DogFit: Domain-guided Fine-tuning for Efficient Transfer Learning of Diffusion Models

Yara Bahram, Mohammadhadi Shateri, Eric Granger
LIVIA, ILLS, ETS Montreal, Canada
yara.mohammadi-bahram@livia.etsmtl.ca, {mohammadhadi.shateri, eric.granger}@etsmtl.ca

Abstract

Transfer learning of diffusion models to new domains with limited data is challenging, as naively fine-tuning the model often results in poor generalization. Test-time guidance methods help mitigate this by offering controllable improvements in image fidelity through a trade-off with sample diversity. However, this benefit comes at a high computational cost, typically requiring dual forward passes during sampling. We propose the Domain-guided Fine-tuning (`DogFit`) method, an effective guidance mechanism for diffusion transfer learning that maintains controllability without incurring additional computational overhead. `DogFit` injects a domain-aware guidance offset into the training loss, effectively internalizing the guided behavior during the fine-tuning process. The domain-aware design is motivated by our observation that during fine-tuning, the unconditional source model offers a stronger marginal estimate than the target model. To support efficient controllable fidelity–diversity trade-offs at inference, we encode the guidance strength value as an additional model input through a lightweight conditioning mechanism. We further investigate the optimal placement and timing of the guidance offset during training and propose two simple scheduling strategies, i.e., *late-start* and *cut-off*, which improve generation quality and training stability. Experiments on DiT and SiT backbones across six diverse target domains show that `DogFit` can outperform state-of-the-art guidance methods in transfer learning in terms of FID and $\text{FD}_{\text{DINOV2}}$ while requiring up to $\sim \times 2$ fewer sampling TFLOPS.

Code — https://github.com/yaramohamadi/DogFit

Introduction

Denoising diffusion probabilistic models (DDPMs) (Sohl-Dickstein et al. 2015; Ho, Jain, and Abbeel 2020) have emerged as powerful generative models, achieving state-of-the-art (SOTA) results in image synthesis (Dhariwal and Nichol 2021; Rombach et al. 2022), video generation (Gupta et al. 2024; Ho et al. 2022), and image editing (Kawar et al. 2023). However, producing high-quality and diverse outputs still requires substantial computational and data resources, particularly with large diffusion backbones and image sizes. In practical applications such as personalized generation with limited data, full model training from scratch is infeasible, making transfer learning a critical tool for target adaptation. A common approach in this setting is to fine-tune a pre-trained diffusion model on the target data. While straightforward, this strategy frequently suffers from overfitting and poor generalization due to the limited size and diversity of the target domain.

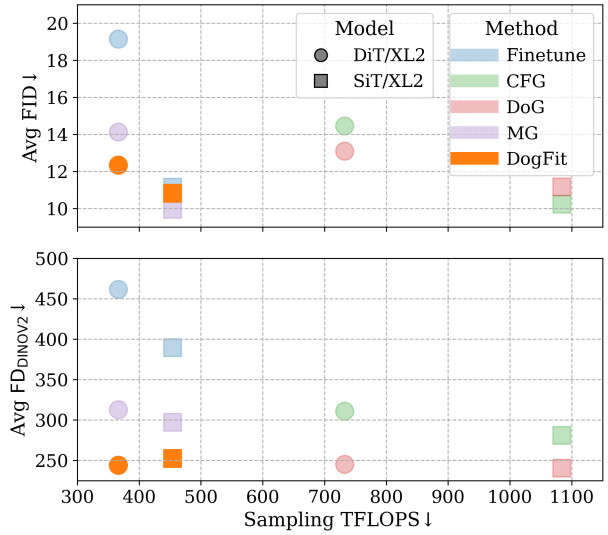


Figure 1: `DogFit` outperforms SOTA guidance methods in majority of cases for diffusion transfer learning, without increasing sampling overhead (TFLOPS). Evaluation is based on FID and $\text{FD}_{\text{DINOV2}}$ averaged across six target datasets.

Classifier-free guidance (CFG) is a standard technique for improving generation quality at inference time by modulating the conditional signal (Ho and Salimans 2022; Karras et al. 2024). This technique is often used in its original form even after target adaptation to steer the model towards better generalization. However, CFG assumes access to well-trained conditional and unconditional models, which is difficult to ensure in transfer scenarios where the target domain is small or lacks labels (Ho and Salimans 2022; Zhong et al. 2025). Recent methods like domain guidance (DoG) address these issues by utilizing the unconditional source model as a

stronger marginal noise estimator, offering improved target domain alignment and generalization (Zhong et al. 2025; Phunyaphibarn et al. 2025). Nevertheless, both CFG and DoG introduce significant computational overhead at sampling time due to their reliance on dual forward passes (Ho and Salimans 2022; Zhong et al. 2025). On the other hand, model guidance (MG) mitigates CFG's runtime cost in in-domain settings by injecting the guidance signal directly into the diffusion training objective (Chen et al. 2025; Tang et al. 2025). However, MG inherits the limitations of CFG in transfer learning due to the underfitting of the unconditional noise estimator (Zhong et al. 2025). Further, MG hard-codes the guidance strength value during training, restricting control at inference. These limitations raise a key question: *can we design effective guidance mechanisms for diffusion transfer learning without incurring computational overhead and still maintaining controllability over the guidance strength?*

This paper proposes domain-guided fine-tuning (`DogFit`), a method that injects a domain-aware guidance offset into the training loss during the fine-tuning process. This allows the diffusion model to directly learn the guided direction, removing the need for additional test-time processes. The domain-aware design is motivated by our observation that during fine-tuning, the source model offers stronger marginal noise estimates than the evolving target model. This design encourages the model to step toward the target domain manifold and avoid out-of-distribution generation. To support efficient control over the guidance strength at inference, `DogFit` conditions the model on the guidance value and exposes it to a range of such values during training. This allows the model to modulate its generation behavior accordingly, enabling fidelity–diversity trade-offs at inference with negligible sampling overhead. We further investigate the optimal placement and timing of the guidance offset during training. Two cost-effective guidance scheduling mechanisms are proposed for `DogFit`: (1) a *late-start* strategy that delays guidance until the model has learned sufficiently stable target representations; and (2) a *cut-off* scheme that restricts guidance to the later denoising steps, where fine-grained domain-specific details are more prevalent. Results over six datasets with different distribution shifts and supervision levels, using two SOTA diffusion backbones, DiT (Peebles and Xie 2023) and SiT (Ma et al. 2024), show that `DogFit` can outperform SOTA guidance methods in terms of FID and $\text{FD}_{\text{DINOV2}}$ while reducing their sampling TFLOPS by up to 2× (Fig.1). Results establish `DogFit` as a practical and efficient guidance method for diffusion transfer learning.

