INSTANTFUSION: A SHARED LATENT INTERFACE FOR HETEROGENEOUS IMAGE GENERATORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion/flow models have emerged as the de facto paradigms for image generation, giving rise to a diverse family of modern generators with increasingly heterogeneous architectures and representations. Despite their remarkable generation capabilities, the cross-model alignment of latent representations and the consistency of their underlying denoising dynamics remain underexplored. In this work, we uncover an intriguing cross-model compatibility: independently trained generators exhibit similar latent spatial organization, while directly handing off intermediate states between unaligned models can still preserve coarse visual content. Motivated by these observations, we propose **InstantFusion**, a lightweight shared latent interface that connects the denoising trajectories of heterogeneous image generators. Specifically, to enable flexible cross-model state translation, model-specific Latent AutoEncoders (**LAE**s) are introduced to project intermediate states from different generators into a shared latent space and decode them into the latent space of a target generator. This shared space is constrained by two complementary objectives: aligning intermediate states derived from the same image at matched noise levels and synchronizing their corresponding local denoising velocities. Once trained, InstantFusion allows denoising to seamlessly transition across frozen generators without requiring a dedicated mapping for every model pair. The versatility of InstantFusion is demonstrated across acceleration, multi-preference composition, cross-architecture on-policy distillation, and cross-model composition, achieving competitive performance across these diverse settings.

1 INTRODUCTION

Diffusion and flow models have emerged as the dominant paradigms for modern image generation (Song et al., 2023; Esser et al., 2024; Wu et al., 2025; Ho et al., 2020; Song et al., 2020; Lipman et al., 2022), enabling increasingly powerful generators with diverse architectural designs. Meanwhile, the model landscape has grown substantially more heterogeneous: generative models differ fundamentally in latent representations and conditioning mechanisms (Esser et al., 2024; Wu et al., 2025), network architectures and parameter scales (Cai et al., 2025; Black Forest Labs, 2025; Podell et al., 2024), and training or distillation recipes (Salimans & Ho, 2022; Song et al., 2023; Luo et al., 2023; Yin et al., 2024). Despite these structural differences, generative models ultimately learn to map noise onto closely related natural-image distributions under textual conditioning. However, they are typically studied and deployed as isolated systems. This isolation leaves it largely unexplored whether their latent representations admit meaningful cross-model alignment and whether their underlying generative dynamics remain compatible across distinct models.

In this work, we observe two indications of compatibility among independently trained generators: First, despite differences in latent dimensionality, tokenization, and parameterization, their intermediate states often exhibit closely related spatial organization (Fig. 2(a)). Second, directly handing an intermediate denoising state from one model to another, without any learned alignment, typically introduces artifacts but can preserve the main subject and coarse spatial layout of the generated image (Fig. 2(b)). These observations suggest that heterogeneous generators may encode partially shared visual structure and that their denoising processes are not entirely disconnected. This naturally raises a more practical question: *can such partial compatibility be turned into a reusable interface for transferring both latent states and denoising dynamics across heterogeneous latent spaces?*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

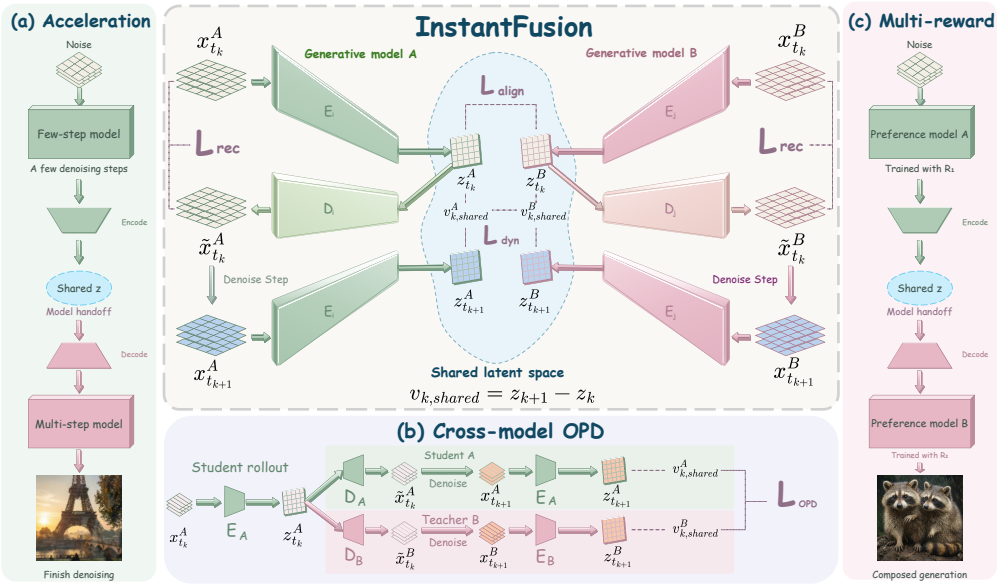

Figure 1: Overview of InstantFusion. During training, model-specific LAEs map the intermediate states of frozen generators into a shared representation space, where both latent states and denoising dynamics are aligned. During inference, a sender state is translated through the shared space into the receiver's latent space, and an adaptive criterion determines when the receiver takes over denoising.

Existing approaches provide only partial answers to this question. Cross-model latent alignment methods establish connections between individual states from different models (Asperti & Tonelli, 2023; Maiorca et al., 2023; Badrinath et al., 2024), but they primarily focus on model-pair-specific mappings, lacking both a shared representation space that can accommodate multiple generators and an explicit treatment of denoising dynamics. Meanwhile, trajectory stitching methods allow various denoisers to take over at different stages of sampling, but generally assume that the participating models operate within the same latent space (Pan et al., 2025). Therefore, transferring denoising trajectories among generators with heterogeneous latent spaces remains an open problem.

Motivated by the analysis above, we introduce **InstantFusion**, a lightweight shared latent interface that connects the denoising trajectories of heterogeneous image generators (Fig. 1). Each frozen generator uses a model-specific Latent AutoEncoder (**LAE**) that maps its intermediate states into a shared representation and reconstructs them in the model's native latent space. To learn this interface, we construct paired intermediate states from the same image at matched noise levels and align their encoded representations. Since state correspondence alone is insufficient for reliable trajectory transfer, additional alignment of the projected local denoising velocities is enforced so that translated states follow the receiver's denoising trajectory. At inference, a sender state is encoded by the sender's encoder and decoded by the receiver's decoder into the receiver's latent space. This design allows new generators to be incorporated through their own LAEs, enabling flexible cross-model transitions without dedicated pairwise mappings.

InstantFusion provides two fundamental interface capabilities: faithful reconstruction and cross-model state transfer. The latter directly enables cross-model composition: composing Qwen-Image (Wu et al., 2025) with SD3 or FLUX improves GenEval-v2 (Kamath et al., 2025) by 13 points and CVTG-2K (Du et al., 2025) by 0.18–0.28 over the corresponding base models. Building on interface-level capabilities, InstantFusion supports three downstream applications shown in Fig. 1: generation acceleration, cross-architecture on-policy distillation, and multi-preference composition. It achieves a 1.8–3.3$\times$ generation speedup while maintaining strong perceptual quality, balances multiple specialized preferences with competitive GenEval-v2 and HPSv2 (Wu et al., 2023) performance, and enables on-policy knowledge transfer from Qwen-Image to the architecturally distinct SD3. Experiments demonstrate that InstantFusion serves as a general interface for capability transfer and composition across heterogeneous image generators.

Our contributions are threefold: (1) Empirical evidence reveals partial compatibility among independently trained image generators, reflected in similar latent spatial organization and the preservation of coarse visual content under direct, zero-shot unaligned handoffs. (2) InstantFusion provides a

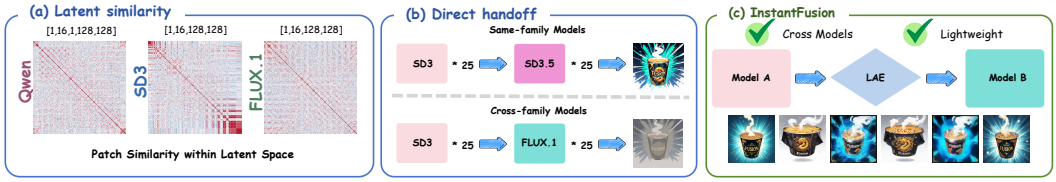

Figure 2: **Cross-model compatibility.** (a) Intermediate states from Qwen-Image, SD3, and FLUX.1 exhibit similar spatial self-similarity patterns. (b) Direct handoff between related models preserves generation quality, while cross-family handoff retains the main subject and coarse layout. (c) These observations motivate a lightweight interface for connecting heterogeneous generators.

lightweight shared latent interface that aligns intermediate representations and local denoising dynamics, enabling seamless denoising trajectory transfer across heterogeneous latent spaces without pair-specific mappings. (3) We demonstrate InstantFusion across generation acceleration, multi-preference composition, cross-architecture on-policy distillation, showing its potential as a general interface for capability transfer among heterogeneous generators.

