Glance: Accelerating Diffusion Models with 1 Sample

Zhuobai Dong[1*], Rui Zhao[2*], Songjie Wu[3], Suyang Hou[3], Junchao Yi[4], Zhengyuan Yang[5], Lijuan Wang[5], and Alex Jinpeng Wang[3✉]

[1] Wuhan University, China
[2] National University of Singapore, Singapore
[3] Central South University, China
[4] University of Electronic Science and Technology of China, China
[5] Microsoft, China
zhuobai@whu.edu.cn, jinpengwang@csu.edu.cn

Abstract. Diffusion models have achieved remarkable success in image generation, yet their deployment remains constrained by the heavy computational cost and the need for numerous inference steps. Previous efforts on fewer-step distillation attempt to skip redundant steps by training compact student models, yet they often suffer from heavy retraining costs and degraded generalization. In this work, we take a different perspective: we accelerate smartly, not evenly, applying smaller speedups to early semantic stages and larger ones to later redundant phases. We instantiate this phase-aware strategy with two experts that specialize in slow and fast denoising phases. Surprisingly, instead of investing massive effort in retraining student models, we find that simply equipping the base model with lightweight LoRA adapters achieves both efficient acceleration and strong generalization. We refer to these two adapters as Slow-LoRA and Fast-LoRA. Through extensive experiments, our method achieves up to **5× acceleration** over the base model while maintaining comparable visual quality across diverse benchmarks. Remarkably, the LoRA experts are trained with only 1 samples on a single V100 within one hour, yet the resulting models generalize strongly on unseen prompts. Code is available at https://zhuobaidong.github.io/Glance/.

Keywords: Diffusion models · Image generation · Efficient training

1 Introduction

Diffusion and flow matching models [2, 22, 33, 53, 55] have shown strong capabilities in generating high-fidelity images, marking a significant advancement in the field of generative modeling. Despite their impressive performance, a notable challenge is the high inference cost due to its iterative denoising nature. To address this issue, various methods are proposed to accelerate the sampling process

*Equal contribution, ✉Corresponding author.

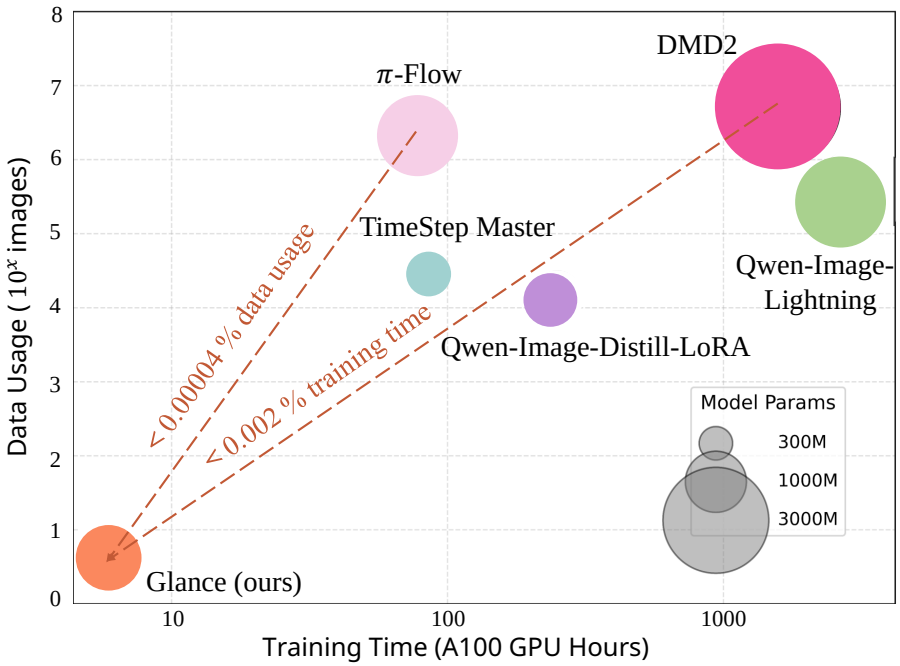

Fig. 1: Comparison of data usage and training time. Glance achieves comparable generation quality with only **1 training samples and within 1 GPU-hour**, demonstrating extreme data and compute efficiency. Note that the x-axis is in logarithmic scale, and values equal to zero are therefore not representable.

of diffusion models, including improving the efficiency of samplers [26, 34, 36, 37] and employing model distillation techniques [15, 16, 27, 35, 50, 54] to reduce the number of inference steps. Recent advancements in trajectory distillation methods and distribution matching techniques [51, 63, 64, 67], often enhanced by adversarial learning at scale, have shown considerable promise in generating high-fidelity images in extremely low steps such as one to four steps.

Despite significant advancements in timestep-distilled diffusion models, it remains unclear how to effectively fine-tune or customize such distilled models. Naively tuning the distilled model with diffusion loss will make the generation results blurry. An alternative approach is to fine-tune or customize the original diffusion model, and then repeat the diffusion distillation process to create a distilled model variant. However, the large computation cost of diffusion distillation, when compared with the customization training used for distillation (cf., 3840 A100 GPU hours for SDXL-DMD2 [63] and 3072 A100 GPU hours for Qwen-Image-Lightning [40]), often makes such distilled model tuning approach less feasible.

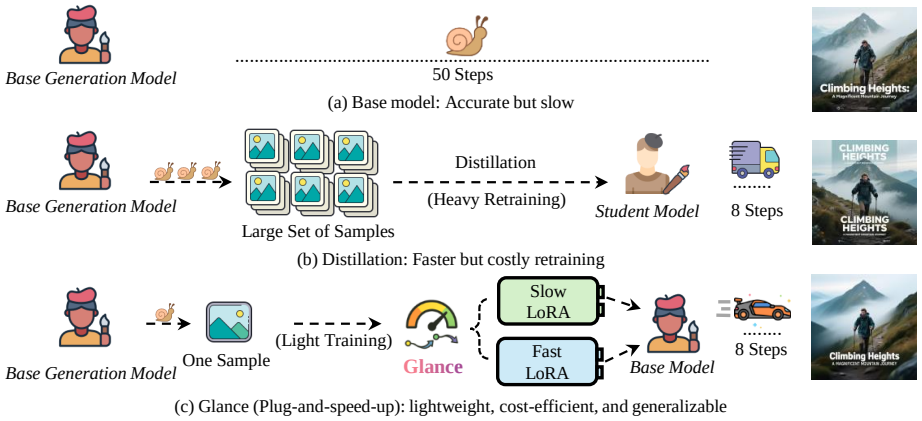

Fig. 2: Comparison of distill and accelerate strategies. Prior distillation pipelines rely on large training sets and costly retraining. Glance requires only one training sample to obtain Slow-LoRA and Fast-LoRA, providing plug-and-play acceleration of the base generation model.