Contributions. **(i)** We propose `DogFit`, a training-time guidance mechanism for diffusion transfer learning that enables controllable improvement over target domain generation fidelity without requiring additional processes or double forward passes at test-time. **(ii)** We show that, during fine-tuning, the source domain model offers stronger marginal estimates than the target model, making it better suited for generating guidance signals. We empirically analyze how the timing and placement of this guidance impact training and propose two scheduling strategies that enhance training stability and generation quality. **(iii)** Extensive experiments are conducted across diverse target datasets on DiT and SiT diffusion backbones, showing that `DogFit` can outperform SOTA guidance methods in transfer learning.

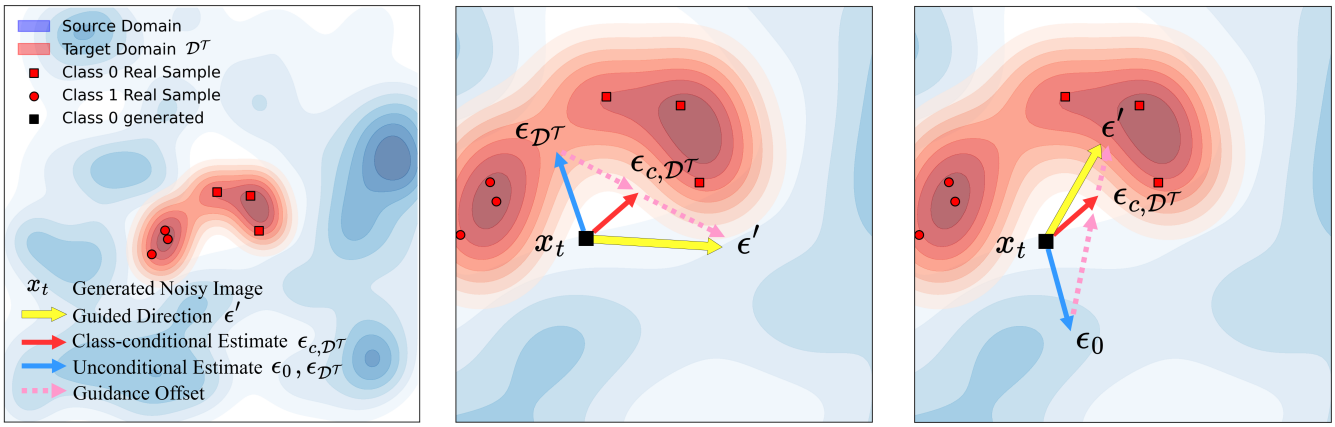
Related Work

Efficient Diffusion Models. The iterative denoising process in diffusion models imposes high sampling-time costs (Shen et al. 2025). To accelerate generation, prior work reduces the number of sampling steps via improved solvers (Song, Meng, and Ermon 2020; Lu et al. 2022), optimized noise schedules (Zheng et al. 2023), or distillation-based methods that train few(one)-step student models to mimic large multi-step teachers (Salimans and Ho 2022; Yin et al. 2024). Recent efforts (Jensen and Sadat 2025; Hsiao et al. 2024; Zhou et al. 2025) distill CFG to eliminate test-time dual passes. Our method, similar to CFG distillation, aims to internalize guidance, but does so without the architectural modifications or post-hoc training stages required by distillation methods. Instead, `DogFit` simply integrates guidance directly into the fine-tuning objective.

Diffusion Transfer Learning. Transfer learning adapts a model pre-trained on a large-scale source domain to a target domain with smaller size and diversity, leveraging learned representations for better generalization (Weiss, Khoshgoftaar, and Wang 2016). In the context of diffusion models, strategies broadly fall into three categories: Parameter-efficient fine-tuning (PEFT) methods reduce training cost by limiting the number of trainable parameters, often via adapters or LoRA (Hu et al. 2022; Xie et al. 2023; Moon et al. 2022). Distillation-based methods preserve pre-trained source priors during adaptation, particularly during early denoising steps (Zhong et al. 2024; Hur et al. 2024). Few-shot image generation methods enable diffusion models to generalize to unseen domains using a handful of images (Zhu et al. 2022; Wang et al. 2024; Cao and Gong 2024; Ouyang et al. 2024; Ruiz et al. 2023). Despite this progress, current methods overlook the role of guidance in shaping the generation trajectory, implicitly assuming that its original form remains optimal after adaptation, a limiting assumption in small, low-diversity target domains. We propose a new guidance method tailored for diffusion transfer that can be layered on top of existing fine-tuning methods for improving generalization without incurring additional sampling cost.

Proposed Method

Let ϵ_{θ_0} denote a diffusion model pre-trained on a large-scale source domain $\mathcal{D}^{\mathcal{S}}$, and let $\mathcal{D}^{\mathcal{T}}$ be a smaller target domain such that $|\mathcal{D}^{T}| \ll |\mathcal{D}^{S}|$ (typically, $1000 < |\mathcal{D}^{T}|$). The goal of transfer learning is to fine-tune a model ϵ_θ (initialized with the weights θ_0 of the pre-trained model) on $\mathcal{D}^{\mathcal{T}}$ such that it generates high-quality, diverse samples aligned with the target distribution $p(x|\mathcal{D}^{\mathcal{T}})$. In this work, we aim to develop a strong guidance mechanism for diffusion transfer learning that is computationally efficient and supports controllable guidance strength.