2 Related Work

We discuss key studies here and defer others to the appendix C.

Text-to-Image Diffusion Models. Diffusion and flow-based models have become the prevailing frameworks for text-to-image generation (Ho et al., 2020; Lipman et al., 2022). Modern systems commonly perform denoising in learned latent spaces, combining pretrained autoencoders with text-conditioned denoisers (Rombach et al., 2022; Podell et al., 2024). Within this general framework, model designs have become increasingly diverse. Recent generators, including SD3, FLUX.1, Qwen-Image, and Z-Image, differ substantially in their latent parameterizations, denoising architectures, and training objectives (Esser et al., 2024; Black Forest Labs, 2025; Wu et al., 2025; Cai et al., 2025). Yet these independently developed generators are typically studied in isolation, leaving the relationships among their intermediate representations and denoising dynamics largely unexplored.

Cross-Model Latent Alignment. A related line of work investigates whether independently trained generative models develop structurally compatible latent representations. These studies establish cross-model correspondences through linear mappings, relative representations, or closed-form transformations (Asperti & Tonelli, 2023; Moschella et al., 2022; Maiorca et al., 2023; Badrinath et al., 2024). More recently, learned bridges have been proposed to connect incompatible teacher and student VAE latent spaces, enabling cross-space diffusion distillation (Nguyen et al., 2026). Nevertheless, these approaches operate primarily on isolated latent states and do not model how corresponding states evolve along denoising trajectories. It therefore remains unclear whether static latent correspondence also implies compatibility between the underlying generative dynamics.

3 Methodology

InstantFusion enables handoff between frozen image generators by aligning their latent representations and denoising dynamics. In Sec. 3.2, we identify spatial self-similarity and preserved coarse semantics across unaligned models. Building on this, Sec. 3.3 equips each generator with a model-specific LAE to align and reconstruct intermediate states, while Sec. 3.4 synchronizes dynamic velocities to harmonize trajectories. For inference, Sec. 3.5 introduces an adaptive strategy tracking predicted clean latents for optimal switching. Fig. 1 provides an overview of our full pipeline.

3.1 Preliminaries

Flow Matching. Flow Matching (FM) (Lipman et al., 2022) learns a vector field that transports noise to data. Let $\mathbf{x}_0 \sim \pi_0$ be a clean latent drawn from the data latent distribution π_0, and let $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ be standard Gaussian noise. At time $t \in [0, 1]$, the linear interpolation is

$$\mathbf{x}_t = (1 - t)\mathbf{x}_0 + t\boldsymbol{\epsilon}, \tag{1}$$

where $t = 0$ denotes clean data and $t = 1$ denotes noise. The vector field $v_\theta(\mathbf{x}_t, t, c)$, parameterized by θ and conditioned on prompt representation c, is trained to predict the path velocity $\boldsymbol{\epsilon} - \mathbf{x}_0$:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{\mathbf{x}_0, c, \boldsymbol{\epsilon}, t}\left[\|v_\theta(\mathbf{x}_t, t, c) - (\boldsymbol{\epsilon} - \mathbf{x}_0)\|_2^2\right]. \tag{2}$$

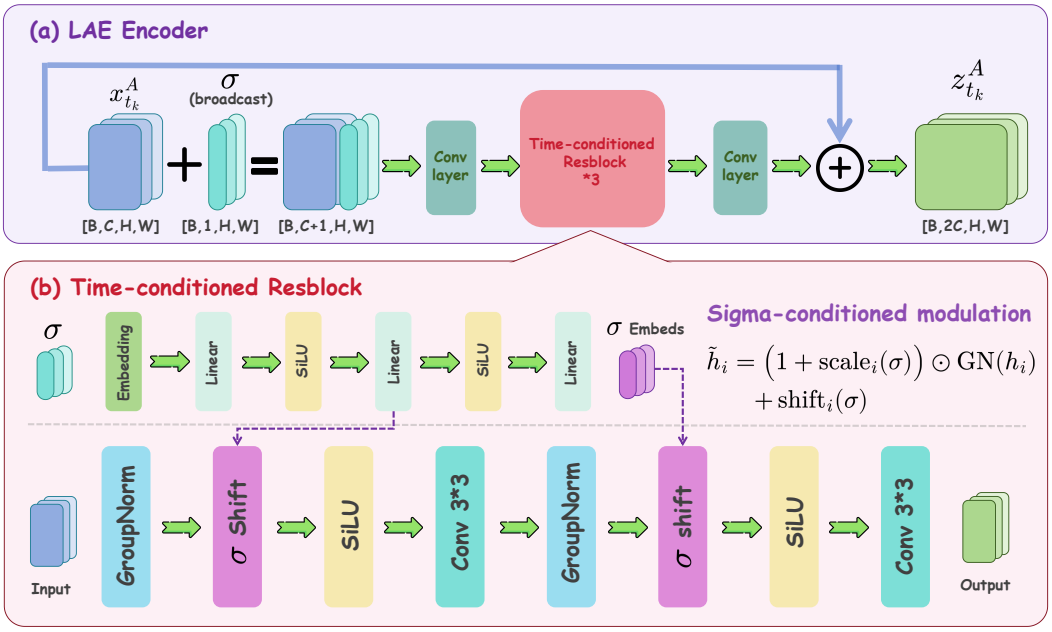

Figure 3: **Architecture of the LAE.** (a) σ denotes the scalar noise level associated with denoising step t_k. It is spatially broadcast and concatenated with the intermediate state $\mathbf{x}_{t_k}^A$. Convolutional layers and time-conditioned residual blocks then map the input to its encoded representation $\mathbf{z}_{t_k}^A$, with a skip connection preserving the original state. (b) Details of the time-conditioned resblock: Each residual block embeds σ and predicts scale and shift parameters that modulate group-normalized features, conditioning the transformation on the current denoising stage.

Here, $\|\cdot\|_2$ denotes the Euclidean norm. Generation integrates the learned field from $t = 1$ to $t = 0$.

3.2 Observation: Shared Structure Across Models

We identify two concrete signals of latent compatibility across independently trained generators. First, intermediate representations maintain equivalent spatial dimensions across model architectures, such as Qwen-Image ($[1, 16, 1, 128, 128]$) alongside SD3 and FLUX.1 ($[1, 16, 128, 128]$), yielding similar spatial patterns that reflect comparable information density (Fig. 2(a)). Second, halting SD3 partway through denoising and handing off the intermediate state to SD3.5 or FLUX.1 to complete the remaining steps produces contrasting outcomes: while same-family handoff yields a seamless image, cross-family handoff introduces local artifacts yet clearly preserves the central cup and its coarse layout (Fig. 2(b)). These artifacts confirm that native latent spaces cannot be directly interchanged, whereas the preserved cup layout suggests that some coarse semantic structure is shared across the latent spaces. These insights directly motivate an interface that aligns and reconstructs intermediate states before harmonizing their denoising dynamics.

3.3 Latent Alignment and Reconstruction

Let $m \in \{1, \ldots, M\}$ index the $M \geq 2$ frozen generators. All generators use the same frozen variational autoencoder (VAE). We encode an image into a clean latent $\mathbf{x}_0$ and construct a common noisy input $\mathbf{x}_t$ using the same noise sample $\boldsymbol{\epsilon}$ and time t.

Each generator is equipped with a lightweight, time-conditioned LAE consisting of an encoder E_m and a decoder D_m, whose architecture is illustrated in Fig. 3.

$$\mathbf{z}_t^m = E_m(\mathbf{x}_t, t), \qquad \widetilde{\mathbf{x}}_t^m = D_m(\mathbf{z}_t^m, t), \tag{3}$$

Cross-model handoff requires the LAE encoders to map corresponding intermediate states to a common representation that can be consumed by any model's decoder. We enforce this correspondence with the pairwise alignment loss, where A and B index distinct generators.

$$\mathcal{L}_{\text{align}} = \frac{2}{M(M-1)} \sum_{1 \leq A < B \leq M} \left\| \mathbf{z}_t^A - \mathbf{z}_t^B \right\|_2^2. \tag{4}$$

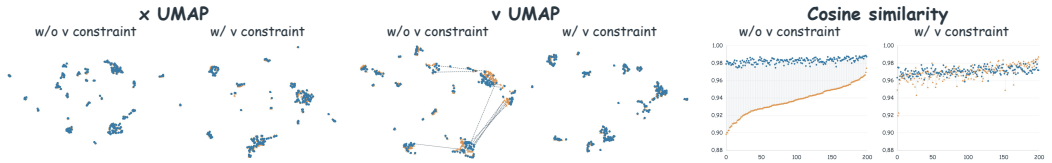

Figure 4: **Effect of denoising-dynamics alignment.** Training without $\mathcal{L}_{\text{dyn}}$ establishes correspondence between intermediate states $\mathbf{x}$, but leaves their associated denoising velocities $\mathbf{v}$ poorly aligned. Adding $\mathcal{L}_{\text{dyn}}$ improves velocity alignment. Each line in the UMAP visualization connects the $\mathbf{v}$ feature points produced by different models when conditioned on the same prompt.