In this work, we take a different perspective by revisiting the denoising dynamics of diffusion models. We observe that the generation trajectory consists of two qualitatively distinct phases: an early semantic phase that determines global structure, and a late redundant phase that primarily refines texture. Uniform acceleration treats all steps equally, yet semantic steps are far more sensitive to perturbation than redundant ones. This motivates a phase-aware acceleration strategy that applies small speedups to semantic steps and large speedups to redundant steps.

To realize this idea, we introduce Glance, implemented as a pair of lightweight LoRA adapters that attach to the pretrained diffusion model. Slow-LoRA stabilizes early semantic formation, while Fast-LoRA accelerates late-stage refinement. Crucially, our method does not require training a new student network and the base model remains unchanged. Both LoRA experts are trained using only **one sample** on a single V100 GPU within **one hour** (Fig. 1).

To demonstrate its scalability, we distill FLUX.1-12B [28] and Qwen-Image-20B [58] text-to-image models into 8- and 10-step students, respectively. Extensive experiments across six text-to-image benchmarks show that Glance exhibits performance curves that closely track those of the base models, indicating strong consistency under accelerated inference. On OneIG-Bench, HPSv2, and GenEval, the performance of Glance reaches **92.60%**, **99.67%**, and **96.71%** of the teacher models, respectively. We further conduct dense ablation studies on slow–fast phase decomposition, timestep allocation, and training data scaling, all of which consistently validate the effectiveness and robustness of our design.

We summarize the contributions as follows:

– We employ a phase-aware acceleration scheme that treats semantic and redundant steps differently for a more natural and stable speedup.

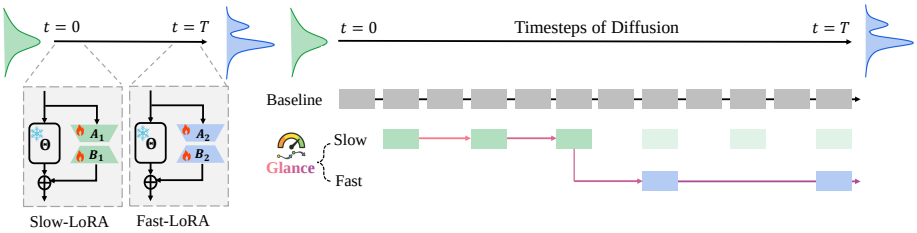

Fig. 3: Visualization of Slow-Fast paradigm. In the slow stage, we sample one timestep every two steps from the first 10 timesteps (i.e., 5 samples in total). In the fast stage, an additional 5 timesteps are uniformly sampled from the remaining 40 steps. During inference, the slow-stage timesteps are executed prior to the fast-stage ones.

– We introduce two lightweight adapters that plug into the base model, enabling effective acceleration with only one sample and one hour of training.
– Glance achieves a **5× speed-up** over FLUX.1 and Qwen-Image while retaining teacher-level performance across six benchmarks.

2 Related Work

Diffusion Models. Recently, diffusion models (DMs) [22,53,56] have become the leading paradigm for visual generation, achieving state-of-the-art performance across a wide range of conditioning modalities, including images [49], depth, edges, poses [41, 66], and text [11, 13, 21, 44, 46, 49]. Advances in large-scale systems such as PixArt [7–9], SD3 [14], Qwen-Image [58], and FLUX [28] further push generation quality, controllability, and multilingual rendering capabilities. Despite these rapid developments, achieving high-fidelity synthesis still requires many denoising steps, resulting in substantial inference cost and limiting their use in real-time or resource-constrained applications.

Diffusion Distillation. Early work [38] directly regresses the teacher's ODE integral in a single step, but the ℓ_2-based x_0 regression often produces overly smooth and blurry results. Progressive distillation methods [15, 35, 50] refine this paradigm via multi-stage training that enlarges step size and lowers NFE by merging teacher steps. While effective, these approaches suffer from error accumulation and substantial computational overhead. Consistency distillation [16,27,54] replaces x_0 regression with velocity-based objectives to enforce trajectory consistency, improving fidelity but requiring costly Jacobian–vector products (JVPs) or inaccurate finite-difference approximations. Distribution matching methods [51, 63, 64, 67]instead adopt score-based or adversarial objectives to align the student's output distribution with the teacher's, achieving high perceptual quality yet prone to mode collapse and instability due to auxiliary discriminators. Recent concurrent studies [6, 40, 57] have distilled Qwen-Image

and FLUX into compact 4-NFE or 8-NFE versions, but these still incur high distillation costs from trajectory-level supervision and extensive teacher sampling. In contrast, our approach achieves comparable or superior generation quality with minimal training cost, avoiding recursive distillation and auxiliary network overhead while maintaining stable optimization.

Efficient Tuning of Diffusion Models. LoRA [23] has been widely adopted for efficient adaptation of diffusion models via low-rank parameter updates [19, 25,29,39,42,47,60–62,66]. Several works integrate LoRA into distillation or controllable generation: DMD [65] enables faster inference through LoRA-based distillation; ControlNeXt [43] and TC-LoRA [10] provide adaptive, time- or condition-dependent LoRA control; Timestep Master [68] assigns LoRA experts to different noise levels for better representation. Recent models such as Qwen-Image-Lightning [40] and Qwen-Image-Distill-LoRA [57] embed LoRA into distilled backbones, producing compact low-step (4–8 NFE) models. However, their generation quality remains limited. We address this by allocating slow and fast LoRA adapters according to denoising phases, yielding substantial improvements.

3 Method

In this section, we introduce Glance, a phase-aware acceleration framework that improves both efficiency and adaptability of diffusion models through slow–fast paradigm. We first revisit the diffusion model and flow-matching formulation as preliminaries, then describe our phase-aware LoRA experts and their learning objectives.

3.1 Preliminary

Diffusion and Flow Matching. Diffusion models [22] learn data distributions by gradually transforming noise into data through a parameterized denoising process. The flow matching formulation [13, 35] interprets diffusion as learning a continuous velocity field that transports a sample from Gaussian noise $x_1 \sim \mathcal{N}(0, I)$ to clean data x_0. At timestep $t \in [0, 1]$, the intermediate state is defined as $x_t = tx_0 + (1 - t)x_1$, and the model predicts the transport velocity $v_\theta(x_t, t, h)$ conditioned on guidance h (e.g., text embedding). The objective is a mean-squared error between the predicted and target velocities:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{x_0, x_1, t, h}\left[\| v_\theta(x_t, t, h) - v_t \|_2^2\right],$$

where v_t is the groundtruth velocity. To achieve superior performance, the diffusion model is often designed with a large number of network parameters that are pre-trained on large-scale web data. Apparently, it is computationally expensive to distill such a big model for step reduction.