(a) Ground Truth. (b) CFG and MG Directions. (c) DoG and **DogFit** Directions.

Figure 2: Transfer learning via guidance mechanisms for synthetic data created using a mixture of Gaussian distributions. (a) The red region defines the target domain, while the blue background represents the source distribution. (b) CFG and MG prioritize class separability without considering the source distribution, often pushing samples toward out-of-distribution areas in the target domain. (c) **DogFit** and DoG utilize source information to emphasize movement towards the core of the target manifold, improving domain alignment. (*) Sampling-time guidance methods (DoG and CFG) operate by computing the guidance offset, whereas training-time guidance methods (**DogFit** and MG) learn the guided direction directly.

Preliminaries

Diffusion Models. Denoising Diffusion Probabilistic Models (DDPMs) (Ho, Jain, and Abbeel 2020) are generative models that learn to reverse a fixed noising process applied over T time-steps. Starting from a clean image x_0, noise is gradually added to produce a sequence of noisy images $\{x_t\}_{t=1}^{T}$. A neural network $\epsilon_\theta(x_t, t)$ is trained to predict the noise ϵ at each time-step t, using this denoising objective:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{t,x_0,\epsilon}\left[\|\boldsymbol{\epsilon}_\theta(x_t, t) - \boldsymbol{\epsilon}\|^2\right]. \tag{1}$$

Time-step t is omitted in subsequent sections when clear from context. In conditional generation, the model receives an auxiliary input c (e.g., class labels or text), producing $\epsilon_\theta(x_t|c)$. SOTA transformer-based diffusion architectures such as DiT (Peebles and Xie 2023) and SiT (Ma et al. 2024) operate in latent space rather than directly on pixel space. For notational simplicity, however, we keep using x to denote the model input.

Guidance in Transfer Learning. While diffusion models can functionally incorporate conditional inputs during training, their outputs often drift under weak conditioning, motivating the use of guidance signals during the reverse process (Dhariwal and Nichol 2021; Ho and Salimans 2022). We present three representative guidance strategies in transfer learning (i.e., CFG, MG, and DoG) covering the main design axes: *where* the guidance signal originates (unconditional branch, or external source model) and *when* it is injected (sampling-time vs. training-time). See visual comparison in Fig. 2.

- *Classifier-Free Guidance (CFG)* (Ho and Salimans 2022; Nichol et al. 2021) improves conditional generation by extrapolating the model's class-conditional prediction further away from its unconditional counterpart-typically using the same model. The reverse process is modified as:

$$\boldsymbol{\epsilon}'_\theta(x_t|c) = \boldsymbol{\epsilon}_\theta(x_t|c) + (w-1)\cdot(\boldsymbol{\epsilon}_\theta(x_t|c) - \boldsymbol{\epsilon}_\theta(x_t)), \tag{2}$$

where increasing the guidance scale $w > 1$ amplifies the effect of the condition during sampling. CFG poses limitations in transfer settings: The unconditional noise estimator is prone to underfitting due to joint training under limited target training data (Chen et al. 2023). Further, CFG often steers generations toward out-of-distribution regions in the target domain (Fig. 2 (b)), and it relies on conditionally labeled data, which may be unavailable in some domains (Zhong et al. 2025).

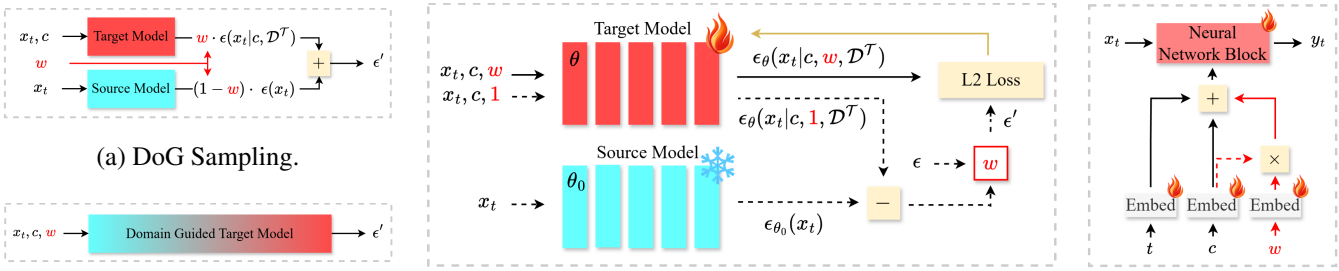
- *Domain Guidance (DoG)* (Zhong et al. 2025) offers a stronger alternative in transfer learning by using the unconditional source model as the marginal noise estimator:

$$\boldsymbol{\epsilon}'_\theta(x_t|c, \mathcal{D}^{\mathcal{T}}) = \tag{3}$$
$$\boldsymbol{\epsilon}_\theta(x_t \mid c, \mathcal{D}^{\mathcal{T}}) + (w-1)\cdot\left(\boldsymbol{\epsilon}_\theta(x_t \mid c, \mathcal{D}^{\mathcal{T}}) - \boldsymbol{\epsilon}_{\theta_0}(x_t)\right),$$

where $\boldsymbol{\epsilon}_\theta(x_t \mid c, \mathcal{D}^{\mathcal{T}})$ is the fine-tuned target model's class-conditional prediction and $\boldsymbol{\epsilon}_{\theta_0}(x_t)$ is the unconditional one of the source model. The source model, having been trained on rich data, provides a stronger reference point than CFG and further improves target domain alignment (Zhong et al. 2025) (Fig. 2 (c)). Like CFG, however, DoG requires two forward passes during sampling.