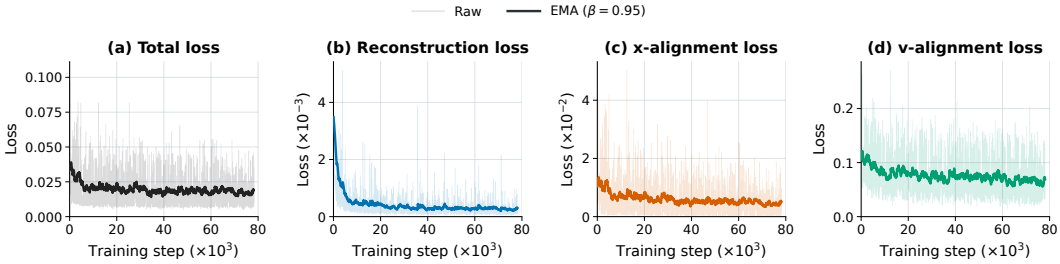

Figure 5: Training dynamics of the model-specific LAEs. The total loss and the three component losses ($\mathcal{L}_{\text{rec}}$, $\mathcal{L}_{\text{align}}$, and $\mathcal{L}_{\text{dyn}}$) decrease steadily and converge during training.

However, minimizing $\mathcal{L}_{\text{align}}$ alone admits a trivial solution that satisfies the pairwise-distance objective while discarding all input-dependent information. We therefore introduce a reconstruction loss to preserve the latent information required for subsequent denoising

$$\mathcal{L}_{\text{rec}} = \frac{1}{M}\sum_{m=1}^{M} \|\widetilde{\mathbf{x}}_t^m - \mathbf{x}_t\|_2^2. \tag{5}$$

where $\widetilde{\mathbf{x}}_t^m$ denotes the reconstructed latent obtained by passing $\mathbf{x}_t$ through model m's LAE encoder and decoder. Moreover, to encourage comparable representation magnitudes across models for the same image and timestep, we introduce a norm-consistency regularization term

$$\mathcal{L}_{\text{reg}} = \frac{2}{M(M-1)} \sum_{1\le A<B\le M} \left(\|\mathbf{z}_t^A\|_2 - \|\mathbf{z}_t^B\|_2\right)^2. \tag{6}$$

3.4 Denoising Dynamics Alignment

To examine whether latent state alignment also aligns denoising dynamics, Fig. 4 analyzes shared states and their one-step displacements using UMAP and cosine similarity. The UMAP visualizations show that corresponding states from different models cluster closely after state alignment, whereas their denoising displacements remain separated. Consistently, cross-model cosine similarity is high for paired states but lower for their displacements. This mismatch motivates explicitly aligning one-step denoising displacements in the shared space.

Let $1 = t_0 > \cdots > t_K = 0$ be a common sampling schedule with K denoising steps. For step $k \in \{0, \ldots, K-1\}$, define $\Delta t_k = t_{k+1} - t_k$. At step k, generator m applies its frozen vector field v_{θ_m} to the reconstructed state, conditioned on its own encoding c^m of the shared prompt

$$\mathbf{x}_{t_{k+1}}^m = \widetilde{\mathbf{x}}_{t_k}^m + \Delta t_k \, v_{\theta_m}(\widetilde{\mathbf{x}}_{t_k}^m, t_k, c^m), \tag{7}$$

where $\mathbf{x}_{t_{k+1}}^m$ denotes the resulting model-specific latent state. Its displacement in the shared space is

$$\Delta\mathbf{z}_k^m = E_m(\mathbf{x}_{t_{k+1}}^m, t_{k+1}) - \mathbf{z}_{t_k}^m. \tag{8}$$

Since all generators use the same time interval, we directly align these displacements:

$$\mathcal{L}_{\text{dyn}} = \frac{2}{M(M-1)} \sum_{1\le A<B\le M} \text{MSE}(\Delta\mathbf{z}_k^A, \Delta\mathbf{z}_k^B), \tag{9}$$

Therefore, the LAEs are jointly optimized with

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{align}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}} + \lambda_{\text{dyn}}\mathcal{L}_{\text{dyn}}, \tag{10}$$

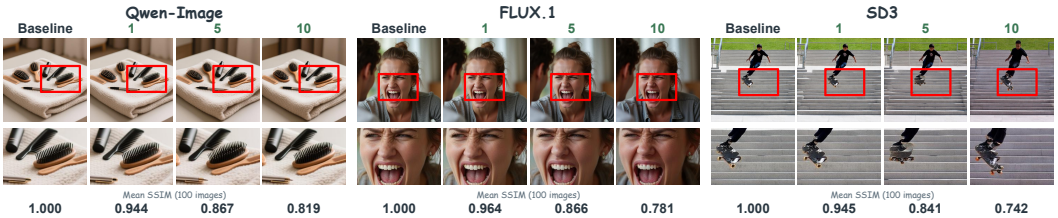

Figure 6: Qualitative comparison of Qwen-Image, FLUX.1, and SD3 outputs after 1, 5, and 10 repeated latent round trips. The original outputs before any round trip serve as baselines.

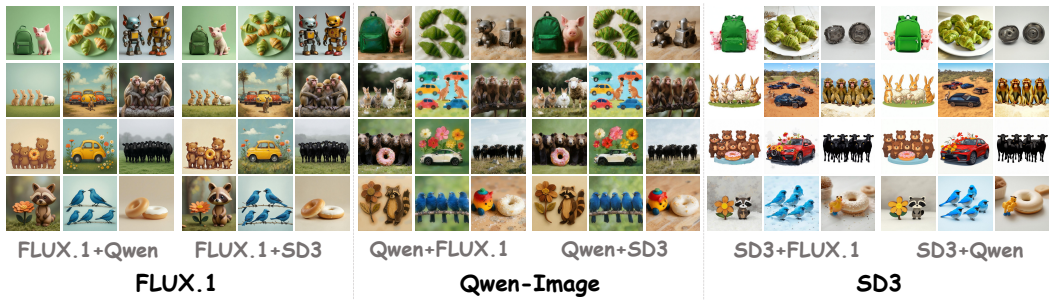

Figure 7: Qualitative results of cross-model latent transfer among heterogeneous generators.

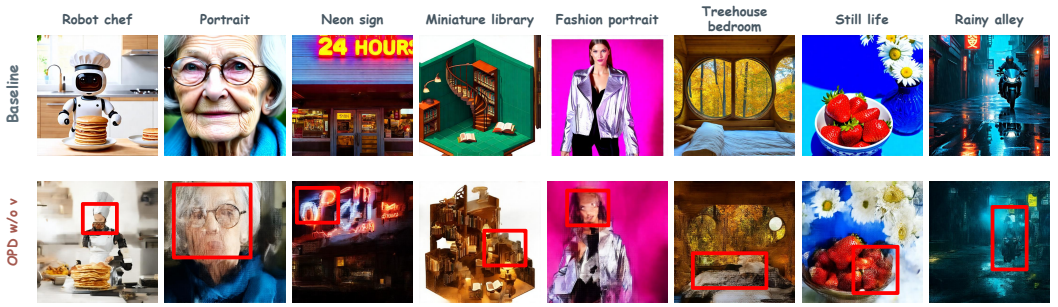

Figure 8: Failure case of cross-model on-policy distillation after removing $\mathcal{L}_{\text{dyn}}$.

where λ_{reg} and λ_{dyn} denote the relative weights assigned to norm regularization and dynamics alignment. We use a small λ_{reg} for training stability. The objective is averaged over sampled training states, while the generators remain frozen and only the LAEs are optimized.

