CollectionLoRA: Collecting 50 Effects in 1 LoRA via Multi-Teacher On-Policy Distillation

Fangtai Wu[1,2], Hailong Guo[2], Shijie Huang[2], Jiayi Song[2,3], Yubo Huang[2], Mushui Liu[1], Zhao Wang[1], Yunlong Yu[1*], Jiaming Liu[2*†], and Ruihua Huang[2]

[1] Zhejiang University, China
[2] Qwen Applications Business Group of Alibaba, China
[3] Xi'an Jiaotong University, China
wft@zju.edu.cn, yuyunlong@zju.edu.cn, liujiaming.ljl@alibaba-inc.com
[*] Corresponding authors. [†] Project lead.

Fig. 1: We propose CollectionLoRA, a multi-teacher on-policy distillation framework capable of consolidating multiple diverse visual effects and few-step inference capabilities into a single LoRA module.

Abstract. Customized image editing equips diffusion models with specific visual transformations using limited paired data, typically via Low-Rank Adaptation (LoRA). However, combining acceleration and effect LoRAs during inference often introduces interference, causing concept bleeding, semantic drift, and degraded stylistic fidelity. In this paper, we propose a unified paradigm that distills numerous customized concepts and fast-inference capabilities into a single LoRA, effectively resolving compositional conflicts while reducing deployment overhead. To address the challenges of multi-concept distillation, particularly gradient vanishing and training instability, we introduce CollectionLoRA, a multi-teacher hybrid distillation framework. Our framework incorporates: (1) Probabilistic Dual-Stream Routing (PDSR) that leverages general-domain data as regularization to enhance robustness. (2) Automated Asymmetric Conditioning (AAC) utilizing Vision-Language Models (VLMs) and

orthogonal trigger words to ensure concept isolation, and (3) the Decoupled Hybrid Objective (DHO) that synergizes Trajectory-Anchored Flow Matching (TA-FM) priors with target simulation and distribution matching, ensuring robust optimization across heterogeneous multi-task distributions. Extensive experiments demonstrate that our unified architecture mitigates multi-LoRA interference and achieves enhanced concept fidelity, effective feature isolation, and high-quality rapid synthesis. Code is available at `https://github.com/Qwen-Applications/CollectionLoRA`.

Keywords: Multi-Teacher Distillation · Multi-Concept Personalization · Few-Step Generation

1 Introduction

Diffusion models [8, 19–21, 24, 27, 29, 38, 40, 42] have recently achieved remarkable success in customized image generation [11, 14, 25, 30, 31, 35, 41, 46, 51, 52]. Leveraging powerful pretrained generative priors, these models can be adapted to perform specific visual transformations (e.g., artistic stylization or domain-specific editing) using only a small number of paired training samples. To enable parameter-efficient adaptation, existing approaches [16, 34] commonly employ Low-Rank Adaptation (LoRA) [15], training independent LoRA modules to capture diverse visual effects.

However, current customized generation pipelines predominantly adopt a modular paradigm, in which multiple effect-specific LoRAs are linearly aggregated with the base diffusion model during inference. In practice, the LoRA corresponding to the desired effect is selected according to the user prompt and combined with the base model, while an additional acceleration LoRA is often incorporated to enable few-step sampling, as illustrated in Fig. 2(a). Although this modular design provides flexibility for extending new visual transformations, it introduces fundamental scalability limitations as the number of effects grows. First, maintaining a large collection of effect-LoRAs leads to prohibitive storage overhead, making deployment on resource-constrained inference devices increasingly difficult. Second, the routing process that maps user instructions to specific LoRAs becomes progressively slower and less reliable as the repository expands, resulting in increased latency and potential matching errors. Third, and more critically, the linear aggregation of multiple LoRAs often disrupts the original feature manifold of the base model. The unintended interactions between effect modules and acceleration modules frequently produce severe interference, manifested as concept bleeding, semantic drift, and degraded stylistic fidelity.

A natural solution to these limitations is to distill multiple customized concepts together with fast-inference capabilities into a single unified LoRA. Such a paradigm removes the need for runtime LoRA composition, thereby reducing storage overhead and eliminating routing complexity. More importantly, it resolves compositional conflicts at the parameter level, enabling stable generation without multi-module interference. However, achieving this goal remains

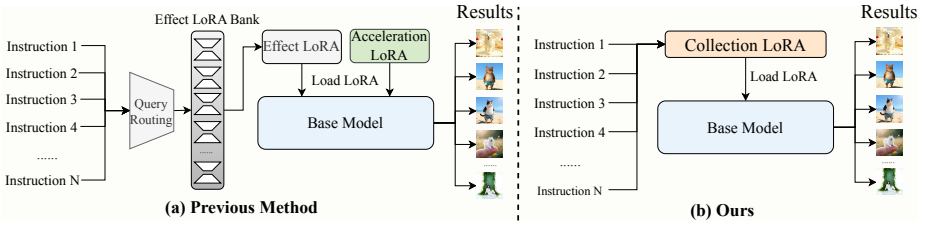


Fig. 2: Comparison between conventional multi-LoRA pipelines and the proposed CollectionLoRA. **Conventional**: Training and deploying task-specific LoRA weights for each concept, which are sequentially composed with an acceleration LoRA during inference. **CollectionLoRA**: Consolidating the acceleration prior and all target concepts into a single unified module through multi-teacher hybrid distillation during training.

challenging. When distilling knowledge from multiple heterogeneous teachers, the training process involves highly non-stationary multi-task distributions. In such settings, traditional Distribution Matching Distillation (DMD) [47, 48] often suffers from gradient vanishing during the early training stage, particularly when the student distribution deviates significantly from that of the teachers. To address these challenges, we propose CollectionLoRA, a multi-teacher hybrid distillation framework that integrates both customized visual transformations and few-step generation capabilities into a single lightweight LoRA. By distilling knowledge from multiple effect-specific teachers together with acceleration priors, CollectionLoRA reduces deployment costs while avoiding the interference caused by sequential LoRA composition, as shown in Fig. 2(b).