- *Model Guidance (MG)* (Tang et al. 2025) removes the need for CFG-style computation during sampling by incorporating its effect directly into the training objective. The model is trained to match a guidance-enhanced noise target:

$$\mathcal{L}_{\text{MG}} = \mathbb{E}_{t,(x_0,c),\epsilon}\,\|\boldsymbol{\epsilon}_\theta(x_t|c) - \boldsymbol{\epsilon}'\|^2, \tag{4}$$
$$\boldsymbol{\epsilon}' = \boldsymbol{\epsilon} + (w-1)\cdot \text{sg}\left(\boldsymbol{\epsilon}_\theta(x_t|c) - \boldsymbol{\epsilon}_\theta(x_t)\right),$$



(a) DoG Sampling.

(b) DogFit Sampling.

(c) DogFit Fine-tuning with w Controllability.

(d) w Conditioning.

Figure 3: Illustration of DogFit during training and sampling :(a–b) DogFit internalizes guidance behavior, removing the need for inference-time double forward passes. (c) During training, the guidance value is treated as an input, allowing inference-time control. The simple case of our method, with fixed guidance, simply requires removing the w and 1 from the target model conditions. (d) w is embedded it as an extra input during fine-tuning, allowing inference-time control. (*) Dashed lines denote paths that do not propagate gradients.

where $\mathrm{sg}(\cdot)$ denotes the stop-gradient operator and ϵ' marks the new noise target. While MG eliminates sampling-time overhead of test-time guidance, it inherits CFG's limitations in transfer learning due to the underfitting of the unconditional noise estimator. Further, it lacks a clear mechanism for allowing diversity-fidelity control at test time-a crucial principle in diffusion guidance.

Domain Guided Fine-Tuning

We propose DogFit, a training-time guidance mechanism for diffusion transfer learning that integrates a domain-aware guidance direction directly into the loss during fine-tuning. The marginal estimate of DogFit is derived from the unconditional source model (Fig.3 (a-b)). It modifies the training loss as:

$$\mathcal{L}_{\texttt{DogFit}} = \mathbb{E}_{t,(x_0,c),\epsilon} \left\| \boldsymbol{\epsilon}_\theta(x_t|c,\mathcal{D}^\mathcal{T}) - \boldsymbol{\epsilon}' \right\|^2, \tag{5}$$
$$\boldsymbol{\epsilon}' = \boldsymbol{\epsilon} + (w-1)\cdot \mathrm{sg}\left(\boldsymbol{\epsilon}_\theta(x_t|c,\mathcal{D}^\mathcal{T}) - \boldsymbol{\epsilon}_{\theta_0}(x_t)\right).$$

The training architecture is illustrated in Fig. 3(c). The training algorithm is shown in Appx. The rest of this subsection provides theoretical insights motivating the design of DogFit (full statements and proofs in Appx.)

Proposition 1. *Suppose a model is trained using the DogFit objective with controllable guidance strength w (Eq. 6), and the following assumptions hold: (i) the model approximately recovers the true noise when no guidance is applied, and (ii) its predictions vary linearly with w. Then, for any value of w, the model replicates the effect of applying DoG at sampling time.*

Proposition 2. *DogFit is equivalent to implicitly applying a domain alignment component to MG's training loss, benefiting from the class-conditional guidance of MG while reducing the risk of out-of-distribution generation.*

This means that ϵ_{θ_0} provides a strong and stable marginal noise estimate not only as an inference-time signal, but also during fine-tuning. Further, the domain alignment term, independent of class conditioning, pulls samples toward the core of the target data manifold, rendering DogFit effective even in the unconditional setting. In summary, it combines the strengths of previous guidance mechanisms in transfer learning into a single framework: it inherits the fast, single-pass sampling of MG and the robust marginal noise estimator of DoG. W

Incorporating Controllability

A crucial feature of guidance is its controllability. To enable efficient fidelity–diversity trade-offs at test time, we introduce a lightweight conditioning mechanism to the model and treat w as an additional input condition during fine-tuning. The training objective with added control is:

$$\mathcal{L}_{\texttt{DogFit}} = \mathbb{E}_{t,(x_0,c,\boldsymbol{w}),\epsilon} \left\| \epsilon_\theta(x_t|c,\boldsymbol{w},\mathcal{D}^\mathcal{T}) - \epsilon' \right\|^2, \tag{6}$$
$$\epsilon' = \epsilon + (\boldsymbol{w}-1)\cdot \mathrm{sg}\left(\epsilon_\theta(x_t|c,\mathbf{1},\mathcal{D}^\mathcal{T}) - \epsilon_{\theta_0}(x_t)\right),$$

where $\epsilon_\theta(x_t \mid c,\mathbf{1},\mathcal{D}^\mathcal{T})$ is the unguided class-conditional prediction. During training, the model must learn to operate across a range of guidance strengths to support controllable fidelity–diversity trade-offs at inference, while still preserving stable behavior in the unguided setting. Our initial experiments revealed that a simple uniform sampling of $\boldsymbol{w}$ during training can reduce diversity and destabilize generation, even at lower guidance levels due to excessive exposure to high w. We sample $\boldsymbol{w}$ from a shifted exponential decaying distribution (SEDD; see distribution CDF in Appx.):

$$\boldsymbol{w} = 1 + z, \quad z \sim \mathcal{P}(z), \tag{7}$$
$$\mathcal{P}(z) = \lambda e^{-\lambda z}, \quad z \geq 0, \tag{8}$$

where λ controls the decay rate. This distribution favors smaller guidance strengths while occasionally introducing larger values, enabling guidance robustness without destabilizing training. As a result, the model generalizes well across the guidance spectrum and supports dynamic controllability at inference (Fig. 5; see comparison with uniform scheduling in Appx.). The next key question is *how to actually incorporate w into the model?*

In conditional diffusion models, the final condition embedding is typically formed by summing the label c and