3.5 Adaptive Model Handoff

The optimal handoff step varies across model pairs, making manual tuning impractical and unscalable as more generators are integrated. Crucially, executing a handoff too early in the trajectory proves ineffective because the sender model has yet to crystallize stable semantic structures from pure noise. To overcome this limitation, the handoff step is determined online, dynamically guided by the temporal consistency of the sender's predicted clean states across consecutive denoising steps.

For a transition from model A to model B, let $\mathbf{x}_{t_k}^A$ denote the current latent state of model A at inference step k. Its translation into the latent space of model B is

$$\mathbf{x}_{t_k}^{A\to B} = D_B(E_A(\mathbf{x}_{t_k}^A, t_k), t_k), \tag{11}$$

For any latent state $\mathbf{x}$ at time t with conditioning c, define model m's clean-state prediction:

$$\mathbf{x}_{0,k} = \mathbf{x} - t\, v_\theta(\mathbf{x}, t_k, c). \tag{12}$$

Here, $\mathbf{x} - t\, v_\theta(\mathbf{x}, t_k, c)$ estimates the clean latent under the linear flow path. We quantify the prediction stability by measuring the agreement between its clean predictions at two consecutive steps:

$$S_k^A = \frac{\left\|\mathbf{x}_{0,k}^A - \mathbf{x}_{0,k-1}^A\right\|_2}{\max\left(\left\|\mathbf{x}_{0,k-1}^A\right\|_2, \kappa\right)}, \qquad \kappa = 10^{-8}. \tag{13}$$

Table 1: **Evaluation of the shared latent interface.** (a) Reconstruction fidelity across repeated latent round-trips. (b) Quantitative generation performance under cross-architecture on-policy distillation. (c) Reconstruction fidelity degradation upon removing the reconstruction loss $\mathcal{L}_{\text{rec}}$.

(a) Repeated latent round trips.

Shared Space	Trips	PSNR ↑	SSIM ↑	LPIPS ↓
SD3–Qwen	1	31.375	0.9417	0.04028
	5	22.977	0.8431	0.12630
	10	20.043	0.7670	0.18892
FLUX.1–Qwen	1	32.579	0.9496	0.03663
	5	22.663	0.8450	0.13064
	10	19.106	0.7621	0.21014
SD3–FLUX.1–Qwen	1	32.397	0.9511	0.03257
	5	23.445	0.8580	0.11006
	10	20.119	0.7808	0.17591

(b) Cross-architecture on-policy distillation.

Models	HPSv2				
	Anime	Concept	Paint	Photo	Avg.
Qwen-Image (Teacher)	29.16	28.22	28.40	25.90	27.92
SD3 Baseline	**28.21**	26.25	26.01	24.59	26.27
SD3–Qwen OPD	27.58	**26.54**	**26.98**	**24.72**	**26.46**

(c) Reconstruction fidelity without $\mathcal{L}_{\text{rec}}$.

Domain	PSNR ↑	SSIM ↑	LPIPS ↓
SD3 Average	10.4743	0.3627	0.9491
Qwen-Image Average	8.6137	0.4794	0.9328

Table 2: Comparison of base models and cross-model combinations, demonstrating that guidance from a small set of high-quality generators can improve the generation quality of weaker models.

Models	New Model	GenEval-v2	CVTG-2K	Clip	OneIG-Bench					
					Alignment	Text	Diversity	Style	Reasoning	Avg.
▼ *Stronger Model*										
Qwen-Image (Wu et al., 2025)	–	**41.73**	0.6699	25.46	0.8304	0.9138	0.2091	0.2230	0.2486	0.4849
▼ *Generation Baselines*										
SD3 (Esser et al., 2024)	–	22.14	0.3682	28.65	0.7934	0.4161	0.2327	0.3596	0.2770	0.4157
FLUX.1 (Black Forest Labs, 2025)	–	22.86	0.5180	27.75	0.7894	0.5736	0.2136	0.3663	0.2653	0.4416
▼ *Hybrid Models (Ours)*										
Qwen-Image + SD3	✓	35.06 ↑12.92	0.6488 ↑0.2806	26.98	0.8072 ↑0.0138	0.6683 ↑0.2522	0.2645 ↑0.0318	0.3748 ↑0.0152	0.2730	0.4776 ↑0.0619
Qwen-Image + FLUX.1	✓	36.54 ↑13.68	0.6983 ↑0.1803	28.86 ↑1.11	0.8262 ↑0.0368	0.7854 ↑0.2118	0.2269 ↑0.0133	0.3785 ↑0.0122	0.2868 ↑0.0215	0.5007 ↑0.0591

Table 3: **Quantitative comparison of generation quality and inference efficiency.** Performance benchmarks across baselines, existing acceleration techniques, and InstantFusion. Furthermore, InstantFusion exhibits orthogonal compatibility when integrated with training-free acceleration methods like DiCache, yielding cumulative speedups without compromising generation fidelity.

Models	New model	Time ↓	GenEval -v2 ↑	Clip ↑	Aesthetic ↑	OneIG-Bench ↑					
						Alignment	Text	Diversity	Style	Reasoning	Avg.
▼ *Generation Baselines*											
Qwen-Image (Wu et al., 2025)	–	18.84	41.73	25.46	5.9797	0.8304	0.9138	0.2091	0.2230	0.2486	0.4849
Qwen-Few-Step	–	0.893	32.94	27.48	6.1430	0.8804	0.9752	0.0885	0.3234	0.3069	0.5148
FLUX.2-Klein (Black Forest Labs, 2025)	–	0.516	25.43	27.71	5.9017	0.5718	0.1863	**0.3930**	0.4059	0.2415	0.3597
Z-Image-Turbo (Cai et al., 2025)	–	3.521	39.21	25.02	4.9892	0.8658	0.9788	0.2193	0.4573	**0.3137**	**0.5669**
T-Stitch (Pan et al., 2025)	–	20.76	7.762	21.24	4.4475	0.3037	0.3349	0.2445	0.2255	0.2208	0.2659
▼ *InstantFusion (Ours)*											
Z-Image-Turbo + Qwen-Image	✓	4.373	**33.34**	27.90	5.3957	0.8043	0.8837	0.3070	0.3933	0.3046	0.5386
FLUX.2-Klein + Qwen-Image	✓	**2.545**	27.19	29.18	5.7441	0.8444	0.9371	0.2454	**0.3977**	0.3089	0.5467
Qwen-Few-Step + Qwen-Image	✓	10.64	32.70	33.15	6.0821	**0.8808**	0.9800	0.1192	0.3205	0.3015	0.5204
Qwen-Few-Step + Qwen-Image + DiCache	✓	5.700	32.46	**33.18**	**6.1076**	0.8780	**0.9812**	0.1190	0.3284	0.3033	0.5220

Given a similarity threshold $\gamma \in [-1, 1]$, we select the first qualifying step. The receiver then continues denoising from $\mathbf{x}^{A\to B}_{t_{k_{\text{switch}}}}$.

$$k_{\text{switch}} = \min\left\{k \in \{0, \dots, K-1\} \;\middle|\; s^A_k \geq \gamma\right\}. \tag{14}$$

4 Experiments

4.1 Experimental Setup

The experiments are conducted on four representative text-to-image generators: SD3 (Esser et al., 2024), Z-Image (Cai et al., 2025), FLUX.1 (Black Forest Labs, 2025), and Qwen-Image (Qwen) (Wu et al., 2025), covering diverse architectures and latent parameterizations. We use 50,000 image-text pairs from the VisPrompt dataset (Yi et al., 2026) for training, while keeping all generator backbones frozen to optimize only the LAEs. Specifically, the LAEs are parameterized with 32 shared channels, 64 hidden channels, and a 128-dimensional σ embedding, trained with an initial learning rate of 10^{-4}. The reconstruction, alignment, dynamics-alignment, and norm regularization losses are balanced with weights of 1.0, 1.0, 0.1, and 10^{-4}, respectively. Further implementation details are provided in the appendix B. As demonstrated in Fig. 5, all training objectives decrease steadily throughout optimization, confirming that the LAEs achieve accurate reconstruction while effectively aligning cross-model latent states and their corresponding denoising dynamics.