Our framework consists of three key components. First, we introduce a Probabilistic Dual-Stream Routing (PDSR) mechanism that jointly optimizes concept-specific supervision and general-domain regularization to alleviate training instability under non-stationary multi-task distributions. By probabilistically mixing customized and general-domain data during training, PDSR stabilizes gradient signals and preserves the base model's generalization ability. Second, we propose an Automated Asymmetric Conditioning (AAC) strategy to mitigate the severe cross-concept interference arising from heterogeneous teacher models. AAC leverages vision–language models to automatically construct asymmetric prompts and assigns orthogonal trigger tokens to different concepts, enforcing semantic separation in the latent space and preventing concept entanglement during distillation. Third, to address the gradient vanishing issue commonly observed in early-stage distribution matching, we design Decoupled Hybrid Objective (DHO) that integrates a Trajectory-Anchored Flow Matching (TA-FM) prior with a Target Simulation (TS) mechanism under dual-timestep constraints. This hybrid objective provides stable trajectory-level supervision while maintaining distribution alignment with multiple teachers, enabling robust optimization under heterogeneous multi-task distributions.

In summary, the main contributions are threefold:

- We propose CollectionLoRA, a unified framework that distills multiple customized visual transformations together with few-step generation capabilities into a single lightweight LoRA.
- To address the optimization challenges arising from heterogeneous multi-teacher supervision and non-stationary multi-task distributions, we integrate **Probabilistic Dual-Stream Routing (PDSR)** and **Automated Asymmetric Conditioning (AAC)** to preserve model generalization and enforce concept-level disentanglement. Furthermore, we introduce a **Decoupled Hybrid Objective (DHO)** that combines **Trajectory-Anchored Flow Matching (TA-FM)** with **Target Simulation (TS)** to stabilize distribution matching and mitigate gradient vanishing during training.
- Extensive experiments demonstrate that CollectionLoRA distills dozens of heterogeneous visual effects together with fast inference capabilities into a single LoRA module. The resulting model eliminates the storage and routing overhead associated with multi-LoRA pipelines while achieving strong concept fidelity, effective feature isolation, and high-quality few-step generation.

2 Related Work

2.1 Customized Image Generation

Customized image generation has emerged as a pivotal task within the broader landscape of image synthesis [30, 45, 49], focusing on enabling pretrained diffusion models to understand specific concepts from limited data and re-render them in diverse contexts. Early optimization-based methods, such as Textual Inversion [11] and DreamBooth [30], paved the way by learning specific tokens or fine-tuning the model for a single subject. Methods like ELITE [39], IP-Adapter [46], InstantID [37], and MoMA [32] treat personalization as a vision-conditioned generation task by training specialized adapters. With the emergence of Diffusion Transformers (DiT) [27] like FLUX [19] and SD3 [8], the paradigm has further evolved toward leveraging strong in-context capabilities. For instance, OmniControl [35] and EasyControl [52] adapt text-to-image models for precise personalized control, while unified models like Bagel [7] attempt to harmonize understanding and generation. Furthermore, large-scale models such as FLUX Kontext [21], Qwen-Image-Edit [40], and FLUX2 [20] also leverage in-context learning for image generation and editing. However, zero-shot adapters often fail on out-of-distribution effects, making custom LoRA training the reliable industrial standard. Yet, directly composing these LoRAs with acceleration modules (e.g., lightx2v [6]) triggers severe feature interference and semantic drift. To resolve this, we distill multiple customized effects into a single, few-step unified LoRA, completely avoiding the conflicts inherent in multi-module composition.

2.2 Few-Step Generation

To address the inference inefficiency of diffusion models, Consistency Models (CMs) [33] and their derivatives [26, 36, 50] enable few-step generation by enforcing trajectory self-consistency. Recently, Distribution Matching Distillation

(DMD) [47, 48] established a superior paradigm by directly minimizing the divergence between the teacher and student distributions. Recent advances further elevate DMD: Decoupled-DMD [23] enhances fine details via independent noise schedules, while DMDR [18] and Flash-DMD [3] integrate reinforcement learning to incorporate external preference rewards safely, surpassing the teacher's performance ceiling. Despite these advancements, existing DMD-based methods are largely confined to single-teacher, homogeneous settings. They suffer from severe training instability and feature collapse when bridging large student-teacher gaps or simultaneously fitting multiple target distributions. To overcome these bottlenecks, we propose CollectionLoRA, a multi-teacher distillation framework designed to stabilize multi-source matching and prevent distribution collapse.

2.3 On-Policy Distillation

Standard offline distillation often suffers from exposure bias and compounding errors [1]. OPD [1, 13, 22] mitigates these issues by applying teacher feedback directly to states visited by the student's own rollouts. While initially validated in Large Language Models for superior stability over scalar-reward reinforcement learning [44], OPD has recently been adapted to continuous visual generation. In diffusion and flow matching (e.g., Flow-OPD [9], D-OPSD [17]), OPD matches the teacher's dense velocity fields along student-sampled trajectories. Technically, while traditional OPD typically aligns the conditional transition distribution step-wise along trajectories, DMD [47] focuses on optimizing the marginal data distribution of generated samples. Following recent literature [4, 12], we conceptually unify DMD under the OPD taxonomy, as both frameworks fundamentally rely on correcting the student's on-policy explored states with teacher signals.

Building upon this paradigm, CollectionLoRA pioneers large-scale multi-teacher distillation to efficiently consolidate 50 to 180 diverse visual effects alongside few-step generation capabilities into a single, unified module.

3 Preliminaries: Distribution Matching Distillation

Distribution Matching Distillation (DMD) aims to train an efficient student generator G_θ such that its generated distribution p_{fake} approximates the target real distribution p_{real} defined by a pre-trained diffusion teacher. To ensure consistency between training and inference in few-step synthesis, DMD2 [47] employs **Backward Simulation** to simulate the inference process. Specifically, starting from pure noise $z \sim \mathcal{N}(0, \mathbf{I})$, the simulation iteratively performs a sequence of denoising and re-noising steps: it predicts a denoised sample $\hat{x}_0$, then adds noise back to reach the next scheduled timestep. This iterative loop continues until a selected timestep t is reached, yielding a simulated clean image that effectively captures the cumulative sampling characteristics of the multi-step inference trajectory.