	Metric	Method	Unlabeled	Labeled						Sampling Cost	
			FFHQ	ArtBench	Caltech	CUB-Birds	Food	Stanford-Cars	Avg.	Passes	TFLOPS
DiT/XL-2	FID ↓	Fine-tuning	15.94	23.36	30.02	9.35	16.75	16.24	19.14	**x1**	**366.14**
		+ CFG (Ho and Salimans 2022)	–	20.83	24.07	5.03	11.77	10.60	14.46	x2	732.28
		+ DoG (Zhong et al. 2025)	13.87	17.30	23.76	**3.65**	10.97	9.77	13.09	x2	732.28
		MG (Tang et al. 2025)	–	19.91	23.71	4.85	11.13	11.03	14.13	**x1**	**366.14**
		DogFit	**10.48**	**16.32**	**21.68**	3.69	10.64	**9.35**	**12.34**	**x1**	**366.14**
		DogFit + Control	13.03	16.98	21.98	3.69	**10.44**	10.05	12.63	**x1**	**366.14**
	FD_{DINOV2} ↓	Fine-tuning	461.45	360.77	529.85	428.21	610.31	378.11	461.45	**x1**	**366.14**
		+ CFG (Ho and Salimans 2022)	–	314.32	401.96	200.92	410.48	227.46	311.03	x2	732.28
		+ DoG (Zhong et al. 2025)	**273.93**	266.25	377.49	**135.74**	314.16	**132.91**	**245.31**	x2	732.28
		MG (Tang et al. 2025)	–	299.06	409.20	220.40	379.39	255.88	312.78	**x1**	**366.14**
		DogFit	282.32	269.06	**377.15**	143.42	**293.24**	147.20	246.01	**x1**	**366.14**
		DogFit + Control	274.67	**261.72**	378.86	144.39	314.12	141.08	248.03	**x1**	**366.14**
SiT/XL-2	FID ↓	Fine-tuning	**7.63**	9.66	26.10	4.92	7.76	10.53	11.79	**x1**	**454.01**
		+ CFG (Ho and Salimans 2022)	-	9.28	23.21	3.49	7.95	9.71	10.73	x2	1083.77
		+ DoG (Zhong et al. 2025)	**7.63**	11.22	23.53	**3.39**	7.95	9.71	11.16	x2	1083.77
		MG (Tang et al. 2025)	-	9.64	**23.10**	3.60	6.84	**8.62**	**10.36**	**x1**	**454.01**
		DogFit	12.44	9.62	23.45	3.52	7.85	10.05	10.91	**x1**	**454.01**
		DogFit + Control	10.8	**9.03**	23.15	3.59	**6.52**	10.10	10.48	**x1**	**454.01**
	FD_{DINOV2} ↓	Fine-tuning	335.15	271.92	519.54	405.07	522.47	283.29	400.46	**x1**	**454.01**
		+ CFG (Ho and Salimans 2022)	-	236.13	406.63	203.47	368.26	190.24	280.94	x2	1083.77
		+ DoG (Zhong et al. 2025)	335.15	**215.12**	**387.89**	160.85	298.65	**139.09**	240.32	x2	1083.77
		MG (Tang et al. 2025)	-	232.17	418.86	229.32	359.90	203.23	288.70	**x1**	**454.01**
		DogFit	**278.10**	222.20	403.44	181.48	296.49	159.83	252.29	**x1**	**454.01**
		DogFit + Control	319.16	230.23	399.20	**152.50**	**234.98**	183.41	**240.06**	**x1**	**454.01**

Table 1: FID and FD_{DINOV2} performance of DogFit against SOTA methods using DiT/XL-2 and SiT/XL-2 backbones. Their sampling costs are also shown in terms of forward passes and TFLOPs.

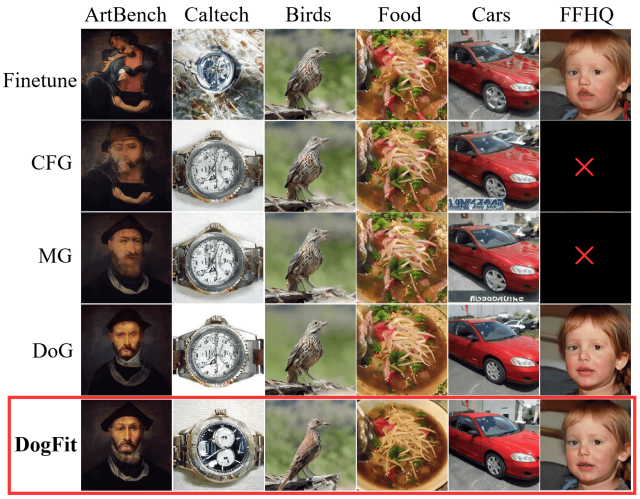


Figure 4: Qualitative comparison of guidance mechanisms on DiT-XL/2 (guidance scale 1.5). ×: not applicable.

time-step t embeddings. Therefore, a natural approach to add w is to embed it and sum it with the other existing embeddings. However, we saw that even slight distributional shifts and directional changes can destabilize training and disrupt the pretrained features. To avoid this, we use w as a label modulator: we embed it as a scalar, multiply it with a detached label c embedding, and add the resulting modulation vector to the condition embedding. This allows w to control the influence of the label without altering its semantic direction, preserving the integrity of the pretrained embedding space and ensuring stable guidance conditioning. For the unconditional case, we use the same formulation while keeping c as a fixed zero vector.

Guidance Scheduling Mechanisms

Inspired by test-time guidance scheduling for improved generation quality (Sadat et al. 2023; Kynkäänniemi et al. 2024), we investigate the optimal timing and placement of training-time guidance using two simple yet effective strategies. First, we introduce a *late-start* mechanism that delays the injection of guidance until iteration $s > \tau_s$, avoiding noisy updates and unstable gradients in the early phases of training. This is particularly beneficial in conditional settings where class embeddings are reinitialized to accommodate mismatched source and target label spaces. By deferring guidance, the model benefits from more stable target representations and condition embeddings, resulting in higher-fidelity updates. Second, we introduce a *cut-off* strategy that disables guidance at higher noise levels ($t > \tau_c$), concentrating domain supervision on the later denoising stages where adaptation to target-specific details is most effective (Choi et al. 2022). This design preserves source-driven diversity in the early denoising steps and prevents the model from over-relying on the target domain to shape the global structure