Low-Rank Adaptation. To alleviate the above difficulty, LoRA [23] has been recently applied for rapid distillation diffusion models on target data [4, 68].

Table 1: Quantitative comparisons on COCO-10k dataset and HPSv2 prompt set.

Model	Distill method	NFE	COCO-10k prompts					HPSv2 prompts		
			Data align.		Prompt align.		Pref. align.	Prompt align.		Pref. align.
			FID↓	pFID↓	CLIP↑	VQA↑	HPSv2.1↑	CLIP↑	VQA↑	HPSv2.1↑
FLUX.1 dev	-	50	27.8	34.9	0.268	0.900	0.309	0.284	0.805	0.314
FLUX Turbo	GAN	8	**26.7**	**32.0**	0.267	0.900	0.308	**0.286**	**0.814**	0.313
Hyper-FLUX	CD+Re	8	29.8	33.3	**0.268**	0.894	0.309	0.285	0.807	0.315
π-Flow (FLUX)	π-ID	8	29.0	35.4	**0.268**	**0.901**	**0.311**	0.285	0.810	**0.316**
Glance **(FLUX)**	Slow-Fast	8	34.2	40.1	0.259	0.879	0.298	0.276	0.895	0.297
Glance **(FLUX)**	Slow-Fast	10	30.4	37.5	0.265	0.891	0.303	0.282	0.799	0.303
Qwen-Image	-	50×2	34.1	45.6	0.282	0.936	0.312	0.302	0.872	0.309
Qwen-Image	-	10×2	39.1	52.1	0.265	0.918	0.287	0.288	0.835	0.305
Qwen-Image	-	8×2	44.2	56.3	0.263	0.873	0.269	0.287	0.761	0.292
Qwen-Image-Lightning	VSD	4	37.5	51.6	0.280	**0.935**	**0.322**	0.299	**0.867**	**0.328**
π-Flow (Qwen)	π-ID	4	**36.0**	**46.1**	**0.281**	0.934	0.314	**0.300**	0.860	0.310
Glance **(Qwen)**	Slow-Fast	8×2	38.9	52.4	0.273	0.926	0.289	0.287	0.842	0.302
Glance **(Qwen)**	Slow-Fast	10×2	37.8	50.3	0.280	0.932	0.313	0.297	0.849	0.308

Table 2: Quantitative comparisons on OneIG-Bench. * denotes unavailable results.

Model	Distill Method	Training Cost		NFE	Alignment↑	Text↑	Diversity↑	Style↑	Reasoning↑
		Data	GPU hour						
FLUX.1 dev	-	-	-	50	0.790	0.556	0.238	0.370	0.257
FLUX Turbo	GAN	1M	*	8	0.791	0.334	**0.234**	**0.370**	0.239
Hyper-FLUX	CD+Re	1.1M	800	8	0.790	**0.530**	0.198	0.369	0.254
π-Flow (FLUX)	π-ID	2.3M	83	8	**0.792**	0.517	**0.234**	0.369	**0.256**
Glance **(FLUX)**	Slow-Fast	1	0.6	8	0.774	0.284	0.196	0.353	0.208
Glance **(FLUX)**	Slow-Fast	1	0.8	10	0.788	0.328	0.204	0.358	0.231
Qwen-Image	-	-	-	50×2	0.880	0.888	0.194	0.427	0.306
Qwen-Image	-	-	-	10×2	0.802	0.693	0.156	0.410	0.290
Qwen-Image	-	-	-	8×2	0.752	0.611	0.148	0.411	0.276
Qwen-Image-Lightning	VSD	0.4M	3072	4	**0.885**	**0.923**	0.116	0.417	**0.311**
π-Flow (Qwen)	π-ID	2.3M	83	4	0.875	0.892	**0.180**	**0.434**	0.298
LoRA (Qwen)	Uniform steps	1	0.8	10×2	0.621	0.332	0.097	0.298	0.193
Glance **(Qwen)**	Slow-Fast	1	0.6	8×2	0.863	0.692	0.162	0.414	0.286
Glance **(Qwen)**	Slow-Fast	1	0.8	10×2	0.868	0.734	0.160	0.421	0.303

Specifically, LoRA introduces low-rank decomposition of an extra matrix, $\Theta' = \Theta + BA$, where $\Theta \in \mathbb{R}^{d\times k}$ denotes the frozen pretrained parameters, and the low-rank matrices $A \in \mathbb{R}^{r\times k}$ and $B \in \mathbb{R}^{d\times r}$ (with $r \ll d, k$) constitute the learnable LoRA parameters.

3.2 Phase-aware LoRA Experts for Phase-wise Denoising

To accelerate the denoising process of pretrained diffusion models while maintaining generative quality, we retain the pretrained parameters Θ and introduce a compact yet effective augmentation: a set of *phase-specific* LoRA adapters. Each adapter specializes in a specific stage of the denoising trajectory, enabling the model to adapt dynamically to varying noise levels and semantic complexities during inference.

Beyond uniform timestep partitioning. Prior works such as Timestep Master [68] have demonstrated the potential of using multiple LoRA adapters

trained over different timestep intervals. However, Uniform partitioning assumes equal contribution from all timesteps, which contradicts the intrinsic non-uniformity of diffusion dynamics. Empirical analyses and prior studies [3] reveal that different timesteps exhibit markedly different levels of semantic importance: in the early, high-noise regime, the model primarily reconstructs coarse global structures and high-level semantics (*low-frequency information*); in contrast, the later, low-noise regime refines textures and details (*high-frequency information*).

Phase-aware partitioning via SNR. To better align expert specialization with the intrinsic dynamics of the diffusion process, we introduce a *phase-aware* partitioning strategy guided by the signal-to-noise ratio (SNR). Unlike timestep indices, the SNR provides a physically meaningful measure of the relative dominance between signal and noise, and it decreases monotonically as denoising progresses. At the beginning of the process (t large, high-noise phase), the latent representation is dominated by noise with a low SNR, making coarse structural recovery the primary objective. In contrast, as t decreases and SNR rises, the model transitions into a low-noise regime focused on texture refinement.

Based on this observation, we define a transition boundary t_s corresponding to an SNR threshold (e.g., half of the initial SNR value). Two phase-specific experts are then employed: a *slow expert* specialized for the high-noise phase ($t \geq t_s$) that focuses on coarse semantic reconstruction, and a *fast expert* for the low-noise phase ($t < t_s$) that enhances fine-grained details. This SNR-guided partition allows each expert to operate in the regime where it is most effective, forming a semantically meaningful decomposition of the denoising process.