This simulated image then serves as the training target, replacing the conventional real data. The generator G_θ is trained to denoise a noisy version of this simulated sample to get the generated sample x_g, with the objective of matching the score functions of the real and fake distributions. The gradient for updating the generator parameters θ is formulated as:

$$\nabla_\theta \mathcal{L}_{DMD} = \mathbb{E}_{z,t,\epsilon}\left[\left(s_{fake}(x_t, t) - s_{real}(x_t, t)\right) \nabla_\theta x_g\right], \tag{1}$$

where x_t is the diffused state of the generated sample x_g. The target score s_{real} is derived from the frozen teacher, while the fake score s_{fake} is estimated by a critic model updated via standard denoising loss. Additionally, DMD2 [47] incorporates a time-conditioned GAN loss to compensate for potential estimation errors in the teacher model. However, adversarial training is inherently unstable, and applying this loss under the limited data constraints of the few-shot personalization setting severely exacerbates mode collapse and overfitting in multi-distribution scenarios. Consequently, this GAN-based objective is excluded from our design. Beyond the challenges associated with adversarial training, directly adapting the distribution matching framework to non-stationary, customized concepts remains susceptible to vanishing gradients, which we address in the subsequent sections.

4 Method

To jointly distill 50 customized effects and few-step generation capabilities into a single lightweight student LoRA, we propose CollectionLoRA, a multi-teacher hybrid distillation framework. This section introduces our Probabilistic Dual-Stream Routing (PDSR) mechanism, the Automated Asymmetric Conditioning (AAC) strategy, and the Decoupled Hybrid Objective (DHO).

4.1 Probabilistic Dual-Stream Routing

To prevent catastrophic forgetting during multi-concept distillation, we propose the PDSR mechanism. At each training step, the framework dynamically routes the optimization process between two parallel logical streams based on a predefined LoRA switching rate, denoted as p_{switch}. Specifically, we sample a random probability $p \sim \mathcal{U}(0, 1)$. If $p < p_{\text{switch}}$, the process is routed to the effect stream. Otherwise, it defaults to the general stream. This threshold-based routing assigns customized optimization objectives to each branch:

- **General Knowledge Stream ($p \geq p_{\text{switch}}$).** Utilizing 20K unannotated images and a frozen base model as the teacher, this stream applies standard target-free distribution matching. This preserves foundational generation priors and prevents the student from overfitting to specific effects.
- **Effect Concept Stream ($p < p_{\text{switch}}$).** Utilizing a lightweight paired dataset, this stream dynamically mounts a specific effect LoRA onto the base model to instantiate the teacher. To avoid optimization collapse during the cold-start phase, we apply a hybrid mechanism combining trajectory anchoring and target-driven distribution matching.

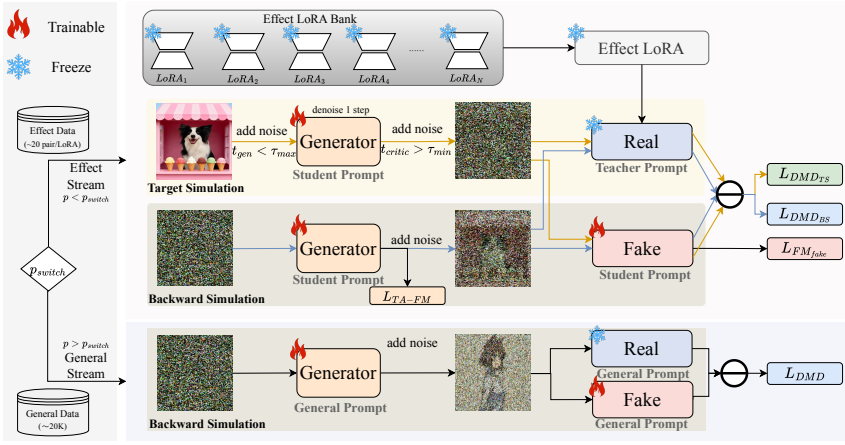


Fig. 3: Overview of the CollectionLoRA training architecture. The framework dynamically routes the optimization process between two parallel streams based on the switching probability p_{switch}. The general stream relies on backward simulation with general data to preserve foundational generation priors via the distribution matching loss $\mathcal{L}_{DMD}$. The effect stream utilizes an effect LoRA bank to instantiate dynamic frozen teachers and combines target simulation bounded by specific timesteps $t_{gen} < \tau_{max}$ and $t_{critic} > \tau_{min}$ with backward simulation. By respectively using teacher prompts and student prompts, the trainable generator is optimized through a combination of distribution matching losses $\mathcal{L}_{DMD_{TS}}$ and $\mathcal{L}_{DMD_{BS}}$ along with the trajectory anchoring loss $\mathcal{L}_{TA-FM}$ while the trainable fake score model is updated via $\mathcal{L}_{FM_{fake}}$.

4.2 Automated Asymmetric Conditioning

To mitigate feature interference and concept leakage during multi-effect integration, we propose the Automated Asymmetric Conditioning (AAC) strategy. Unlike standard distillation, AAC uses different prompts for the teacher and student models to ensure clear concept isolation. Specifically, each effect teacher T^i_{effect} uses its original training prompt c^i_{teacher} to generate high-quality target images y. To avoid semantic confusion in the student model, we use a Vision-Language Model (VLM) to automatically generate a descriptive caption c^i_{vlm} for each effect. This automated process removes the need for manual prompt engineering. We then assign a unique orthogonal trigger word v^i to each effect and construct the student condition as:

$$c^i_{\text{student}} = [v^i, c^i_{\text{vlm}}]. \tag{2}$$

4.3 Decoupled Hybrid Objective

We propose the DHO to mitigate forgetting and stabilize the cold-start phase of effect distillation. Specifically, when training is routed to the general stream, the

student model leverages its pre-trained prior via a standard Backward Simulation (BS) mechanism, as shown in Fig. 3. In this setup, the generator samples an image x_g from Gaussian noise, which is subsequently evaluated by the teacher and critic model to get score s_{real} and s_{fake} to compute the distribution matching loss $\mathcal{L}_{\text{DMD\_BS}}$ as in the Eq. 1. Conversely, when routed to the effect stream, the student is tasked with learning new visual concepts without prior initialization. Standard distribution matching is ineffective in this scenario, as the student-teacher distribution shift during the cold-start phase leads to vanishing gradients. To stabilize optimization and ensure perceptual quality, we design a hybrid objective. **(1) Trajectory Anchoring via Flow Matching.** To bridge the initial distribution gap, we construct a flow matching objective. Specifically, we utilize target images y pre-generated by the effect teacher and regularize the student to follow the flow trajectory towards y:

$$\mathcal{L}_{TA-FM} = \|G_\theta(y_t, t, c_{\text{student}}) - (y - \epsilon)\|_2^2, \tag{3}$$

where $y_t = ty + (1-t)\epsilon$ denotes the noisy intermediate state driven by Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$, and $(y - \epsilon)$ represents the target vector field. $\mathcal{L}_{TA-FM}$ forces the student to align with this target trajectory, preventing gradient divergence during the cold-start phase. While $\mathcal{L}_{TA-FM}$ stabilizes early training, relying solely on this regression loss introduces an over-smoothing bias that degrades high-frequency textures. To address this, we introduce a distribution matching loss based on target simulation, guided by target images y. **(2) Distribution Matching via Target Simulation (TS).** In this branch, the target image y is first diffused to timestep t_{gen} and subsequently denoised in a single step via the student generator G_θ to produce the simulated output $\hat{y}$. $\hat{y}$ is then further corrupted with noise to a critic timestep t_{critic} (denoted as $\hat{y}_{t_{critic}}$). We compute the gradient update $\nabla_\theta \mathcal{L}_{\text{DMD\_TS}}$ by substituting the variables into Eq. 1, where x_g, x_t, and t are replaced by $\hat{y}$, $\hat{y}_{t_{\text{critic}}}$, and t_{critic}, respectively. To ensure effective teacher guidance and mitigate the instability of heterogeneous distillation, we impose dual-timestep constraints:

- **Generator Upper-Bound ($t_{gen} < \tau_{max}$).** This bound restricts the forward diffusion to preserve the teacher's prior, preventing the denoising process from degenerating into unguided generation.
- **Critic Lower-Bound ($t_{critic} > \tau_{min}$).** This bound ensures sufficient noise injection during evaluation, which amplifies the divergence between the real and fake score predictions to provide reliable gradient guidance from the teacher.

4.4 Overall Objective

To balance initial trajectory alignment with late-stage distribution matching, we implement a step-dependent dynamic weighting schedule. The overall optimization objective is:

$$\begin{aligned}\mathcal{L}_{total} = & \mathbb{1}_{\{general\}} \mathcal{L}_{DMD\_BS} \\ & + \mathbb{1}_{\{effect\}} \left(\lambda_1(s)\mathcal{L}_{TA-FM} + \lambda_2(s)\mathcal{L}_{DMD\_TS} + \mathcal{L}_{DMD\_BS}\right).\end{aligned} \tag{4}$$

where the indicator functions $\mathbb{1}_{\{general\}}$ and $\mathbb{1}_{\{effect\}}$ denote the stream selected for the current training step and $\lambda_1(s)$ decreases over time to reduce the influence of the initial trajectory alignment, while $\lambda_2(s)$ increases to focus the model on refining fine-grained visual details in later stages.

5 Experiment

5.1 Experimental Setup

Training Datasets. Our dual-stream framework utilizes two distinct datasets. For the effect stream, we curate a dataset of 50 specific effects covering complex tasks like style transfer, scene replacement, and subject modification. Each effect includes approximately 20 image pairs, categorized into animal and portrait domains. For the general stream, we collect 20K source images and use an MLLM to generate diverse editing instructions (e.g., add, remove, adjust, style). This dataset is used solely for the base Distribution Matching Distillation (DMD) training, thus requiring no target output images.

Evaluation Dataset. To evaluate effect-specific performance, we introduce EffectBench. Aligned with our training data, it comprises animal and portrait categories. We use Gemini-2.5 Pro and Qwen-Image [40] to generate 100 diverse test images per category, ensuring high variance in subject types, actions, scenes, and camera distances. This yields an evaluation protocol of 5,000 instructions per model.

Baseline Methods. We adopt Qwen-Image-Edit-2509 [40] as the base model and compare our approach against two standard paradigms: (1) Base Model + Effect LoRA, and (2) Base Model + Effect LoRA + Acceleration LoRA. For acceleration, we utilize the popular Qwen-Image-Edit-Lightning LoRA released by lightx2v [6]. To evaluate multi-concept injection capabilities, we also construct a strong baseline denoted as 50-in-1 (FM), which optimizes a unified LoRA on all aggregated training data using a standard flow matching objective.

Fig. 4: Evaluation of subject consistency metrics. While DINO often assigns high scores to failed generations based on superficial similarity, VSA strictly penalizes semantic failures with a zero score, providing a robust measure of actual subject consistency in complex stylizations.

Table 1: Quantitative comparison on our EffectBench.

Setting	Method	CLIP (↑)	DreamSim (↓)	DINO (↑)	VSA (↑)	EditReward (↑)	BCR (↓)	NFE (↓)
Single Effect	Base	0.726	0.434	0.611	4.075	1.007	0.141	40 × 2
	Base+Lightning	0.717	0.441	**0.612**	3.901	0.986	0.168	8
50 Effects in 1	FM + Lightning	0.703	0.468	0.611	4.150	0.929	0.217	8
	Ours	**0.727**	**0.425**	0.600	**4.380**	**1.052**	**0.087**	**8**

Table 2: Deployment costs across numbers of LoRAs.