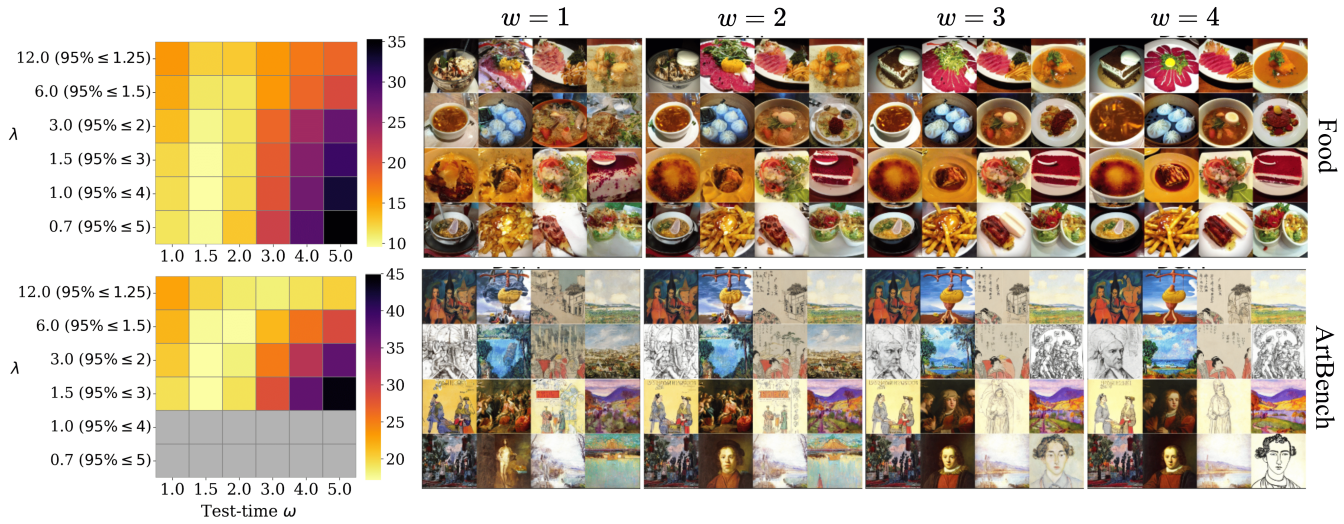


Figure 5: Effect of controllable guidance at test time for the Food and ArtBench domains. (Left) FID behavior across test-time ws with different training-time λs. (Right) Corresponding generated samples for fixed $\lambda = 3$ (95% of sampled w values in $[1, 2]$), across varying test-time w. Gray cells indicate extreme FID values (≈350) resulting from collapsed generations.

of the image. This becomes crucial when source and target domains share similar structural priors (e.g. FFHQ faces).

Results and Discussion

Setup. `DogFit` is validated on six challenging datasets with diverse domain characteristics and label availability for a comprehensive assessment: Food101 (Bossard, Guillaumin, and Van Gool 2014), Caltech101 (Griffin et al. 2007), CUB-200-201 (Wah et al. 2011), ArtBench10 (Liao et al. 2022), Stanford-Cars (Krause et al. 2013), and FFHQ256 (Karras, Laine, and Aila 2019). These datasets range from fine-grained species (Wah et al. 2011) to abstract art-styles (Liao et al. 2022) and face generation (Karras, Laine, and Aila 2019), covering a variety of distribution shifts in appearance, texture, and structure. We compare `DogFit` to standard fine-tuning, CFG, DoG, and MG. All methods are applied on DiT-XL/2 (Peebles and Xie 2023) and SiT-XL/2 (Ma et al. 2024) backbones pre-trained on ImageNet at 256×256 resolution for 7M steps with Fréchet Inception Distance (FID) (Heusel et al. 2017) scores of 2.27 (DiT) and 2.06 (SiT). We generate 10,000 images using 50 sampling steps (Peebles and Xie 2023; Xie et al. 2023), fixing guidance strength to 1.5 for fair comparison, and fine-tune for 24,000 training steps (Zhong et al. 2025) with batch size of 32 on 2 A100 GPUs for 5 hours. We calculate FID and $\text{FD}_{\text{DINOV2}}$ (Stein et al. 2023) between the generated images and the full datasets to evaluate the generation quality. Further, we compute Precision and Recall (Kynkäänniemi et al. 2019) for analyzing fidelity and diversity.

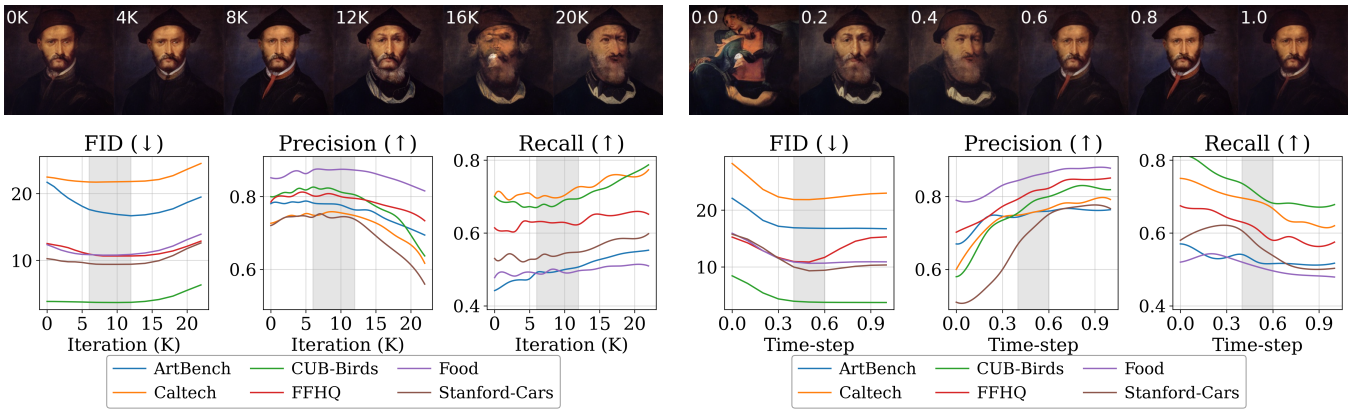
Comparison with State-of-the-art Methods