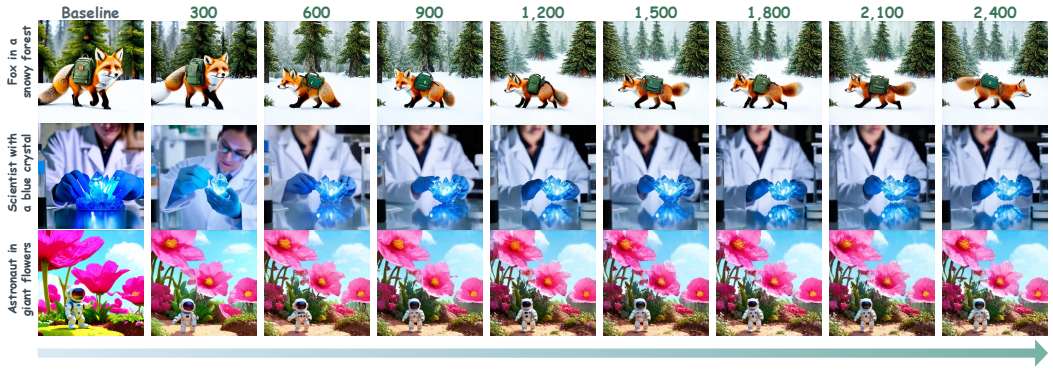

Figure 9: Evolution of generated images over training steps during cross-model OPD.

Table 4: Evaluation results of reward-specialized GRPO models and model combination strategies.

Models	New Model	GenEval-v2	HPSv2					HPSv3				
			Anime	Concept	Paint.	Photo	Avg.	Anime	Concept	Paint.	Photo	Avg.
▼ *Reward-Specialized GRPO Baselines*												
GRPO baseline (HPSv2)	–	28.83	32.21	31.67	31.91	28.95	31.18	12.14	12.88	13.30	9.94	12.07
GRPO baseline (HPSv3)	–	26.53	31.42	30.57	30.93	27.63	30.14	12.31	13.06	13.48	9.73	12.15
GRPO baseline (GenEval-v2)	–	35.89	28.55	27.28	27.47	25.47	27.19	10.98	11.30	11.71	7.69	10.42
▼ *Direct V Fusion*												
HPSv2 + GenEval-v2	–	29.39	31.10	30.58	30.86	27.40	29.99	11.62	12.34	12.78	8.94	11.42
HPSv2 + HPSv3	–	28.58	32.11	31.50	31.72	28.75	31.02	12.33	13.03	13.46	10.22	12.26
▼ *Multi-preference*												
HPSv2 + GenEval-v2	–	30.49	31.86	30.89	31.22	28.41	30.60	11.54	11.92	12.36	9.55	11.34
HPSv2 + HPSv3	–	24.53	28.83	27.36	27.57	25.70	27.37	11.58	11.99	12.34	8.04	10.99
▼ *InstantFusion (Ours)*												
GenEval-v2 + HPSv2	✓	**30.97**	31.91	31.37	31.52	27.92	30.68	12.09	12.72	13.07	9.61	11.87
HPSv2 + HPSv3	✓	28.94	**32.18**	**31.66**	**31.89**	**28.69**	**31.11**	**12.26**	**13.00**	**13.36**	**10.13**	**12.19**

4.2 Performance of LAE

Reconstruction Fidelity. We first jointly train one LAE per generator to project its native latent space into a shared representation. To verify how much information is preserved by the shared representation, clean latents are mapped into the shared space and reconstructed back into their native domains. As reported in Tab. 1a, single-trip encoding achieves high-fidelity reconstruction (e.g., PSNR $>$ 31 dB, SSIM $>$ 0.94), demonstrating that the shared space retains sufficient detail, though performance degrades as trips increase. Qualitative results are illustrated in Fig. 6.

Cross-Model Transfer. Beyond preserving native latents, the shared interface enables cross-generator exchange of intermediate states. To assess this capability, intermediate representations are translated via the shared space to resume denoising across SD3, FLUX.1, and Qwen-Image. As detailed in Tab. 2, guidance from Qwen-Image improves the GenEval-v2 scores of SD3 and FLUX.1 by 12.92 and 13.68, respectively, reaching absolute scores of 35.06 and 36.54. The corresponding combinations also improve CVTG-2K by 0.2806 and 0.1803, and the OneIG-Bench (Chang et al., 2026) average by 0.0619 and 0.0591. Together with the qualitative results in Fig. 7, these gains demonstrate effective state transfer across heterogeneous architectures.

4.3 Applications of LAE

Generation Acceleration. Tab. 3 shows that InstantFusion accelerates Qwen-Image by combining it with efficient generators. Pairing Qwen-Image with Z-Image-Turbo or FLUX.2-Klein reduces inference time from 18.84s to 4.373s and 2.545s, while achieving GenEval-v2 scores of 33.34, 27.19 and OneIG-Bench averages of 0.5386, 0.5467, without the quality collapse observed in T-Stitch (Pan et al., 2025). InstantFusion is further compatible with existing acceleration methods: applying DiCache (Bu et al., 2026) to InstantFusion reduces latency from 10.64s to 5.70s, while GenEval-v2 changes only from 32.70 to 32.46 and the OneIG-Bench average marginally increases.

Cross-Architecture On-Policy Distillation. We leverage InstantFusion for cross-architecture on-policy distillation (OPD), transferring generative priors from a stronger teacher (Qwen-Image) to a student (SD3) (Jiang et al., 2026; Li et al., 2026; Fang et al., 2026; Zhou et al., 2026; Fu et al., 2026b). As shown in Fig. 1(b), InstantFusion projects intermediate states from both models into the shared latent space, enabling the teacher to compute velocity targets directly along the student's

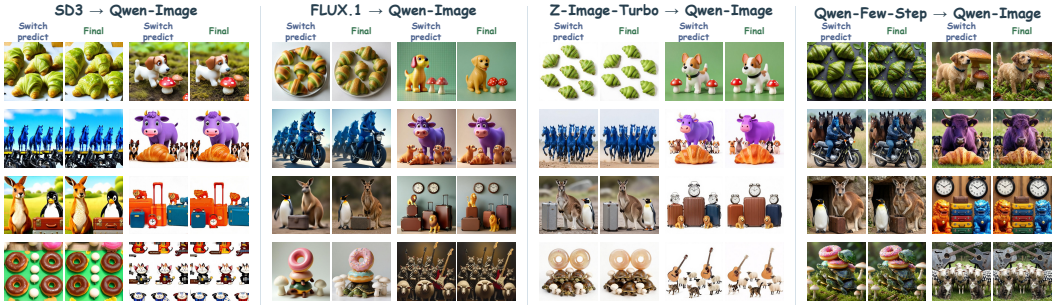

Figure 10: Comparison between the predicted clean images at adaptively selected handoff steps and the corresponding fully denoised outputs, demonstrating preservation of global structure.

Table 5: Capabilities enabled by progressively adding the training objectives. Each column indicates whether the resulting interface supports the corresponding capability.

Training Objectives	Cross-Model State Translation	Information-Preserving Reconstruction	Denoising-Dynamics Consistency
$\mathcal{L}_{\text{align}}$	✓	×	×
$\mathcal{L}_{\text{align}} + \mathcal{L}_{\text{rec}}$	✓	✓	×
InstantFusion	✓	✓	✓

Table 6: Comparison between the baseline and InstantFusion for cross-model transfer. InstantFusion consistently outperforms the baseline across all evaluated metrics

Transfer Direction	Method	GenEval-v2	CVTG-2K	HPSv2 Avg.
Qwen-Image → SD3	Conv. layer baseline	27.86	0.2744	20.50
	InstantFusion	35.06	0.6488	25.34
Qwen-Image → FLUX.1	Conv. layer baseline	17.37	0.3566	14.79
	InstantFusion	36.54	0.6983	25.03

trajectory. Consequently, Tab. 1b shows that cross-model OPD improves SD3 from 26.27 to 26.46 in HPSv2, with qualitative evolutions illustrated in Fig. 9.

Multi-Preference Composition. Tab. 4 evaluates the composition of independently trained reward experts (Lee et al., 2025; Rame et al., 2023). When combining experts optimized for GenEval-v2 and HPSv2, InstantFusion achieves scores of 30.97 and 30.68, outperforming velocity fusion and closely matching jointly trained multi-reward GRPO. This result shows that InstantFusion can combine existing single-reward experts without the additional optimization required by joint multi-reward training. We observe the same trend when composing HPSv2 and HPSv3 experts, indicating that the shared latent space preserves the respective preferences learned by the individual experts.

4.4 Ablation and Analysis

Ablation of Training Objectives. The three objectives serve complementary functions: $\mathcal{L}_{\text{align}}$ enables cross-model state translation, $\mathcal{L}_{\text{rec}}$ preserves information fidelity for accurate reconstruction, and $\mathcal{L}_{\text{dyn}}$ aligns denoising trajectories across architectures (Tab. 5). As shown in Fig. 8, omitting $\mathcal{L}_{\text{dyn}}$ causes training divergence during cross-architecture OPD, even with aligned latent spaces. Conversely, removing $\mathcal{L}_{\text{rec}}$ allows cross-model translation but significantly impairs reconstruction quality. These distinct failure modes demonstrate that all three loss terms are essential.