Metric	Method	10 LoRAs	20 LoRAs	50 LoRAs	100 LoRAs	150 LoRAs
Routing Latency	baseline	6.88s/q	6.95 s/q	7.09s/q	7.22s/q	9.18s/q
	ours	**0s/q**	**0s/q**	**0s/q**	**7.22s/q**	**9.18s/q**
LoRA Loading Latency × Switch Count	baseline	1.2s*200	1.2s*200	1.2s*200	1.2s*200	1.2s*200
	ours	**0s**	**0s**	**0s**	**1.2s*108**	**1.2s*136**
Routing Accuracy	baseline	99%	94%	87%	85%	76%
	ours	**100%**	**100%**	**100%**	**90%**	**82%**
Storage Overhead	baseline	2.2G * 10	2.2G * 20	2.2G * 50	2.2G * 100	2.2G * 150
	ours	**2.2G**	**2.2G**	**2.2G**	**2.2G * 2**	**2.2G * 3**

Evaluation Metrics. We assess generation quality using several metrics: CLIP [28] and DreamSim [10] for style alignment, DINO [2] for subject consistency, and EditReward [43] for instruction-following and overall image quality. We also use an MLLM to calculate the Bad Case Rate (BCR), representing the proportion of effect failures across all test images. Because traditional metrics like DINO often fail to capture subject consistency in complex effect scenarios (Fig. 4), we introduce the Valid Subject Alignment (VSA) metric. VSA uses the MLLM for a two-stage evaluation: it first verifies if the target effect was successfully applied. If it fails, the consistency score defaults to zero and if successful, the MLLM evaluates the underlying subject consistency. Further details are provided in the supplementary material.

Implementation Details. All effect LoRA training is performed at a 1024x1024 resolution using the ai-toolkit [5] with a learning rate of 1e-4. Single-effect LoRAs are trained for 2,000 steps, while the 50-in-1 (FM) baseline is trained for 30,000 steps. In our method, both the generator and the fake score model are fine-tuned exclusively via LoRA with a learning rate of 1e-4. The update frequency ratio of the fake score model to the generator is 5:1, and the generator is trained for 5,000 steps. By default, the stream switching probability p_{switch} is 0.5, with timestep bounds $\tau_{\max} = 750$ and $\tau_{\min} = 500$. All experiments are conducted on 8 NVIDIA H800 GPUs.

5.2 Quantitative Evaluation

Quantitative Comparison with Baselines. Table 1 compares our joint distillation approach (50 Effects in 1, NFE=8) against independent single-effect models (NFE=80) and a naive multi-task baseline (FM + Lightning). Our method achieves state-of-the-art style alignment and overall quality (CLIP: 0.727, DreamSim: 0.425, EditReward: 1.052), remarkably outperforming the computationally expensive single-task teacher and mitigating the degradation typical of multi-concept fusion. Benefiting from our AAC and DHO strategies, we effectively suppress parameter interference, yielding a Bad Case Rate (BCR) of just 0.087—substantially lower than both the baseline (0.217) and the teacher (0.141). Regarding subject consistency, although a marginal decrease is observed in the traditional DINO metric (0.600 vs. 0.611)—which often struggles with severe visual transformations as shown in Fig. 4—our method achieves the highest Valid Subject Alignment (VSA) score (4.380). This demonstrates that our unified model robustly triggers effects, minimizes failure penalties, and better preserves foundational structures under extreme concept compression.

Deployment Cost Analysis. To evaluate deployment overhead, we simulated 200 queries on a single GPU against a VLM-routed baseline, as shown in Tab. 2. For 10-50 LoRAs, CollectionLoRA bypasses routing entirely, achieving 0s latency, 100% accuracy, and a constant 2.2GB storage overhead, circumventing the baseline's linear storage growth and accuracy degradation. At 100-150 LoRAs, our approach reverts to VLM routing but still drastically reduces storage to 2% of the baseline, decreases model switches (136 vs. 200), and maintains higher accuracy (82% vs. 76%). Consequently, its minimal memory footprint and switching burdens highlight significant cost-saving potential for large-scale multi-GPU deployments.

5.3 Qualitative Evaluation

Fig. 5 compares qualitative results. In the **Texture & Detail Loss panel**, the accelerated baseline (Base+Lightning) exhibits semantic drift (e.g., identity shifts, col. 1) and structural degradation (e.g., detail loss, col. 3). Furthermore, the joint distillation baseline (FM+Lightning) oversmooths due to regression losses, losing high-frequency textures (e.g., pet fur, cols. 1-2) and realism. In contrast, our constrained Target Simulation mechanism restores fine-grained textures and physical realism while preserving structural consistency. The **Style Interference panel** shows style degradation from module serialization and parameter crowding. For Base+Lightning, directly concatenating acceleration and effect LoRAs disrupts the feature manifold and degrades style purity. Second, without concept isolation, FM+Lightning outputs (cols. 5-6) show concept bleeding from the color palettes and brushstrokes of other effects. CollectionLoRA resolves this by isolating latent effects, synthesizing pure, crosstalk-free styles even under 50-effect, 8-step inference. The **Generalization Collapse panel** illustrates the generalization limits of few-shot fine-tuning. Due to minimal training data, baseline models yield distorted structures and poses for out-of-distribution

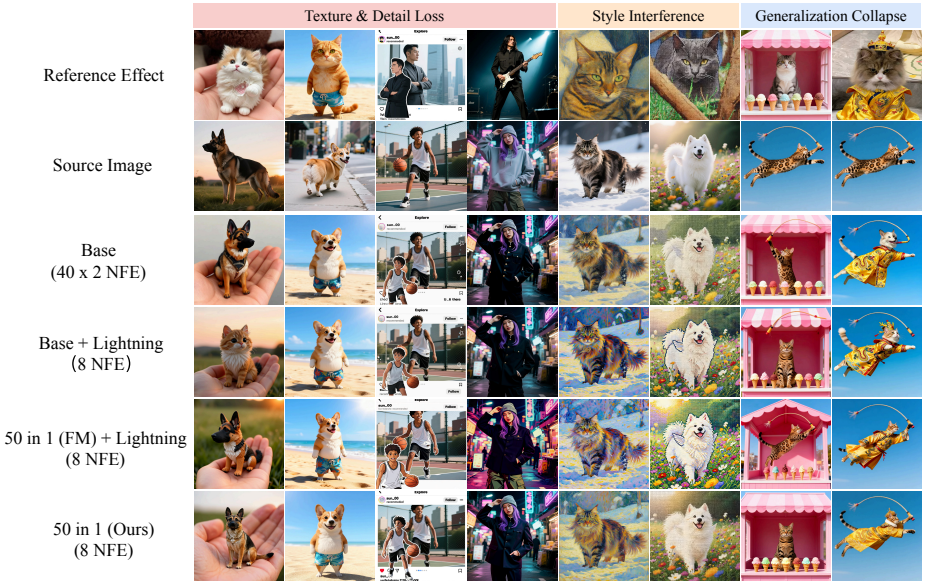


Fig. 5: Qualitative comparison of CollectionLoRA against baseline methods. The visual results indicate that CollectionLoRA effectively mitigates texture and detail loss, style interference, and generalization collapse commonly observed in the alternative pipelines.

