieves the best FID scores across all sampling steps. Notably, our method consistently outperforms baselines in low-step sampling settings (e.g., 25 steps), demonstrating its effectiveness and robustness under constrained sampling budgets.

Sensitivity of the CLR scale. To investigate the impact of

Steps	CUB bird			SUN		
	CFG	DoG	WAT	CFG	DoG	WAT
25	9.69	4.60	**3.65**	24.34	19.87	**15.66**
50	5.37	**3.52**	3.55	14.13	11.69	**9.76**
100	4.27	**3.35**	3.66	10.07	8.71	**8.38**

Table 4: Results on varying sampling steps. FID ↓

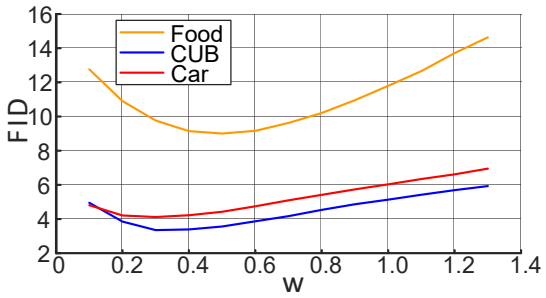


Figure 3: The impact of different CLR scales.

the CLR scale on generation quality, we conduct experiments with different CLR scales across three datasets, as shown in Fig. 3. The results indicate that performance peaks around a scale of 0.5 for all datasets. While we adopt a conventional setting of $w = 0.5$ in the table for a fair comparison with CFG, our method indeed achieves the best results near this value. Moreover, we observe that our method exhibits a similar sensitivity to the scale parameter as DoG. Specifically, when w is too large or too small, the generation quality deteriorates. This suggests that knowledge retention from the pretrained diffusion model should be maintained within a reasonable range, with the fine-tuned model playing the dominant role in generation.

Conclusion

In this paper, we investigate the underlying principles of knowledge transfer in diffusion models. Starting from the distributional differences in diffusion trajectories across varying data distributions, we leverage existing theoretical insights and empirical observations to identify the transferable and generalizable knowledge regions of a pre-trained model. Based on our analysis, we propose a data-level approach to mitigate distribution discrepancies by performing class-wise matching for each conditional branch of the diffusion model. Building on this matching, we introduce a novel sampling strategy that linearly combines the conditional branches of the pre-trained and fine-tuned models. Unlike classifier-free guidance, our approach does not rely on the use of an unconditional branch or require its separate training. This design not only reduces model complexity and parameter count but also shortens training time. Experimental results validate the effectiveness of our theoretical analysis and demonstrate the superiority of our proposed method.

Acknowledgments

This work is supported in part by National Key R&D Program of China (2024YFA1011900), National Natural Science Foundation of China (Grant No. 62376291), Guangdong Basic and Applied Basic Research Foundation (2023B1515020004), Science and Technology Program of Guangzhou (2024A04J6413), and the Fundamental Research Funds for the Central Universities, Sun Yat-sen University (24xkjc013).

References

Bossard, L.; Guillaumin, M.; and Van Gool, L. 2014. Food-101–mining discriminative components with random forests. In *European Conference on Computer Vision*, 446–461. Springer.

Bouvrie, J.; and Slotine, J.-J. 2019. Wasserstein contraction of stochastic nonlinear systems. *arXiv preprint arXiv:1902.08567*.

Corso, G.; Stärk, H.; Jing, B.; Barzilay, R.; and Jaakkola, T. 2022. Diffdock: Diffusion steps, twists, and turns for molecular docking. *arXiv preprint arXiv:2210.01776*.

Cuturi, M. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34: 8780–8794.

Griffin, G.; Holub, A.; Perona, P.; et al. 2007. Caltech-256 object category dataset. Technical report, Technical Report 7694, California Institute of Technology Pasadena.

Gu, X.; Yang, L.; Sun, J.; and Xu, Z. 2023. Optimal transport-guided conditional score-based diffusion model. *Advances in Neural Information Processing Systems*, 36: 36540–36552.

Hang, T.; Gu, S.; Li, C.; Bao, J.; Chen, D.; Hu, H.; Geng, X.; and Guo, B. 2023. Efficient diffusion training via min-snr weighting strategy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7441–7451.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30.