Adaptive Handoff Analysis. To verify whether the adaptive criterion selects a suitable transition step, we compare the clean image predicted at the selected handoff step with the final generated image. As shown in Fig. 10, this early prediction already preserves the coarse spatial layout of the final output. Crucially, this structural consistency indicates that the selected handoff often occurs after the global structure has largely stabilized.

LAE Architecture Design. We conduct an ablation to validate the effectiveness of our time-conditioned residual blocks in LAE. Replacing these time-conditioned blocks with standard convolutional layers leads to a substantial drop in cross-model handoff performance (Tab. 6), reducing GenEval-v2 by 7.20 when combining Qwen-Image with SD3 and by 19.17 with FLUX.1. This performance gap under an identical layer highlights the representational efficiency of our design.

5 Conclusion

In this work, we introduced **InstantFusion**, a shared latent interface that connects heterogeneous image generators by aligning their latent states. Without modifying the frozen generators, InstantFusion enables cross-model state transfer. Nevertheless, results across a diverse range of downstream tasks further demonstrate the flexibility and broad applicability of InstantFusion. Future work may develop the learned shared space into a reusable pretrained interface, such that new generators can be incorporated with minimal adaptation and benefit from a common starting point for model composition, knowledge transfer, and downstream optimization.

AI USE STATEMENT

In this work, we used generative AI tools for minor language polishing. We have not used generative AI tools for research ideation or experimental design, and the rest of the required disclosure tasks are not applicable to this work. Additionally, we did not use generative AI tools for tasks with recommended disclosure. We have reviewed all AI-assisted work. All AI-assisted edits were limited to minor language polishing. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses only publicly available datasets. To the best of our knowledge, the use of these datasets does not raise concerns related to discrimination, bias, fairness, privacy, or security. No new data involving human participants were collected for this study.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we will publicly release the code and checkpoints upon publication.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Andrea Asperti and Valerio Tonelli. Comparing the latent space of generative models. *Neural Computing and Applications*, 35(4):3155–3172, 2023.

Charumathi Badrinath, Usha Bhalla, Alex Oesterling, Suraj Srinivas, and Himabindu Lakkaraju. All roads lead to rome? exploring representational similarities between latent spaces of generative image models. *arXiv preprint arXiv:2407.13449*, 2024.

Black Forest Labs. Flux.2: Frontier visual intelligence. Black Forest Labs Official Website, 2025. URL https://blackforestlabs.ai/flux-2-frontier-visual-intelligence/. Release; FLUX.2 provides significant improvements in visual quality, prompt adherence, and architectural efficiency.

Jiazi Bu, Pengyang Ling, Yujie Zhou, Yibin Wang, Yuhang Zang, Dahua Lin, and Jiaqi Wang. Dicache: Let diffusion model determine its own cache. In *International Conference on Learning Representations*, volume 2026, pp. 73778–73803, 2026.

Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Aiming Hao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Yuming Jiang, et al. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

Jingjing Chang, Yixiao Fang, Peng Xing, Shuhan Wu, Wei Cheng, Rui Wang, Xianfang Zeng, Gang Yu, and Hai-Bao Chen. Oneig-bench: Omni-dimensional nuanced evaluation for image generation. *Advances in Neural Information Processing Systems*, 38, 2026.

Nikai Du, Zhennan Chen, Zhizhou Chen, Shan Gao, Xi Chen, Zhengkai Jiang, Jian Yang, and Ying Tai. Textcrafter: Accurately rendering multiple texts in complex visual scenes, 2025. URL https://arxiv.org/abs/2503.23461.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Zhen Fang, Wenxuan Huang, Yu Zeng, Yiming Zhao, Shuang Chen, Kaituo Feng, Yunlong Lin, Lin Chen, Zehui Chen, Shaosheng Cao, et al. Flow-opd: On-policy distillation for flow matching models. *arXiv preprint arXiv:2605.08063*, 2026.

Siming Fu, Zheming Fu, Ruizhe He, Hualiang Wang, Jie Huang, Xiaoxiao Ma, Mingchen Zhong, Weihu Huang, Xiaoxuan He, and Haojun Xu. Any-opd: Heterogeneous on-policy distillation for flow-matching models via representation-space bridging. *arXiv preprint arXiv:2608.03316*, 2026a.

Siming Fu, Haojun Xu, Ruizhe He, Zheming Fu, Hualiang Wang, Jie Huang, Xiaoxiao Ma, Mingchen Zhong, Weihu Huang, Xiaoxuan He, et al. Poly-opd: Heterogeneous multi-teacher on-policy distillation for capability-selectable flow models. *arXiv preprint arXiv:2608.04349*, 2026b.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: On-policy distillation of large language models. *arXiv preprint arXiv:2306.08543*, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Dengyang Jiang, Xin Jin, Dongyang Liu, Zanyi Wang, Mingzhe Zheng, Ruoyi Du, Xiangpeng Yang, Qilong Wu, Zhen Li, Peng Gao, Harry Yang, and Steven Hoi. D-opsd: On-policy self-distillation for continuously tuning step-distilled diffusion models. *arXiv preprint arXiv:2605.05204*, 2026.

Amita Kamath, Kai-Wei Chang, Ranjay Krishna, Luke Zettlemoyer, Yushi Hu, and Marjan Ghazvininejad. Geneval 2: Addressing benchmark drift in text-to-image evaluation. *arXiv preprint arXiv:2512.16853*, 2025.

Kyungmin Lee, Xiahong Li, Qifei Wang, Junfeng He, Junjie Ke, Ming-Hsuan Yang, Irfan Essa, Jinwoo Shin, Feng Yang, and Yinxiao Li. Calibrated multi-preference optimization for aligning diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18465–18475. IEEE, 2025.

Quanhao Li, Junqiu Yu, Kaixun Jiang, Yujie Wei, Zhen Xing, Pandeng Li, Ruihang Chu, Shiwei Zhang, Yu Liu, and Zuxuan Wu. Diffusionopd: A unified perspective of on-policy distillation in diffusion models. *arXiv preprint arXiv:2605.15055*, 2026.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Enshu Liu, Xuefei Ning, Zinan Lin, Huazhong Yang, and Yu Wang. Oms-dpm: Optimizing the model schedule for diffusion probabilistic models. In *International Conference on Machine Learning*, pp. 21915–21936. PMLR, 2023.

Feng Liu, Shiwei Zhang, Xiaofeng Wang, Yujie Wei, Haonan Qiu, Yuzhong Zhao, Yingya Zhang, Qixiang Ye, and Fang Wan. Timestep embedding tells: It's time to cache for video diffusion model. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7353–7363. IEEE, 2025a.

Jiacheng Liu, Chang Zou, Yuanhuiyi Lyu, Junjie Chen, and Linfeng Zhang. From reusing to forecasting: Accelerating diffusion models with taylorseers. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15853–15863. IEEE, 2025b.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.

Valentino Maiorca, Luca Moschella, Antonio Norelli, Marco Fumero, Francesco Locatello, and Emanuele Rodolà. Latent space translation via semantic alignment. *Advances in Neural Information Processing Systems*, 36:55394–55414, 2023.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. *arXiv preprint arXiv:2209.15430*, 2022.

Anh Nguyen, Ngan Nguyen, Duc Vu, Trung Dao, Viet Nguyen, Quan Dao, Kien Nguyen, Chi Tran, Phong Nguyen, Khoi Nguyen, et al. Cross-space distillation: Teaching one-step students with modern diffusion teachers. *arXiv preprint arXiv:2606.32020*, 2026.

Zizheng Pan, Bohan Zhuang, De-An Huang, Weili Nie, Zhiding Yu, Chaowei Xiao, Jianfei Cai, et al. T-stitch: Accelerating sampling in pre-trained diffusion models with trajectory stitching. In *International Conference on Learning Representations*, volume 2025, pp. 6103–6137, 2025.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, volume 2024, pp. 1862–1874, 2024.

Alexandre Rame, Guillaume Couairon, Corentin Dancette, Jean-Baptiste Gaya, Mustafa Shukor, Laure Soulier, and Matthieu Cord. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. *Advances in Neural Information Processing Systems*, 36:71095–71134, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

Junchao Yi, Rui Zhao, Jiahao Tang, Weixian Lei, Linjie Li, Qisheng Su, Zhengyuan Yang, Lijuan Wang, Xiaofeng Zhu, and Alex Jinpeng Wang. Flowinone: Unifying multimodal generation as image-in, image-out flow matching. *arXiv preprint arXiv:2604.06757*, 2026.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6613–6623. IEEE, 2024.