Tab.1 summarizes our main quantitative findings. `DogFit` consistently performs comparatively or superior to prior methods in both FID and $\text{FD}_{\text{DINOV2}}$, while maintaining the efficiency of one-pass sampling. Unlike CFG and MG, `DogFit` remains effective in unconditional settings like FFHQ due to its class-agnostic domain-alignment direction. On SiT, CFG and MG slightly outperform `DogFit` in FID, but `DogFit` outperforms both CFG and MG on $\text{FD}_{\text{DINOV2}}$, underscoring the value of utilizing diverse metrics for comprehensive evaluation. Importantly, equipping `DogFit` with controllable guidance introduces no performance degradation while increasing the parameter count by only 1.33 million parameters (less than 2% of the initial network parameters), leaving no notable impact on sampling TFLOPS. Additional results on Precision and Recall are provided in Appx. Fig. 4 presents a visual comparison of generated samples across target domains on DiT. Guidance-based methods consistently outperform naïve fine-tuning by producing more realistic and coherent images. `DogFit` generates sharp, domain-relevant details on par with or better looking than DoG, while avoiding off-distribution artifacts occasionally seen with CFG and MG (e.g., distorted faces or tires). SiT qualitative results are provided in Appx.

Ablation Studies

Controllable Guidance We study the effect of λ for sampling the guidance scalar w from the SED during training (Eq. 8). As shown in Fig. 5 (left), sparse exposure to higher guidance is sufficient for generalization and achieving a diversity-fidelity tradeoff. In contrast, excessive exposure to large guidance can harm the model's ability to retain pretrained features, leading to instability and collapsed generations. Based on these findings, we adopt $\lambda = 3$ (95% of sampled ws in $[1, 2]$) for all experiments. Fig. 5 (right) illustrates how varying guidance strength at test time effectively modulates the trade-off between fidelity and diversity, increasing sample quality while reducing variation with higher guidance values (more analysis in Appx.

Guidance Scheduling. We analyze the impact of late-start



(a) Late-start. **(b)** Cut-off.

Figure 6: Ablation study on guidance schedules in `DogFit`. (a) Varying the late-start threshold τ_s to control when guidance begins. (b) Varying the cut-off threshold τ_c to restrict guidance to later denoising steps. Performed using FID on DiT-XL/2.

# Steps	Food FID↓		Art FID↓	
	MG	DogFit	MG	DogFit
10	60.60	**57.91**	108.26	**88.84**
25	21.13	**18.97**	39.80	**29.11**
50	11.13	**10.64**	19.91	**16.32**
100	**7.60**	8.09	13.38	**12.70**

Table 2: Varying the number of sampling steps for MG and `DogFit`.

DiT-XL/2 Variant	FID↓		Train Params (M)
	Food	Art	
`DogFit`	**10.64**	16.32	675.42 (100%)
+ DiffFit	11.86	**15.80**	**0.75 (0.11%)**

Table 3: Using `DogFit` with PEFT.

and cut-off on generation quality and diversity. Fig. 6(a) shows how varying the late-start threshold τ_s impacts FID, Precision, and Recall. Delaying guidance improves FID and especially Precision at first, as early stabilization allows formation of more accurate directional cues. However, setting τ_s too high leaves insufficient time for the model to internalize guidance, reducing it to naïve fine-tuning and lowering precision. We set $\tau_s = 12$k for FID and $\tau_s = 6$k for $\text{FD}_{\text{DINOV2}}$ in all experiments. Fig. 6(b) shows the effect of cut-off threshold τ_c on FID, Precision, and Recall. Delaying guidance to later denoising steps enhances diversity, particularly for target datasets with strong global structural similarity to the source domain, such as FFHQ, Cars, and Caltech. However, setting τ_c too low limits exposure to guidance, again reducing the model to naïve fine-tuning. While the optimal τ_c may depend on the source–target similarity, we use $\tau_c = 0.5$ for FID and $\tau_c = 1$ for $\text{FD}_{\text{DINOV2}}$ in our experiments ($t \in [0, 1]$). See $\text{FD}_{\text{DINOV2}}$ analysis in Appx.

Domain-guided Parameter-efficient Fine-tuning. We demonstrate that `DogFit` is fully compatible with DiffFit, a leading PEFT method for DiT models (Xie et al. 2023). As shown in Tab.3, combining the two yields high-quality generations with lower cost in both training and test-time. This highlights `DogFit`'s practicality as a simple drop-in enhancement for efficient fine-tuning mechanisms.

Sampling Steps. Tab.2 reports the effect of varying the number of sampling steps during generation. `DogFit` outperforms MG in nearly all settings, with especially notable gains at lower step counts. This underscores the effectiveness of `DogFit`'s target alignment in low-resource scenarios where fast sampling is critical.

Conclusion

We introduce `DogFit`, a training-time guidance strategy for diffusion model transfer learning. It eliminates the high sampling-time cost of test-time guidance while preserving strong generalization behavior and controllability. `DogFit` leverages the pre-trained source model to inject high quality, domain-aligned guidance signals directly into the fine-tuning loss, and relies on a lightweight conditioning mechanism for inference fidelity–diversity adjustments. Our training-time guidance scheduling strategies further enhance training stability and generation quality. Extensive experiments across diverse target datasets and models show that `DogFit` can surpass the performance and efficiency of SOTA guidance methods, establishing it as a practical solution for scalable diffusion model transfer.

Limitations and Future Work. While effective, the proposed guidance scheduling strategies rely on manually chosen, fixed thresholds. Future work explores adaptive or data-driven approaches that respond to training dynamics or domain-specific properties. Moreover, extension of `DogFit` it to broader settings, i.e., text-to-image, audio, and video generation remains an exciting direction.

Acknowledgments

This research was supported by the Natural Sciences and Engineering Research Council of Canada, and the Digital Research Alliance of Canada.