(OOD) inputs (e.g., the flying cat). By using a dual-stream routing mechanism to introduce general-domain data as structural regularization, our method preserves the base model's generalization and structural fidelity for rare subjects.

5.4 Ablation Study

Quantitative Performance Analysis. We evaluate each component under a 50-in-1 concurrent distillation setting, as shown in Tab. 3. Comparing Exp. (1) and Exp. (2), AAC reduces the BCR from 0.378 to 0.207, demonstrating that orthogonal trigger words effectively mitigate concept bleeding. Exp. (3) achieves the highest CLIP (0.736) and DreamSim (0.420) scores, confirming that TS mitigates the oversmoothing bias of regression losses. In Exp. (5), TA-FM increases VSA from 4.018 to 4.380 and minimizes BCR to 0.087. This highlights TA-FM's role in stabilizing optimization under non-stationary distributions through microscopic trajectory guidance. Finally, removing PDSR in Exp. (4) leads to a substantial decline in EditReward, whereas its inclusion restores the score to 1.052 in Exp. (5), proving its role in maintaining generalization and preventing catastrophic forgetting.

Qualitative Assessment. Fig. 6 presents visual comparisons of our method. Exp. (1) lacks AAC and suffers from semantic collapse. AAC decouples the effect and general streams and this separation ensures accurate task triggering shown

Table 3: Ablation study of the proposed components under the 50-in-1 concurrent setting. ✓indicates the application of the specific module. The best results for each metric are highlighted in **bold**.

Exp.	PDSR	AAC	TS	TA-FM	CLIP ↑	DreamSim ↓	DINO ↑	VSA ↑	EditReward ↑	BCR ↓
(1)	✓				0.725	0.434	0.514	2.756	0.989	0.378
(2)	✓	✓			0.732	0.427	0.525	3.720	1.008	0.207
(3)	✓	✓	✓		**0.736**	**0.420**	0.541	4.018	0.979	0.199
(4)		✓	✓	✓	0.727	0.426	0.590	4.248	0.976	0.108
(5)	✓	✓	✓	✓	0.727	0.425	**0.600**	**4.380**	**1.052**	**0.087**

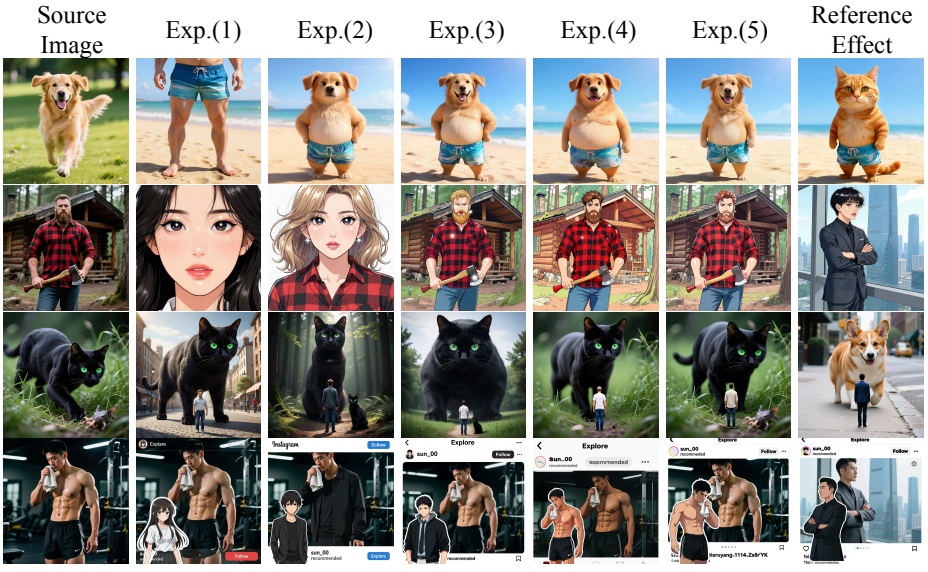


Fig. 6: Qualitative ablation study. The progressive integration of our core components systematically resolves semantic collapse, restores high-frequency textures, and ensures strict structural consistency.

in Exp. (2). TS restores high-frequency textures in Exp. (3). These details include skin and garment folds. Standard flow matching typically blurs such textures. TA-FM significantly improves structural consistency in Exp. (5). It provides stable microscopic anchoring. Consequently, the model preserves the spatial layout and avoids pose distortions. Finally, PDSR resolves the background blending issues in Exp. (4). This module achieves higher visual harmony between effects and complex environments.

LoRA Scaling & Upper Bound. Tab. 4 reports CLIP scores as the number of effects scales from 10 to 180. Our method consistently outperforms All-in-1 (FM) + Lightning across all settings, demonstrating that our hybrid distillation objective effectively mitigates the quality degradation caused by naive multi-concept fusion. At smaller scales (10–50 effects), our method even surpasses all base-

Fig. 7: Qualitative ablation of training dynamics. Integrating TA-FM and TS significantly accelerates structural convergence and enhances generative fidelity compared to the baseline.

Table 4: CLIP Score comparison across numbers of LoRAs

	10 LoRAs	**20 LoRAs**	**50 LoRAs**	**100 LoRAs**	**180 LoRAs**
Base	0.735	0.724	0.726	**0.723**	**0.724**
Base+Lightning	0.716	0.712	0.717	0.717	0.722
All in 1 (FM) + Lightning	0.725	0.722	0.703	0.694	0.689
All in 1 (Ours)	**0.741**	**0.723**	**0.727**	0.716	0.709