Wei Zhou, Xiongwei Zhu, Zelin Xu, Bo Dong, Lixue Gong, Yongyuan Liang, Meng Chu, Leigang Qu, Lingdong Kong, Wei Liu, et al. Danceopd: On-policy generative field distillation. *arXiv preprint arXiv:2606.27377*, 2026.

A Appendix

The supplementary material is organized as follows. Sec. B provides implementation and training details for the shared latent interface and cross-model OPD; Sec. C reviews additional related work; Sec. D reports extended quantitative results, including additional benchmarks, efficiency analysis, and rollout scheduling ablations; Sec. E presents additional qualitative examples and handoffs at different denoising steps; and Sec. F discusses the limitations of our method.

B Training details

Tables 7 and 8 summarize the training configurations. The first stage optimizes only the model-specific LAEs, while cross-model OPD freezes the teacher, reference model, and latent interface and updates only the student LoRA.

Table 7: **Training hyperparameters for the shared latent interface.**

Hyperparameter	Setting
Architecture	ResBlock LAE
Latent channels	16 native; 32 shared
Hidden channels	64
ResBlocks depth	3
σ embedding size	128
Trainable parameters	1.532M
Optimizer	AdamW
Learning rate	10^{-4}
Weight decay	0.01
Batch size	1 per GPU; accumulation 4; effective batch size 32
Precision	BF16
GPUs	8
Time sampling	1,000 timesteps; exponential shift $\mu = \ln 3$
Optimization	LAEs only

Table 8: **Training hyperparameters for cross-model OPD.**

Hyperparameter	Setting
Sampling steps	20
Supervision steps	1, 2, 4, 8
Optimizer	AdamW
Learning rate	10^{-5}
Weight decay	0.01
Batch size	1 per GPU; accumulation 2; effective batch size 16
Precision	BF16
GPUs	8
LoRA configuration	Rank 64; alpha 64; Kaiming initialization
LoRA targets	Attention QKV/output and FFN projections
Reference loss weight	0.1
Training length	Up to 50 epochs
Optimization	SD3 LoRA only

C Additional Related Work

Efficient Generation. Existing acceleration methods reduce either the number or the computational cost of denoising evaluations. Few-step methods compress long denoising trajectories into one or a few evaluations through distillation or consistency learning (Salimans & Ho, 2022; Song et al., 2023; Yin et al., 2024; Luo et al., 2023), while caching-based methods reduce per-step computation by reusing or forecasting intermediate features across adjacent timesteps (Liu et al., 2025a; Bu et al., 2026; Liu et al., 2025b). Model scheduling methods such as OMS-DPM and T-Stitch further assign different pretrained models to different denoising stages (Liu et al., 2023; Pan et al., 2025). These studies demonstrate the potential of multi-model acceleration, but have primarily considered models whose intermediate states are directly interchangeable. Composing independently developed generators with heterogeneous latent spaces remains largely unexplored.

Diffusion On-Policy Distillation. On-policy distillation (OPD) has proven effective in large language models (Gu et al., 2023; Agarwal et al., 2024). Its extensions to diffusion and flow models follow two main directions: on-policy self-distillation uses privileged information to construct a stronger teacher from the student itself (Jiang et al., 2026), while teacher-based OPD transfers denoising capabilities from stronger pretrained models along student trajectories (Li et al., 2026; Fang et al., 2026; Zhou et al., 2026). Most methods, however, either assume that teacher and student belong to the same model family or bridge heterogeneous models by comparing their decoded outputs in a frozen visual representation (Fu et al., 2026a;b). How to directly transfer a teacher's denoising dynamics to a student with an incompatible latent space therefore remains an open question.

Multi-Preference Composition. Multi-preference generation is commonly formulated as learning a single model that balances several objectives. Prior methods either jointly optimize one generator with multiple reward signals or separately fine-tune preference-specific models and merge their parameters (Lee et al., 2025; Rame et al., 2023). The former encodes preference trade-offs through multi-objective training, whereas the latter relies on compatible architectures and weight-space interpolation. In both cases, preference composition remains tied to model training or parameter manipulation, limiting the flexible reuse of independently trained experts.

D Extended Quantitative Results

D.1 Ablation on Three-Stage Rollout Scheduling

To verify whether the shared latent interface remains stable under multi-stage model cascading, we evaluate all six permutation schedules among SD3, FLUX.1, and Qwen-Image across a three-stage generation pipeline (Tab. 9). Across GenEval-v2 and HPSv2 metrics, the generation quality remains highly consistent regardless of the specific model execution order.

Table 9: **Robustness of the shared latent interface across multi-stage rollout schedules.** We evaluate different execution orders across three heterogeneous generators (SD3, FLUX.1, and Qwen-Image). The consistent performance across arbitrary transition permutations confirms the stability and invariance of the joint shared latent space under cascading cross-model transfers.

Rollout Schedule	GenEval-v2	HPSv2				
		Anime	Concept	Paint.	Photo	Avg.
SD3 → FLUX.1 → Qwen	19.99	27.04	25.41	25.46	24.69	25.65
SD3 → Qwen → FLUX.1	24.93	24.35	22.80	22.72	21.91	22.95
FLUX.1 → SD3 → Qwen	25.22	27.86	25.41	25.30	24.87	25.86
FLUX.1 → Qwen → SD3	27.96	25.68	22.85	22.96	22.68	23.54
Qwen → SD3 → FLUX.1	30.76	26.60	24.14	24.23	24.06	24.76
Qwen → FLUX.1 → SD3	33.13	27.11	24.62	24.98	23.71	25.11

D.2 Per-Model Reconstruction Analysis Beyond Average Metrics

While Table 1a summarizes the average reconstruction fidelity across generators, Table 10 provides a granular per-model breakdown—reporting PSNR, SSIM, and LPIPS across 1, 5, and 10 latent round trips. This detailed evaluation highlights how reconstruction errors accumulate over sequential cross-model translations, confirming the stability limits of the shared latent space.

D.3 Inference Efficiency Analysis

To determine whether the shared latent interface introduces measurable inference overhead, we compare the end-to-end latency of individual generators and LAE-enabled hybrid generation. Two-model hybrids allocate the steps evenly, while the three-model hybrid uses a $17/17/16$ allocation. After three warm-up runs, each configuration is measured over 20 runs, with the results reported in Table 11. Since each handoff only translates the latent state without invoking an additional denoiser evaluation, InstantFusion introduces no extra theoretical NFE.

E Extended Qualitative Results

E.1 Handoffs at Different Denoising Steps

To evaluate the trajectory-wide alignment of the proposed LAE across the generation process, we conduct cross-model handoffs at arbitrary intermediate timesteps. As depicted in Figs. 11 and 12, transitions between source and target generators execute reliably at virtually any stage of denoising.

Table 10: Reconstruction fidelity under different shared-space configurations.

Shared Space	Reconstructed Model	Switch	PSNR ↑	SSIM ↑	LPIPS ↓
TaylorSeer ($\mathcal{N} = 6$)	FLUX.1	–	28.945	0.6556	0.4020
SD3–Qwen	Qwen	1	32.418	0.9433	0.03811
		5	24.653	0.8719	0.10096
		10	22.255	0.8253	0.13709
	SD3	1	30.331	0.9401	0.04244
		5	21.301	0.8143	0.15164
		10	17.830	0.7086	0.24075
	Average	1	31.375	0.9417	0.04028
		5	22.977	0.8431	0.12630
		10	20.043	0.7670	0.18892
FLUX.1–Qwen	FLUX.1	1	33.253	0.9575	0.02897
		5	21.512	0.8326	0.13814
		10	16.886	0.7171	0.24954
	Qwen	1	31.904	0.9416	0.04428
		5	23.813	0.8573	0.12314
		10	21.325	0.8070	0.17074
	Average	1	32.579	0.9496	0.03663
		5	22.663	0.8450	0.13064
		10	19.106	0.7621	0.21014
SD3–FLUX.1–Qwen	FLUX.1	1	34.159	0.9639	0.02168
		5	23.799	0.8657	0.10111
		10	19.571	0.7814	0.17821
	Qwen	1	32.276	0.9441	0.03814
		5	23.982	0.8670	0.10392
		10	21.459	0.8191	0.14158
	SD3	1	30.757	0.9453	0.03788
		5	22.553	0.8412	0.12516
		10	19.328	0.7418	0.20794
	Average	1	32.397	0.9511	0.03257
		5	23.445	0.8580	0.11006
		10	20.119	0.7808	0.17591

Table 11: **End-to-end inference latency and theoretical NFEs.** Latency is measured at 1024×1024 resolution, averaged over 20 runs, and reported in seconds per image. For hybrid configurations, the NFE entries show how the 50 denoising evaluations are allocated across models.