Surprising effectiveness of extremely small training sets. To evaluate whether phase-wise LoRAs can recover accelerated inference, we initially conducted an overfitting-style experiment using only **10 training samples**. Unexpectedly, the model rapidly learned a faithful approximation of the accelerated sampling trajectory. Even more remarkably, reducing the dataset to **a single training sample** still produced a stable acceleration behavior.

We attribute this data efficiency to the nature of flow matching. By directly predicting the target velocity field along the diffusion trajectory, the training objective bypasses redundant score-matching steps. Consequently, essential structural knowledge for fast inference can be distilled from only a few examples.

Necessity of carefully designed timestep skipping. Despite this promising data efficiency, subsequent ablation studies reveal that timestep skipping is far from arbitrary. Although few-step students can imitate the teacher behavior in aggregate, not all timesteps contribute equally to the reconstruction dynamics; naive skipping strategies can severely degrade performance.

To this end, we conducted a comprehensive investigation of different specialization schemes. We first explored assigning multiple timesteps to the slow stage LoRA adapters while keeping a single adapter for the fast stage, and vice versa. We also tested a degenerate configuration where a single LoRA was trained across the entire trajectory. However, these variants either lacked the expressiveness to capture high-noise complexity or failed to exploit temporal locality in the low-noise refinement phase.

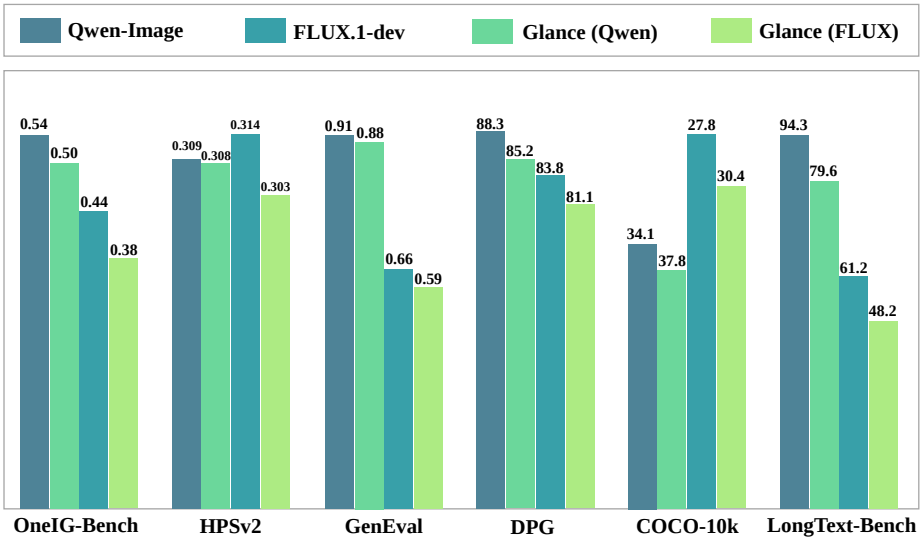

Fig. 4: Comparison between Image Generation Benchmarks. Glance further shows performance trajectories that closely follow those of the corresponding base models.

Our experiments ultimately show that separating the trajectory into a dedicated slow region and a dedicated fast region yields the most robust specialization. This design preserves sufficient capacity for modeling the challenging high-noise dynamics while enabling lightweight refinement in later steps, achieving a compact yet effective acceleration mechanism.

Flow-matching supervision. Each phase-specific LoRA expert is trained under a flow-matching supervision scheme that aligns its predicted denoising direction with the underlying data flow. We denote by $i \in 1, 2$ the index of the LoRA expert, corresponding to the two phase-specific adapters trained in our framework. Given the noisy latent x_t obtained during the diffusion process, the i-th LoRA expert predicts a velocity field $\hat{v}_{\Theta,B_i,A_i}(x_t, t, h)$, which is supervised against the ground-truth flow vector $v_t^\star$. The training objective is defined as follows:

$$\mathcal{L}_{\text{FM}}(t; i) = \mathbb{E}_{x,t,h}\left[\, w(t)\, \|\hat{v}_{\Theta,B_i,A_i}(x_t, t, h) - v_t^\star\|_2^2 \,\right],$$

where $w(t)$ denotes an optional timestep-dependent weighting function. By restricting the training samples of each expert to its assigned denoising phase, the model effectively learns to specialize on distinct noise levels. The resulting mixture of phase-aware LoRA experts collectively improves the denoising speed, forming the foundation of our proposed *slowfast* paradigm.

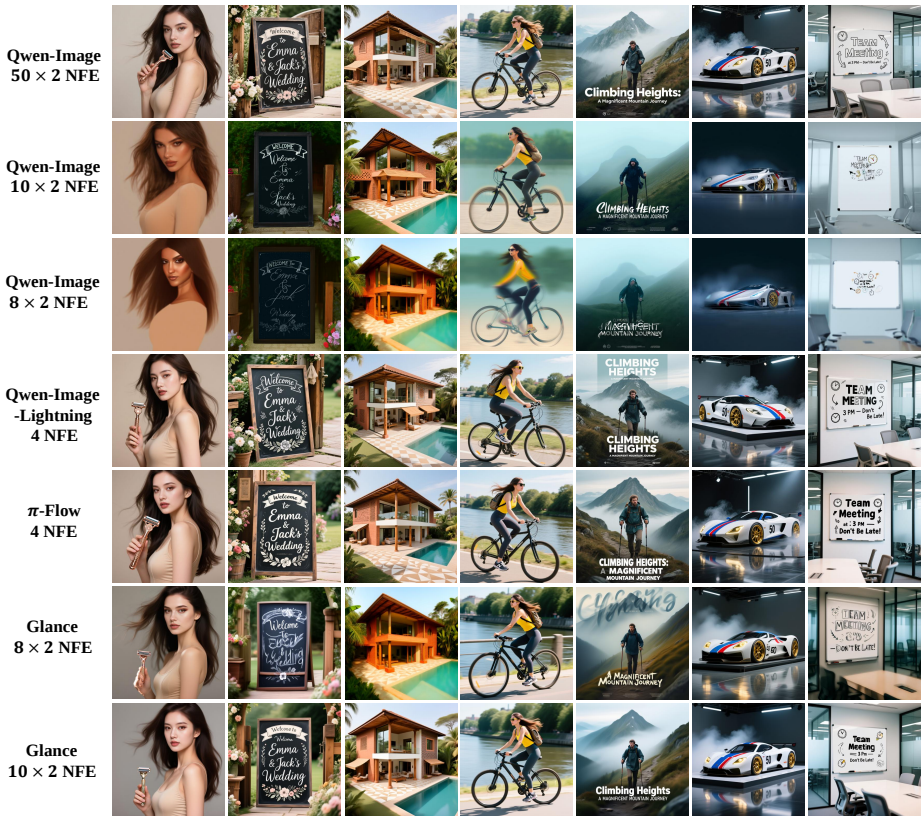

Fig. 5: Visual comparison of different Slow–Fast configurations. All images are generated from the same initial noise using the 50-step base model, our 8/10-step students, and other few-step models. Glance preserves semantic fidelity under strong acceleration, while additional steps progressively enhance fine details.