Table 5: CLIP score comparison for incremental effect addition

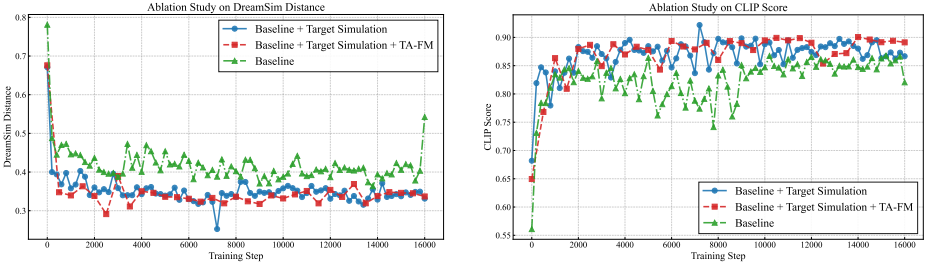
	50 LoRAs	51 LoRAs	52 LoRAs	53 LoRAs	54 LoRAs
Base+Lightning	0.717	0.720	0.721	0.724	0.724
Ours	**0.727**	**0.726**	**0.728**	**0.727**	**0.725**

lines including the single-task Base model. As the number of effects increases to 100–180, a moderate performance drop is observed, which is expected given the increased optimization difficulty under extreme concept compression. Nevertheless, our method maintains competitive performance against Base+Lightning, confirming that CollectionLoRA scales gracefully to a large number of effects without catastrophic quality degradation.

Incremental Effect Extension. Tab. 5 validates the incremental extension capability of CollectionLoRA. Starting from the 50-effect model, adding effects 51_{st}-54_{th} via lightweight fine-tuning (100 steps for generator) consistently outperforms Base+Lightning across all settings, with CLIP scores remaining stable in the range of 0.725-0.728. The absence of catastrophic forgetting confirms that CollectionLoRA supports effect extension without retraining from scratch.

Training Dynamics and Stability. As shown in Fig. 8, our full configuration offers significant stability advantages. Compared to the baseline, the model with TA-FM converges faster in the early phase, with DreamSim distance decreasing rapidly. While the baseline exhibits optimization oscillations after 8,000 steps

due to non-stationary multi-task distributions, TA-FM ensures a stable training trajectory. Qualitative backtracking shown in Fig. 7 confirms these results: the baseline remains underfitted at 12,000 steps, failing to recover core structures like the pink booth. In contrast, with TA-FM, our model captures the global structure within 200 steps and achieves high-fidelity synthesis by step 5,000, demonstrating enhanced training efficiency and perceptual quality.



(a) DreamSim Distance **(b)** CLIP Score

Fig. 8: Quantitative ablation of training dynamics. The integration of TS and TA-FM significantly accelerates convergence and stabilizes the optimization trajectory compared to the fluctuating baseline.

6 Conclusion

We propose CollectionLoRA, a unified multi-teacher distillation framework that integrates diverse customized visual effects and few-step inference into a single module, eliminating the storage overhead and concept interference (e.g., semantic drift) inherent in traditional multi-LoRA deployments. To address training instability in few-shot multi-concept distillation, our framework features three components: Probabilistic Dual-Stream Routing (PDSR) for structural regularization, Automated Asymmetric Conditioning (AAC) for latent concept isolation, and a Decoupled Hybrid Objective (DHO) to stabilize optimization and restore high-frequency details. Extensive experiments demonstrate that CollectionLoRA achieves superior concept fidelity, feature isolation, and high-quality synthesis surpassing single-task teachers, all while maintaining few-step generation capabilities.

Acknowledgment

This research was supported in part by the Key R&D Program of Zhejiang Province 2026C01016, the National Natural Science Foundation of China under Grant 62576313, CASIC Guangxin Intelligent Technology, and Key R&D Program of Ningbo under Grant 2025Z128.

References

1. Agarwal, R., Vieillard, N., Zhou, Y., Stanczyk, P., Ramos, S., Geist, M., Bachem, O.: On-policy distillation of language models: Learning from self-generated mistakes (2024), https://arxiv.org/abs/2306.13649
2. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers (2021), https://arxiv.org/abs/2104.14294
3. Chen, G., Huang, S., Liu, K., Zhu, J., Qu, X., Chen, P., Cheng, Y., Sun, Y.: Flash-dmd: Towards high-fidelity few-step image generation with efficient distillation and joint reinforcement learning. arXiv preprint arXiv:2511.20549 (2025)
4. Chern, E., Hu, Z., Tang, B., Su, J., Chern, S., Deng, Z., Liu, P.: Livetalk: Real-time multimodal interactive video diffusion via improved on-policy distillation (2025), https://arxiv.org/abs/2512.23576
5. aitookit Contributors: aitookit. https://github.com/ostris/ai-toolkit (2025)
6. Contributors, L.: Lightx2v: Light video generation inference framework. https://github.com/ModelTC/lightx2v (2025)
7. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., et al.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025)
8. 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: Forty-first international conference on machine learning (2024)
9. Fang, Z., Huang, W., Zeng, Y., Zhao, Y., Chen, S., Feng, K., Lin, Y., Chen, L., Chen, Z., Cao, S., Zhao, F.: Flow-opd: On-policy distillation for flow matching models (2026), https://arxiv.org/abs/2605.08063
10. Fu, S., Tamir, N., Sundaram, S., Chai, L., Zhang, R., Dekel, T., Isola, P.: Dreamsim: Learning new dimensions of human visual similarity using synthetic data (2023), https://arxiv.org/abs/2306.09344
11. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)
12. Gu, Y., Fang, G., Jiang, Y., Mao, W., Han, S., Cai, H., Shou, M.Z.: Anyflow: Any-step video diffusion model with on-policy flow map distillation (2026), https://arxiv.org/abs/2605.13724
13. Gu, Y., Dong, L., Wei, F., Huang, M.: Minillm: On-policy distillation of large language models (2026), https://arxiv.org/abs/2306.08543
14. He, W., Fu, S., Liu, M., Wang, X., Xiao, W., Shu, F., Wang, Y., Zhang, L., Yu, Z., Li, H., et al.: Mars: Mixture of auto-regressive models for fine-grained text-to-image synthesis. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 17123–17131 (2025)
15. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models (2021), https://arxiv.org/abs/2106.09685
16. Huang, S., Song, Y., Zhang, Y., Guo, H., Wang, X., Shou, M.Z., Liu, J.: Photodoodle: Learning artistic image editing from few-shot pairwise data. arXiv preprint arXiv:2502.14397 (2025)
17. Jiang, D., Jin, X., Liu, D., Wang, Z., Zheng, M., Du, R., Yang, X., Wu, Q., Li, Z., Gao, P., Yang, H., Hoi, S.: D-opsd: On-policy self-distillation for continuously tuning step-distilled diffusion models (2026), https://arxiv.org/abs/2605.05204