References

Bossard, L.; Guillaumin, M.; and Van Gool, L. 2014. Food-101–mining discriminative components with random forests. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13*, 446–461. Springer.

Cao, Y.; and Gong, S. 2024. Few-shot image generation by conditional relaxing diffusion inversion. In *European Conference on Computer Vision*, 20–37. Springer.

Chen, H.; Jiang, K.; Zheng, K.; Chen, J.; Su, H.; and Zhu, J. 2025. Visual Generation Without Guidance. *arXiv preprint arXiv:2501.15420*.

Chen, M.; Huang, K.; Zhao, T.; and Wang, M. 2023. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In *International Conference on Machine Learning*, 4672–4712. PMLR.

Choi, J.; Lee, J.; Shin, C.; Kim, S.; Kim, H.; and Yoon, S. 2022. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11472–11481.

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.

Gupta, A.; Yu, L.; Sohn, K.; Gu, X.; Hahn, M.; Li, F.-F.; Essa, I.; Jiang, L.; and Lezama, J. 2024. Photorealistic video generation with diffusion models. In *European Conference on Computer Vision*, 393–411. Springer.

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.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing 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. 2022. Video diffusion models. *Advances in Neural Information Processing Systems*, 35: 8633–8646.

Hsiao, Y.-T.; Khodadadeh, S.; Duarte, K.; Lin, W.-A.; Qu, H.; Kwon, M.; and Kalarot, R. 2024. Plug-and-play diffusion distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13743–13752.

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.

Hur, J.; Choi, J.; Han, G.; Lee, D.-J.; and Kim, J. 2024. Expanding expressiveness of diffusion models with limited data via self-distillation based fine-tuning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5028–5037.

Jensen, C. P.; and Sadat, S. 2025. Efficient Distillation of Classifier-Free Guidance using Adapters. *arXiv preprint arXiv:2503.07274*.

Karras, T.; Aittala, M.; Kynkäänniemi, T.; Lehtinen, J.; Aila, T.; and Laine, S. 2024. Guiding a diffusion model with a bad version of itself. *Advances in Neural Information Processing Systems*, 37: 52996–53021.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4401–4410.

Kawar, B.; Zada, S.; Lang, O.; Tov, O.; Chang, H.; Dekel, T.; Mosseri, I.; and Irani, M. 2023. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6007–6017.

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.

Kynkäänniemi, T.; Aittala, M.; Karras, T.; Laine, S.; Aila, T.; and Lehtinen, J. 2024. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. *arXiv preprint arXiv:2404.07724*.

Kynkäänniemi, T.; Karras, T.; Laine, S.; Lehtinen, J.; and Aila, T. 2019. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32.

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

Lu, C.; Zhou, Y.; Bao, F.; Chen, J.; Li, C.; and Zhu, J. 2022. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35: 5775–5787.

Ma, N.; Goldstein, M.; Albergo, M. S.; Boffi, N. M.; Vanden-Eijnden, E.; and Xie, S. 2024. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, 23–40. Springer.

Moon, T.; Choi, M.; Lee, G.; Ha, J.-W.; and Lee, J. 2022. Fine-tuning diffusion models with limited data. In *NeurIPS 2022 Workshop on Score-Based Methods*.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

Ouyang, Y.; Xie, L.; Zha, H.; and Cheng, G. 2024. Transfer Learning for Diffusion Models. *arXiv preprint arXiv:2405.16876*.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4195–4205.

Phunyaphibarn, P.; Lee, P. Y.; Kim, J.; and Sung, M. 2025. Unconditional Priors Matter! Improving Conditional Generation of Fine-Tuned Diffusion Models. *arXiv preprint arXiv:2503.20240*.

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.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 22500–22510.

Sadat, S.; Buhmann, J.; Bradley, D.; Hilliges, O.; and Weber, R. M. 2023. CADS: Unleashing the diversity of diffusion models through condition-annealed sampling. *arXiv preprint arXiv:2310.17347*.

Salimans, T.; and Ho, J. 2022. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*.

Shen, H.; Zhang, J.; Xiong, B.; Hu, R.; Chen, S.; Wan, Z.; Wang, X.; Zhang, Y.; Gong, Z.; Bao, G.; et al. 2025. Efficient Diffusion Models: A Survey. *arXiv preprint arXiv:2502.06805*.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, 2256–2265. pmlr.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

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.

Tang, Z.; Bao, J.; Chen, D.; and Guo, B. 2025. Diffusion Models without Classifier-free Guidance. *arXiv preprint arXiv:2502.12154*.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, Pasadena, CA, USA.

Wang, X.; Lin, B.; Liu, D.; Chen, Y.-C.; and Xu, C. 2024. Bridging data gaps in diffusion models with adversarial noise-based transfer learning. In *Forty-first International Conference on Machine Learning*.

Weiss, K.; Khoshgoftaar, T. M.; and Wang, D. 2016. A survey of transfer learning. *Journal of Big data*, 3: 1–40.

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.

Yin, T.; Gharbi, M.; Zhang, R.; Shechtman, E.; Durand, F.; Freeman, W. T.; and Park, T. 2024. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6613–6623.

Zheng, K.; Lu, C.; Chen, J.; and Zhu, J. 2023. Improved techniques for maximum likelihood estimation for diffusion odes. In *International Conference on Machine Learning*, 42363–42389. PMLR.

Zhong, J.; Guo, X.; Dong, J.; and Long, M. 2024. Diffusion tuning: Transferring diffusion models via chain of forgetting. *arXiv preprint arXiv:2406.00773*.

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*.

Zhou, Z.; Chen, D.; Wang, C.; Chen, C.; and Lyu, S. 2025. DICE: Distilling Classifier-Free Guidance into Text Embeddings. *arXiv preprint arXiv:2502.03726*.

Zhu, J.; Ma, H.; Chen, J.; and Yuan, J. 2022. Few-shot image generation with diffusion models. *arXiv preprint arXiv:2211.03264*.