4 Experiments

This section presents a comprehensive evaluation of Glance on the text-to-image generation task. We first report quantitative results compared with competitive baselines, followed by detailed ablation analyses. We then discuss the generalization behavior of the model and its sensitivity to data scale.

4.1 Experimental Setup

Distillation Setup. We distill two large-scale text-to-image generators, FLUX.1-12B [28] and Qwen-Image-20B [58], into compact Slow-Fast students. During distillation, the base parameters inherited from the teacher are kept frozen, while only the LoRA adapters are optimized. Following Qwen-Image-Distill-LoRA [57],

we extend the adapter placement beyond the standard attention projections. Specifically, LoRA modules are injected not only into the query, key, value, and output projections, but also into auxiliary projection layers and modality-specific MLPs across both visual and textual branches. This broader integration allows the student to more effectively capture cross-modal dependencies and retain generation fidelity despite its compact capacity. For consistency, the results reported in Table 1 and 2 are both trained using the **fox sample** illustrated in Fig. 6. Moreover, the teacher baseline adopts FlowMatchEulerDiscreteScheduler [1], the official sampler of Qwen-Image.

Evaluation protocol. We conduct a comprehensive evaluation on 1024^2 high-resolution image generation from three distinct prompt sets: (a) 10K captions from the COCO 2014 validation set [31], (b) 3200 prompts from the HPSv2 benchmark [59], (c) 1120 prompts from OneIG-Bench [5], (d) 553 prompts from the GenEval benchmark [18], (e) 1065 prompts from the DPG-Bench [24], and (f) 160 prompts from the LongText-Bench [17]. For the COCO and HPSv2 sets, we report common metrics including FID [20], patch FID (pFID) [30], CLIP similarity [45], VQAScore [32], and HPSv2.1 [59]. On COCO prompts, FIDs are computed against real images, reflecting data alignment. On HPSv2, CLIP and VQAScore measure prompt alignment, while HPSv2 captures human preference alignment. For OneIG-Bench, GenEval, DPG-Bench, and LongText-Bench, we adopt their official evaluation protocols and report results based on their respective benchmark metrics.

4.2 Main Results

We compare Glance against other few-step student models distilled from the same teacher. For FLUX, we compare against: 8-NFE Hyper-FLUX [48], trained with consistency distillation (CD) and reward models (Re); 8-NFE FLUX Turbo [52], based on GAN-like adversarial distillation; π-Flow (FLUX) [6], trained with policy-based imitation distillation (π-ID). For Qwen-Image, we compare with the 4-NFE Qwen-Image Lighting based on variational score distillation (VSD) and π-Flow (Qwen). To further evaluate under extremely low-data conditions, we additionally implement a LoRA-based uniform timestep distillation approach that uniformly distill time steps using only 1 sample. The results are shown in Table 1 and 2. In Appendix A, we also explore the Qwen-Image-Edit model [58] in the image-editing domain.

Overall Performance. As illustrated in Figure 4, Glance exhibits performance curves that closely track those of the base models (Qwen-Image and Flux) across all benchmarks, indicating strong consistency under the accelerated setting. The detailed quantitative summaries in Table 1 and Table 2 confirm this trend: Glance achieves nearly the same generation quality as the 50-step base models while running **5× faster**. Despite the aggressive acceleration, it maintains competitive FID, CLIP, and HPSv2 scores across all benchmarks, showing no clear degradation in visual fidelity or prompt alignment. Moreover, even when trained with only **1 sample** within less than one GPU hour, Glance delivers results comparable to other few-step distillation approaches that require

large-scale datasets and heavy computation. These results underscore the strong **efficiency–performance balance** achieved by our approach, validating that the proposed Slow-Fast LoRA framework can preserve high generation quality under *minimal supervision and limited computational resources*.

Visual Fidelity. To further examine the trade-off between speed and quality, Figure 5 presents qualitative comparisons among our Glance models (8- and 10-step), the base teacher models (8-, 10-, and 50-step), and other 4-step distillation models such as Qwen-Image-Lightning and π-Flow, all under identical noise initialization. Even at only eight steps, Glance maintains the teacher's global semantics and color composition with minimal loss of fidelity. Increasing the step count gradually restores fine textures and small structures, indicating that phase-aware LoRA adaptation preserves the denoising trajectory of teacher despite extreme acceleration. These observations align with our quantitative findings: the 5× faster model achieves nearly the **same quality as the teacher** while requiring only a fraction of the computation.

4.3 Ablation Study

We conduct comprehensive ablation studies to analyze the key factors that contribute to the effectiveness of the proposed Slow–Fast design. Unless otherwise stated, all experiments are evaluated on OneIG-Bench using Qwen-image as the teacher model.

The importance of Slow-Fast Design. To verify the effectiveness of the proposed phase-aware design, we divide the diffusion process into two distinct denoising stages and systematically vary the LoRA assignment strategy. Specifically, we experiment with five configurations to assess the role of Slow-LoRA and Fast-LoRA under different timestep allocations. We experiment with five configurations: (1) phase-aware Slow3–Fast5 setup (ours), (2) Slow3 + Base5, (3) Base3 + Fast5, (4) single LoRA at identical timesteps, and (5) single LoRA uniformly sampled across eight timesteps.

As summarized in Table 3, our asymmetric Slow–Fast configuration achieves the **highest performance across all metrics**, demonstrating its superior balance between quality and efficiency. This confirms that aligning LoRA updates with the semantic-to-refinement progression of denoising leads to more effective **knowledge transfer**. In this process, the model performs slow adaptation during early stages and fast refinement during later stages, achieving better specialization than uniform or single-expert alternatives. Among all variants, the **Single (uniform)** setup performs the worst, confirming the necessity of phase-wise specialization. Notably, we observe that the early-stage Slow-LoRA contributes more significantly to final image quality, underscoring the importance of coarse-to-fine adaptation in guiding generation.

If more samples help? To investigate the effect of data composition on LoRA adaptation, we first randomly select 1 text–image pairs from the Qwen-Image-Self-Generated-Dataset [12] dataset as the minimal training set. We then scale up the dataset while keeping the total number of training epochs fixed. As shown in Table 4, increasing the number of training samples from 1 to 10 and further

Table 3: Slowfast stage ablation study.