Metric	Individual Generators			Hybrid Generation			
	SD3	FLUX.1	Qwen	FLUX.1→SD3	SD3→Qwen	FLUX.1→Qwen	FLUX.1→SD3→Qwen
Latency (s/image)	3.61	10.70	20.82	7.16	12.21	15.7	11.53
Theoretical NFE	50	50	50	$25 + 25$	$25 + 25$	$25 + 25$	$17 + 17 + 16$

This flexibility confirms that our LAE maintains robust state and dynamics alignment throughout the entire trajectory, rather than being restricted to specific timestep windows.



Figure 11: Cross-model state handoffs at different denoising timesteps (Qwen-Image, Flux.1).



Figure 12: Cross-model state handoffs at different denoising timesteps (Qwen-Image, SD3).

E.2 Additional Results across Applications

To further evaluate the versatility and qualitative fidelity of InstantFusion, this section provides extended visual comparisons across diverse downstream configurations. Specifically, Figs. 13, 14, and 15 depict single-trip and multi-trip latent reconstructions across SD3, FLUX.1, and Qwen-Image, verifying that our shared representation consistently preserves spatial structures across distinct autoencoder designs. Next, Figs. 16 and 17 showcase cross-architecture handoffs from Qwen-Image into SD3 and FLUX.1 respectively, highlighting robust state migration between heterogeneous model families. Fig. 18 illustrates the synergistic acceleration. Finally, Figs. 19 and 20 demonstrate multi-preference composition results across distinct reward alignments—combining instruction adherence (GenEval-v2) with aesthetic preference (HPSv2), and coupling multi-domain aesthetic metrics (HPSv2 and HPSv3), respectively.

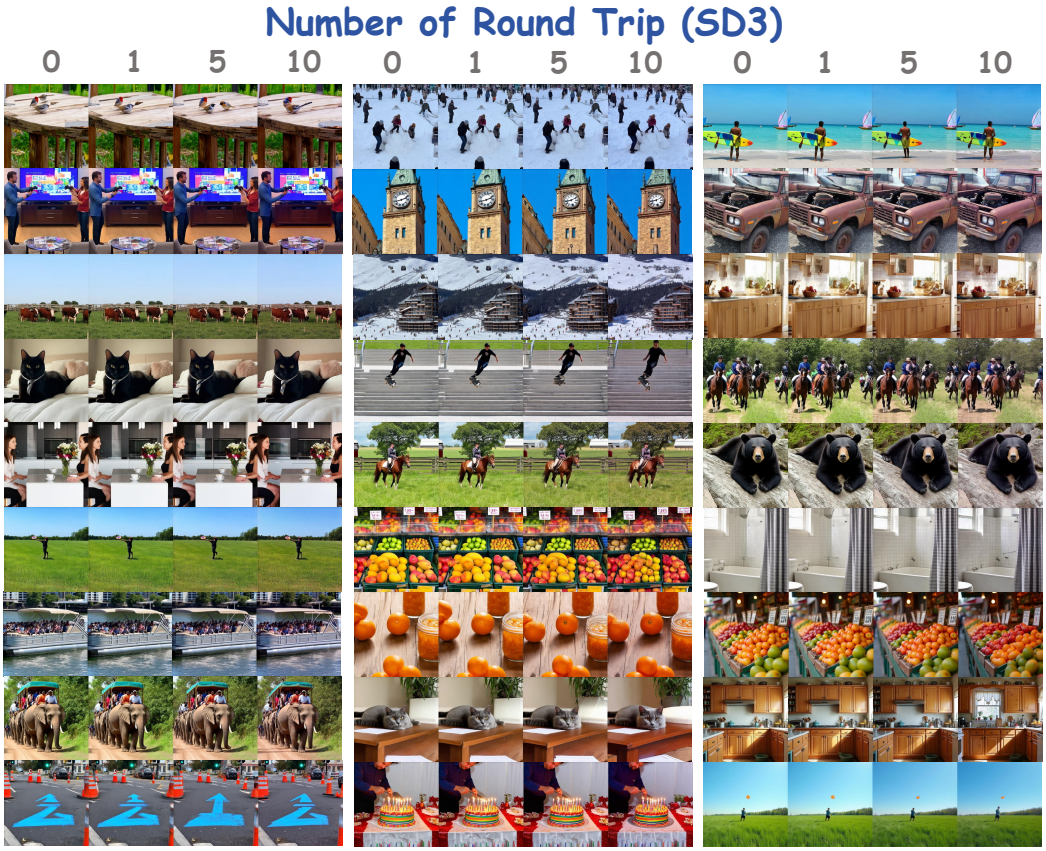

Figure 13: Qualitative comparison of shared latent reconstruction (SD3).

F Limitations

While InstantFusion provides a lightweight interface for cross-architecture state and dynamics translation without pair-specific mappings, several limitations remain. Currently, our formulation relies on a shared autoencoder and compatible time parameterization across generators. Furthermore, executing hybrid inference requires hosting multiple frozen backbones simultaneously, incurring elevated memory and storage overhead. Extending this interface across heterogeneous autoencoders and optimizing dynamic model loading/offloading represent key directions for future research.



Figure 14: Qualitative comparison of shared latent reconstruction (FLUX.1).



Figure 15: Qualitative comparison of shared latent reconstruction (Qwen-Image).



Figure 16: Cross-architecture state handoffs initiated from Qwen-Image to SD3.

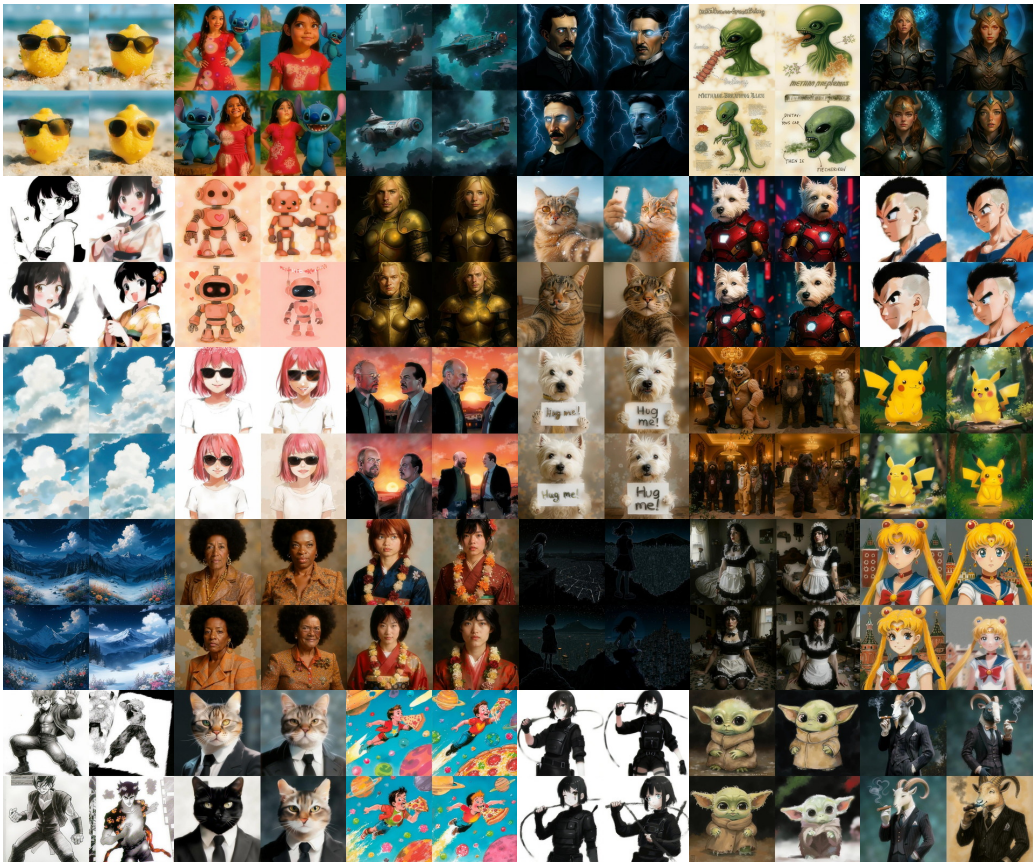

Figure 17: Cross-architecture state handoffs initiated from Qwen-Image to FLUX.1.



Figure 18: Qualitative results under generation acceleration.



Figure 19: Multi-preference composition cases combining GenEval-v2 and HPSv2.



Figure 20: Multi-preference composition cases combining HPSv2 and HPSv2=3.