Ho, J.; Chan, W.; Saharia, C.; Whang, J.; Gao, R.; Gritsenko, A.; Kingma, D. P.; Poole, B.; Norouzi, M.; Fleet, D. J.; et al. 2022a. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information PProcessing systems*, 33: 6840–6851.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Ho, J.; Salimans, T.; Gritsenko, A.; Chan, W.; Norouzi, M.; and Fleet, D. J. 2022b. Video diffusion models. *Advances in Neural Information Processing Systems*, 35: 8633–8646.

Hoogeboom, E.; Satorras, V. G.; Vignac, C.; and Welling, M. 2022. Equivariant diffusion for molecule generation in 3d. In *International Conference on Machine Learning*, 8867–8887. PMLR.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Huang, Q.; Ko, T.; Zhuang, Z.; Tang, L.; and Zhang, Y. 2025. HiRA: Parameter-efficient hadamard high-rank adaptation for large language models. In *The Thirteenth International Conference on Learning Representations*.

Kim, M.; Ki, D.; Shim, S.-W.; and Lee, B.-J. 2025. Adaptive Non-Uniform Timestep Sampling for Accelerating Diffusion Model Training. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2513–2522.

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision workshops*, 554–561.

Li, M.; Zhai, Y.-M.; Luo, Y.-W.; Ge, P.-F.; and Ren, C.-X. 2020. Enhanced transport distance for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13936–13944.

Liao, P.; Li, X.; Liu, X.; and Keutzer, K. 2022. The artbench dataset: Benchmarking generative models with artworks. *arXiv preprint arXiv:2206.11404*.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4195–4205.

Peyré, G.; Cuturi, M.; et al. 2019. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6): 355–607.

Picek, L.; Šulc, M.; Matas, J.; Jeppesen, T. S.; Heilmann-Clausen, J.; Læssøe, T.; and Frøslev, T. 2022. Danish fungi 2020-not just another image recognition dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1525–1535.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494.

Song, Y.; Dhariwal, P.; Chen, M.; and Sutskever, I. 2023. Consistency models.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Stein, G.; Cresswell, J.; Hosseinzadeh, R.; Sui, Y.; Ross, B.; Villecroze, V.; Liu, Z.; Caterini, A. L.; Taylor, E.; and

Loaiza-Ganem, G. 2023. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. *Advances in Neural Information Processing Systems*, 36: 3732–3784.

Terpin, A.; Lanzetti, N.; Gadea, M.; and Dorfler, F. 2024. Learning diffusion at lightspeed. *Advances in Neural Information Processing Systems*, 37: 6797–6832.

Villani, C.; et al. 2008. *Optimal transport: old and new*, volume 338. Springer.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset.

Wang, K.; Shi, M.; Zhou, Y.; Li, Z.; Yuan, Z.; Shang, Y.; Peng, X.; Zhang, H.; and You, Y. 2025. A closer look at time steps is worthy of triple speed-up for diffusion model training. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 12934–12944.

Xiao, J.; Hays, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, 3485–3492. IEEE.

Xie, E.; Yao, L.; Shi, H.; Liu, Z.; Zhou, D.; Liu, Z.; Li, J.; and Li, Z. 2023. Difffit: Unlocking transferability of large diffusion models via simple parameter-efficient fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4230–4239.

Xu, M.; Yu, L.; Song, Y.; Shi, C.; Ermon, S.; and Tang, J. 2022. Geodiff: A geometric diffusion model for molecular conformation generation. *arXiv preprint arXiv:2203.02923*.

Zaken, E. B.; Ravfogel, S.; and Goldberg, Y. 2021. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv preprint arXiv:2106.10199*.

Zhong, J.; Guo, X.; Dong, J.; and Long, M. 2024. Diffusion tuning: Transferring diffusion models via chain of forgetting. *Advances in Neural Information Processing Systems*, 37: 114574–114600.

Zhong, J.; Zhang, X.; Wang, J.; and Long, M. 2025. Domain guidance: A simple transfer approach for a pre-trained diffusion model. *arXiv preprint arXiv:2504.01521*.