Model	Alignment↑	Text↑	Diversity↑	Style↑	Reasoning↑
Slow3–Fast5	**0.849**	**0.614**	**0.152**	**0.396**	**0.284**
Slow3 + Base5	0.805	0.567	0.123	0.368	0.255
Base3 + Fast5	0.747	0.521	0.125	0.372	0.243
Single (identical)	0.702	0.453	0.110	0.342	0.218
Single (uniform)	0.621	0.332	0.097	0.298	0.193

to 100 does not lead to notable performance gains. Most metrics remain nearly unchanged, while the *style* score even slightly declines, suggesting that simply enlarging the dataset without enhancing its diversity or phase alignment may weaken stylistic consistency. These results indicate that for phase-aware LoRA adaptation, data quality and phase alignment are more crucial than scale, and that even a single well-chosen sample can achieve effective adaptation.

Table 4: Data ablation study

Model	Alignment↑	Text↑	Diversity↑	Style↑	Reasoning↑
1 sample	0.868	0.734	0.160	**0.421**	0.303
10 samples	0.874	**0.758**	0.163	0.414	0.296
100 samples	**0.876**	0.753	**0.165**	0.418	**0.306**

Table 5: Timestep ablation study

Model	Alignment↑	Text↑	Diversity↑	Style↑	Reasoning↑
Slow3 + Fast5	0.863	0.692	0.162	0.414	0.286
Slow5 + Fast5	0.868	0.734	0.160	0.421	0.303
Slow5 + Fast10	**0.874**	**0.813**	**0.175**	**0.422**	**0.305**

Timestep Ablation. In this experiment, we study the influence of the number of timesteps equipped with LoRA adapters while keeping the data selection and scale fixed at 1 text–image pairs. We progressively increase the number of timesteps on which LoRA modules are attached, examining how temporal coverage of adaptation affects generation quality. As shown in Table 5, with more LoRA-equipped timesteps, the overall performance steadily improves, particularly showing a clear gain in the *text* metric. This trend indicates that **broader temporal adaptation enables the model to capture more phase-specific denoising dynamics**, leading to improved reconstruction quality.

4.4 Discussion

Generalization Ability of Different Single Training Samples. We conduct an ablation study using four distinct single-image settings to examine the generalization behavior of our framework under extreme data scarcity. Specifically, we first select *in-distribution* samples from [12], where the samples are generated from Qwen-image base model. We select three sample with representing diverse semantic and structural characteristics: (1) **Fox**: an anthropomorphic fox with vibrant red fur and white facial features, (2) **Valley landscape**: a unique natural scene featuring a winding spiral valley, and (3) **Bookstore**: a text-rich

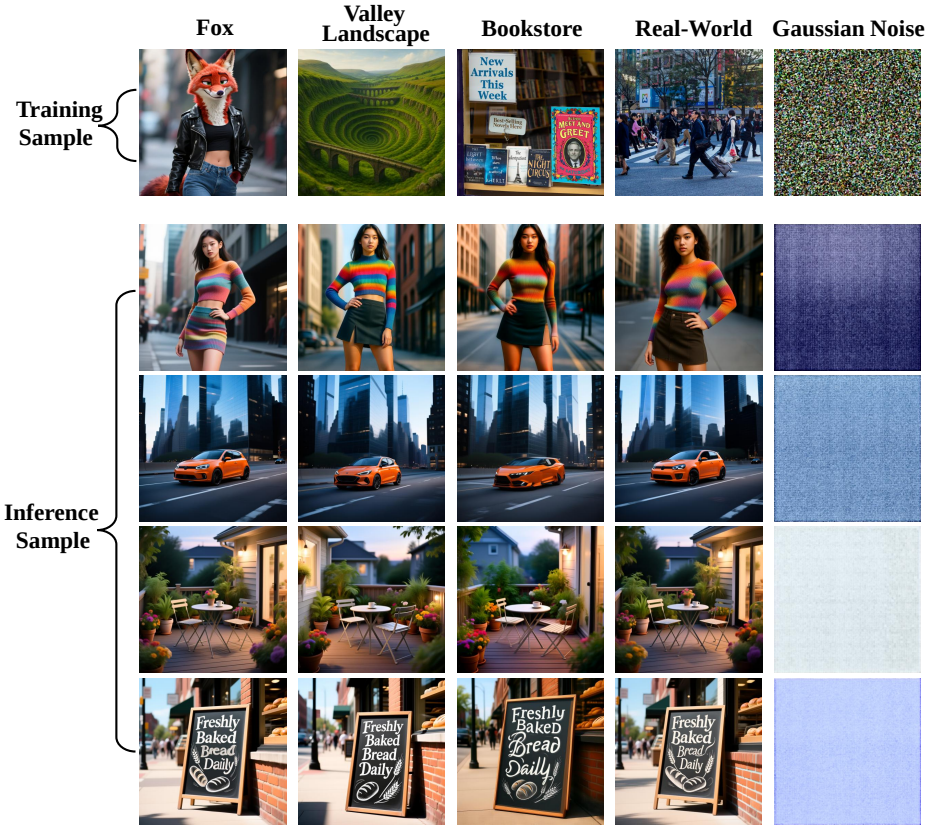

Fig. 6: Qualitative results from the one-sample training setting. Even trained on a single image, the model generalizes well to unseen prompts, producing coherent and detailed results across diverse scenes.

storefront window densely filled with books and signage. Additionally, an *out-of-distribution (OOD)* real-world image, depicting a bustling city crosswalk filled with pedestrians moving in different directions on a clear day, is included to assess robustness beyond the training domain. For completeness, we also experiment with an extreme setting where the model is trained purely on **Gaussian noise**, serving as a control to isolate the effect of meaningful visual content. All models are trained on different samples using the same configuration to assess their generalization ability.

From Table 6 and Fig. 6 we observe: *i*. While the three single-sample settings exhibit noticeable stylistic differences, their quantitative performance remains relatively close. *ii*. The Bookstore (text-rich) model exhibits the weakest generalization, while the fox-trained model demonstrates the strongest. The fox-based LoRA yields consistent improvements across all evaluation metrics, and Figure 6 further illustrates that it produces images with better convergence and richer

Table 6: 1 data sample study.