18. Jiang, D., Liu, D., Wang, Z., Wu, Q., Li, L., Li, H., Jin, X., Liu, D., Li, Z., Zhang, B., et al.: Distribution matching distillation meets reinforcement learning. arXiv preprint arXiv:2511.13649 (2025)
19. Labs, B.F.: Flux. https://github.com/black-forest-labs/flux (2024)
20. Labs, B.F.: FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2 (2025)
21. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., et al.: Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. arXiv preprint arXiv:2506.15742 (2025)
22. Li, Y., Zuo, Y., He, B., Zhang, J., Xiao, C., Qian, C., Yu, T., ang Gao, H., Yang, W., Liu, Z., Ding, N.: Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe (2026), https://arxiv.org/abs/2604.13016
23. Liu, D., Gao, P., Liu, D., Du, R., Li, Z., Wu, Q., Jin, X., Cao, S., Zhang, S., Li, H., et al.: Decoupled dmd: Cfg augmentation as the spear, distribution matching as the shield. arXiv preprint arXiv:2511.22677 (2025)
24. Liu, M., Ma, Y., Yang, Z., Dan, J., Yu, Y., Zhao, Z., Hu, Z., Liu, B., Fan, C.: Llm4gen: Leveraging semantic representation of llms for text-to-image generation. In: Proceedings of the AAAI conference on Artificial Intelligence. vol. 39, pp. 5523–5531 (2025)
25. Liu, M., She, D., Pang, J., Huang, Q., Ying, J., He, W., Hou, Y., Fu, S.: Tfcustom: Customized image generation with time-aware frequency feature guidance. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2714–2723 (2025)
26. Luo, S., Tan, Y., Huang, L., Li, J., Zhao, H.: Latent consistency models: Synthesizing high-resolution images with few-step inference. arXiv preprint arXiv:2310.04378 (2023)
27. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
28. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021), https://arxiv.org/abs/2103.00020
29. 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)
30. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: 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. pp. 22500–22510 (2023)
31. She, D., Fu, S., Liu, M., Jin, Q., Wang, H., Liu, M., Jiang, J.: Mosaic: Multi-subject personalized generation via correspondence-aware alignment and disentanglement. arXiv preprint arXiv:2509.01977 (2025)
32. Song, K., Zhu, Y., Liu, B., Yan, Q., Elgammal, A., Yang, X.: Moma: Multimodal llm adapter for fast personalized image generation. In: European Conference on Computer Vision. pp. 117–132. Springer (2024)
33. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models (2023), https://arxiv.org/abs/2303.01469
34. Song, Y., Liu, C., Shou, M.Z.: Omniconsistency: Learning style-agnostic consistency from paired stylization data (2025), https://arxiv.org/abs/2505.18445

35. Tan, Z., Liu, S., Yang, X., Xue, Q., Wang, X.: Ominicontrol: Minimal and universal control for diffusion transformer (2025), https://arxiv.org/abs/2411.15098
36. Wang, F.Y., Huang, Z., Bergman, A., Shen, D., Gao, P., Lingelbach, M., Sun, K., Bian, W., Song, G., Liu, Y., et al.: Phased consistency models. Advances in neural information processing systems **37**, 83951–84009 (2024)
37. Wang, Q., Bai, X., Wang, H., Qin, Z., Chen, A., Li, H., Tang, X., Hu, Y.: Instantid: Zero-shot identity-preserving generation in seconds. arXiv preprint arXiv:2401.07519 (2024)
38. Wang, Y., Liu, M., He, W., Zhang, L., Huang, Z., Zhang, G., Shu, F., Tao, Z., She, D., Yu, Z., et al.: Mint: Multi-modal chain of thought in unified generative models for enhanced image generation. arXiv e-prints pp. arXiv–2503 (2025)
39. Wei, Y., Zhang, Y., Ji, Z., Bai, J., Zhang, L., Zuo, W.: Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 15943–15953 (2023)
40. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
41. Wu, F., Liu, M., He, W., Wang, Z., Yu, Y.: Dcoar: Deep concept injection into unified autoregressive models for personalized text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 29265–29274 (2026)
42. Wu, J., Lian, H., Yang, J., Hao, D., Tian, Y., Tong, Y., Zhu, J., Chen, B., Qi, Q., Zhang, A., et al.: Loomvideo: Unifying multimodal inputs into video generation and editing. arXiv preprint arXiv:2606.06042 (2026)
43. Wu, K., Jiang, S., Ku, M., Nie, P., Liu, M., Chen, W.: Editreward: A human-aligned reward model for instruction-guided image editing (2026), https://arxiv.org/abs/2509.26346
44. Yang, W., Liu, W., Xie, R., Yang, K., Yang, S., Lin, Y.: Learning beyond teacher: Generalized on-policy distillation with reward extrapolation (2026), https://arxiv.org/abs/2602.12125
45. Yang, Z., Shen, G., Hou, L., Liu, M., Wang, L., Tao, X., Wan, P., Zhang, D., Chen, Y.C.: Rectifiedhr: Enable efficient high-resolution image generation via energy rectification. arXiv e-prints pp. arXiv–2503 (2025)
46. 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)
47. 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)
48. 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)
49. Ying, J., Liu, M., Wu, Z., Zhang, R., Yu, Z., Fu, S., Cao, S.Y., Wu, C., Yu, Y., Shen, H.L.: Restorerid: Towards tuning-free face restoration with id preservation. arXiv preprint arXiv:2411.14125 (2024)
50. Zhai, Y., Lin, K., Yang, Z., Li, L., Wang, J., Lin, C.C., Doermann, D., Yuan, J., Wang, L.: Motion consistency model: Accelerating video diffusion with disentangled motion-appearance distillation. Advances in Neural Information Processing Systems **37**, 111000–111021 (2024)

51. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models (2023), https://arxiv.org/abs/2302.05543
52. Zhang, Y., Yuan, Y., Song, Y., Wang, H., Liu, J.: Easycontrol: Adding efficient and flexible control for diffusion transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19513–19524 (2025)