Model	Alignment↑	Text↑	Diversity↑	Style↑	Reasoning↑
Fox	**0.868**	0.734	**0.160**	**0.421**	**0.303**
Valley Landscape	0.842	0.712	0.146	0.409	0.299
Bookstore (text-rich)	0.797	**0.751**	0.131	0.373	0.267
Real-World (OOD)	0.857	0.728	0.153	0.420	0.298

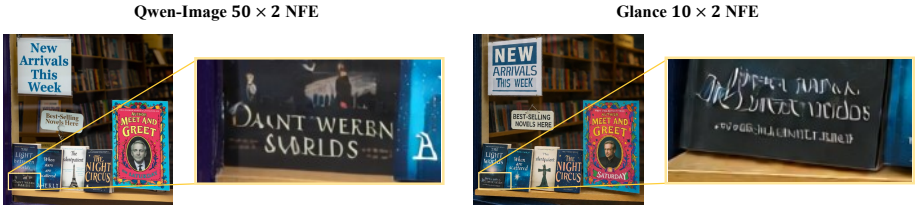

Fig. 7: Text-render failure cases. Glance struggles on extremely small text, producing blurred or distorted characters.

fine-grained details. However, the gap between them is modest. *iii.* Surprisingly, the Real-world image (OOD) model also achieves competitive results, with scores on most metrics only slightly lower than those of the Fox-based LoRA, suggesting that *out-of-domain samples can still provide meaningful transferable cues for denoising adaptation. iv.* When trained solely on Gaussian noise, the model fails to produce any meaningful images, indicating that effective denoising requires exposure to data that resembles natural image distributions.

Overall, while different samples lead to consistent performance trends, the quantitative differences are minor. This shows that Glance is not particularly sensitive to which single image is used for training, and can still learn transferable denoising behaviors from extremely limited but coherent supervision.

Failure Analysis. Although Glance closely tracks base model performance across all evaluated benchmarks, a consistent gap remains in text rendering. As shown in Table 2 and Fig. 4, Glance falls behind Qwen-Image and FLUX by 0.154 and 0.228 on the Text-Render metric, and by 14.7 and 13.0 points on LongText-Bench, respectively. To better understand this discrepancy, we conducted a detailed inspection of the generated samples. As illustrated in Fig. 7, failure cases predominantly occur in images containing extremely dense or very small text, where the model struggles to preserve sharp character boundaries, often producing blurred strokes or local artifacts. In contrast, when the text is shorter and occupies a larger spatial extent in the image, Glance is able to reproduce it faithfully. These suggest that high-frequency textual details are harder for the student to capture than general visual content, Such fine-grained details require precise spatial alignment and are harder to capture during distillation, especially under few-step constraints.

5 Conclusion

We present Glance, a lightweight distillation framework that accelerates diffusion inference through a phase-aware Slow–Fast design. The well-studied LoRA adapters for distinct denoising phases efficiently capture both global semantics and local refinements. Glance enables high-quality generation with only eight steps, achieving an **5× speed-up** over the base model. Despite being trained with as few as one image and a few GPU hours, Glance maintains comparable visual fidelity and exhibits strong generalization to unseen prompts. These results highlight that data- and compute-efficient distillation can retain the expressive capacity of large diffusion models without sacrificing quality. We believe Glance serves as a strong candidate for accelerating large-scale diffusion models, particularly in data-scarce applications.

Acknowledgements

This project is supported by National Natural Science Foundation of China (grant number 62502544).

References

1. Flowmatcheulerdiscretescheduler. `https://huggingface.co/docs/diffusers/api/schedulers/flow_match_euler_discrete` (2025), accessed: 2025-09-27
2. Albergo, M.S., Vanden-Eijnden, E.: Building normalizing flows with stochastic interpolants. arXiv preprint arXiv:2209.15571 (2022)
3. Anagnostidis, S., Bachmann, G., Kim, Y., Kohler, J., Georgopoulos, M., Sanakoyeu, A., Du, Y., Pumarola, A., Thabet, A., Schönfeld, E.: Flexidit: Your diffusion transformer can easily generate high-quality samples with less compute. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 28316–28326 (2025)
4. Chadebec, C., Tasar, O., Benaroche, E., Aubin, B.: Flash diffusion: Accelerating any conditional diffusion model for few steps image generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 15686–15695 (2025)
5. Chang, J., Fang, Y., Xing, P., Wu, S., Cheng, W., Wang, R., Zeng, X., Yu, G., Chen, H.B.: Oneig-bench: Omni-dimensional nuanced evaluation for image generation. arXiv preprint arXiv:2506.07977 (2025)
6. Chen, H., Zhang, K., Tan, H., Guibas, L., Wetzstein, G., Bi, S.: pi-flow: Policy-based few-step generation via imitation distillation. arXiv preprint arXiv:2510.14974 (2025)
7. Chen, J., Ge, C., Xie, E., Wu, Y., Yao, L., Ren, X., Wang, Z., Luo, P., Lu, H., Li, Z.: Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. arXiv preprint arXiv:2403.04692 (2024)
8. Chen, J., Wu, Y., Luo, S., Xie, E., Paul, S., Luo, P., Zhao, H., Li, Z.: Pixart-δ: Fast and controllable image generation with latent consistency models. arXiv preprint arXiv:2401.05252 (2024)

9. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., et al.: Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In: International Conference on Learning Representations (2024)
10. Cho, M., Ohana, R., Jacobsen, C., Jothi, A., Chen, M.H., Mao, Z.M., Can, E.: Tc-lora: Temporally modulated conditional lora for adaptive diffusion control. arXiv preprint arXiv:2510.09561 (2025)
11. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
12. DiffSynth-Studio: Qwen-image-self-generated-dataset. `https://www.modelscope.cn/datasets/DiffSynth-Studio/Qwen-Image-Self-Generated-Dataset` (2025), accessed: 2025-09-24
13. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. arXiv preprint arXiv:2403.03206 (2024)
14. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: International Conference on Machine Learning (2024)
15. Frans, K., Hafner, D., Levine, S., Abbeel, P.: One step diffusion via shortcut models. arXiv preprint arXiv:2410.12557 (2024)
16. Geng, Z., Deng, M., Bai, X., Kolter, J.Z., He, K.: Mean flows for one-step generative modeling. arXiv preprint arXiv:2505.13447 (2025)
17. Geng, Z., Wang, Y., Ma, Y., Li, C., Rao, Y., Gu, S., Zhong, Z., Lu, Q., Hu, H., Zhang, X., Linus, Wang, D., Jiang, J.: X-omni: Reinforcement learning makes discrete autoregressive image generative models great again (2025), `https://arxiv.org/abs/2507.22058`, accessed: 2025-09-29
18. Ghosh, D., Hajishirzi, H., Schmidt, L.: Geneval: An object-focused framework for evaluating text-to-image alignment. Advances in Neural Information Processing Systems **36**, 52132–52152 (2023)
19. Gu, Y., Wang, X., Wu, J.Z., Shi, Y., Chen, Y., Fan, Z., Xiao, W., Zhao, R., Chang, S., Wu, W., et al.: Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. Advances in Neural Information Processing Systems **36** (2024)
20. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NeurIPS (2017)
21. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022)
22. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
23. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
24. Hu, X., Wang, R., Fang, Y., Fu, B., Cheng, P., Yu, G.: Ella: Equip diffusion models with llm for enhanced semantic alignment. arXiv preprint arXiv:2403.05135 (2024)
25. Huang, K., Sun, K., Xie, E., Li, Z., Liu, X.: T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. Advances in Neural Information Processing Systems **36**, 78723–78747 (2023)
26. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models **35**, 26565–26577 (2022)

27. Kim, D., Lai, C.H., Liao, W.H., Murata, N., Takida, Y., Uesaka, T., He, Y., Mitsufuji, Y., Ermon, S.: Consistency trajectory models: Learning probability flow ode trajectory of diffusion. arXiv preprint arXiv:2310.02279 (2023)
28. Labs, B.F.: Flux. `https://github.com/black-forest-labs/flux` (2024), accessed: 2025-09-18
29. Lin, H., Cho, J., Zala, A., Bansal, M.: Ctrl-adapter: An efficient and versatile framework for adapting diverse controls to any diffusion model. arXiv preprint arXiv:2404.09967 (2024)
30. Lin, S., Wang, A., Yang, X.: Sdxl-lightning: Progressive adversarial diffusion distillation (2024), `https://arxiv.org/abs/2402.13929`, accessed: 2025-09-25
31. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
32. Lin, Z., Pathak, D., Li, B., Li, J., Xia, X., Neubig, G., Zhang, P., Ramanan, D.: Evaluating text-to-visual generation with image-to-text generation. In: ECCV (2024)
33. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
34. Liu, L., Ren, Y., Lin, Z., Zhao, Z.: Pseudo numerical methods for diffusion models on manifolds. arXiv preprint arXiv:2202.09778 (2022)
35. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022)
36. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. Advances in Neural Information Processing Systems **35**, 5775–5787 (2022)
37. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. arXiv preprint arXiv:2211.01095 (2022)
38. Luhman, E., Luhman, T.: Knowledge distillation in iterative generative models for improved sampling speed. arXiv preprint arXiv:2101.02388 (2021)
39. Lyu, M., Yang, Y., Hong, H., Chen, H., Jin, X., He, Y., Xue, H., Han, J., Ding, G.: One-dimensional adapter to rule them all: Concepts diffusion models and erasing applications. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7559–7568 (2024)
40. ModelTC: Qwen-Image-Lightning: Speed up qwen-image model with distillation. `https://github.com/ModelTC/Qwen-Image-Lightning` (2025), accessed: 2025-11-07
41. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 4296–4304 (2024)
42. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 4296–4304 (2024)
43. Peng, B., Wang, J., Zhang, Y., Li, W., Yang, M.C., Jia, J.: Controlnext: Powerful and efficient control for image and video generation. arXiv preprint arXiv:2408.06070 (2024)
44. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image

synthesis. In: The Twelfth International Conference on Learning Representations (2023)
45. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763 (2021)
46. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)
47. Ran, L., Cun, X., Liu, J.W., Zhao, R., Zijie, S., Wang, X., Keppo, J., Shou, M.Z.: X-adapter: Adding universal compatibility of plugins for upgraded diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8775–8784 (2024)
48. Ren, Y., Xia, X., Lu, Y., Zhang, J., Wu, J., Xie, P., WANG, X., Xiao, X.: Hyper-SD: Trajectory segmented consistency model for efficient image synthesis. In: NeurIPS (2024), https://openreview.net/forum?id=O5XbOoi0x3, accessed: 2025-10-21
49. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
50. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. arXiv preprint arXiv:2202.00512 (2022)
51. Sauer, A., Lorenz, D., Blattmann, A., Rombach, R.: Adversarial diffusion distillation. In: European Conference on Computer Vision. pp. 87–103. Springer (2024)
52. Sauer, A., Lorenz, D., Blattmann, A., Rombach, R.: Adversarial diffusion distillation. In: ECCV. pp. 87–103 (2024)
53. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. pmlr (2015)
54. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models. arXiv preprint arXiv:2303.01469 (2023)
55. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. Advances in neural information processing systems **32** (2019)
56. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: International Conference on Learning Representations (2020)
57. Team, M.: Diffsynth-studio: Enjoy the magic of diffusion models! https://github.com/modelscope/DiffSynth-Studio (2025), gitHub repository, accessed: 2025-09-13
58. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., ming Yin, S., Bai, S., Xu, X., Chen, Y., Chen, Y., Tang, Z., Zhang, Z., Wang, Z., Yang, A., Yu, B., Cheng, C., Liu, D., Li, D., Zhang, H., Meng, H., Wei, H., Ni, J., Chen, K., Cao, K., Peng, L., Qu, L., Wu, M., Wang, P., Yu, S., Wen, T., Feng, W., Xu, X., Wang, Y., Zhang, Y., Zhu, Y., Wu, Y., Cai, Y., Liu, Z.: Qwen-image technical report (2025), https://arxiv.org/abs/2508.02324, accessed: 2025-09-15
59. Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., Li, H.: Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. arXiv preprint arXiv:2306.09341 (2023)
60. Xie, E., Yao, L., Shi, H., Liu, Z., Zhou, D., Liu, Z., Li, J., Li, Z.: Difffit: Unlocking transferability of large diffusion models via simple parameter-efficient fine-tuning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4230–4239 (2023)

61. Xing, Z., Dai, Q., Hu, H., Wu, Z., Jiang, Y.G.: Simda: Simple diffusion adapter for efficient video generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7827–7839 (2024)
62. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)
63. Yin, T., Gharbi, M., Park, T., Zhang, R., Shechtman, E., Durand, F., Freeman, B.: Improved distribution matching distillation for fast image synthesis. Advances in neural information processing systems **37**, 47455–47487 (2024)
64. Yin, T., Gharbi, M., Zhang, R., Shechtman, E., Durand, F., Freeman, W.T., Park, T.: One-step diffusion with distribution matching distillation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6613–6623 (2024)
65. Yin, T., Gharbi, M., Zhang, R., Shechtman, E., Durand, F., Freeman, W.T., Park, T.: One-step diffusion with distribution matching distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6613–6623 (2024)
66. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)
67. Zhou, M., Zheng, H., Wang, Z., Yin, M., Huang, H.: Score identity distillation: Exponentially fast distillation of pretrained diffusion models for one-step generation. In: Forty-first International Conference on Machine Learning (2024)
68. Zhuang, S., Guo, Y., Ding, Y., Li, K., Chen, X., Wang, Y., Wang, F., Zhang, Y., Li, C., Wang, Y.: Timestep master: Asymmetrical mixture of timestep lora experts for versatile and efficient diffusion models in vision. arXiv preprint arXiv:2503.07416 (2025)