NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Does FLUX Already Know How to Perform Physically Plausible Image Composition?

ICLR 2026 Presentation

Shilin Lu[1,*], Zhuming Lian[1,*], Zihan Zhou[1],
Shaocong Zhang[1], Chen Zhao[2], Adams Wai-Kin Kong[1]

[1]Nanyang Technological University, [2]Nanjing University
*Equal Contribution

(Credit to Zhuming Lian)

Background

Image Composition

Goal: Insert a reference object into a target scene while maintaining realism.

Challenges:

- Complex lighting (shadows, reflections, backlighting)
- Handling different resolutions
- Maintaining object identity and background consistency

Recent diffusion models such as **FLUX** have strong generative priors.
Key question:

Can these priors already support realistic image composition?

Motivation

Limitations of Existing Methods

Training-free methods mainly rely on:

- **Image inversion:** forces object pose from reference image and limits flexibility in new scenes
- **Attention manipulation:** unstable and sensitive to hyperparameters

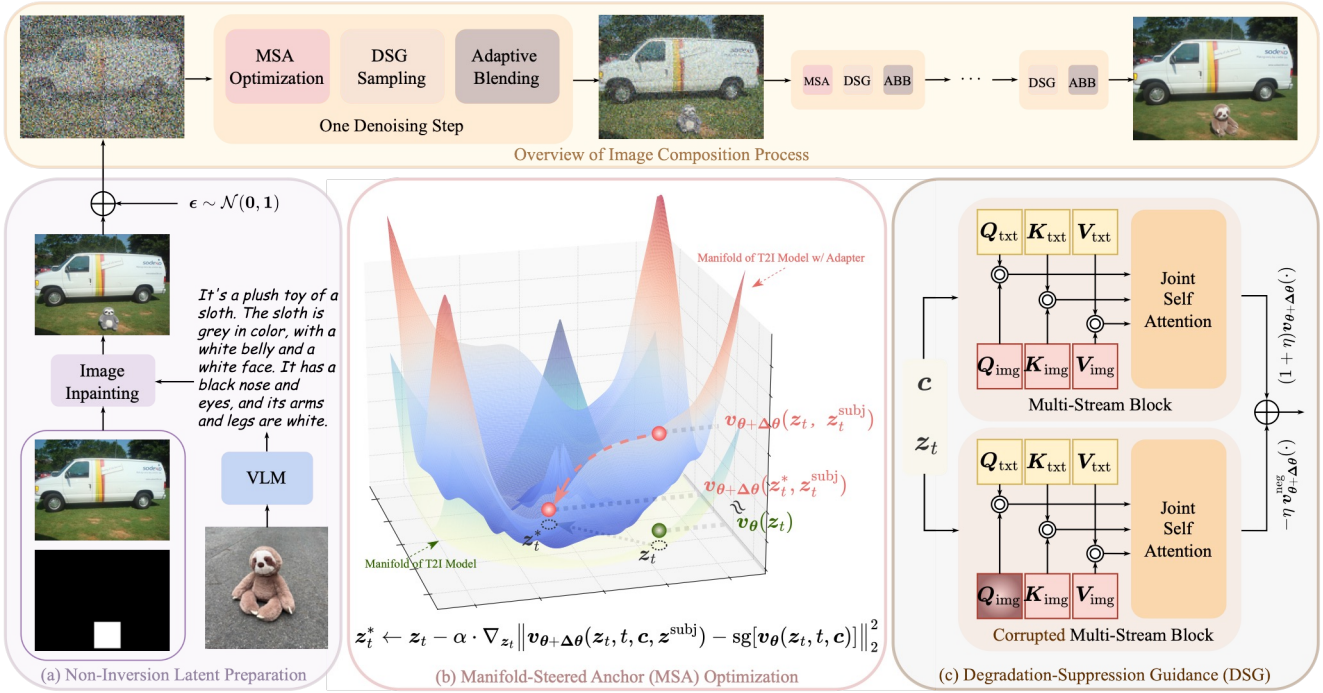
Therefore, we propose a **training-free composition framework – SHINE** (Seamless High-fidelity Insertion with Neutralized Errors)

Method Overview

1 Manifold-Steered Anchor (MSA) (Figure (b))

Align predictions from the base diffusion model and the customization adapter.

Goal: preserve identity while keeping scene structure.



Method Overview

2 Degradation-Suppression Guidance (DSG)

Construct negative guidance from degraded outputs.

Observation: Blurring image query features in self-attention produces degraded predictions.
We use a **CFG-like method** to guide the model **away from these directions**.

$$\boldsymbol{v}_t^{\text{dsg}} = \boldsymbol{v}_{\boldsymbol{\theta}+\Delta\boldsymbol{\theta}}(\boldsymbol{z}_t, t, \boldsymbol{c}, \boldsymbol{z}^{\text{subj}}) + \eta\big(\boldsymbol{v}_{\boldsymbol{\theta}+\Delta\boldsymbol{\theta}}(\boldsymbol{z}_t, t, \boldsymbol{c}, \boldsymbol{z}^{\text{subj}}) - \boldsymbol{v}^{\text{neg}}_{\boldsymbol{\theta}+\Delta\boldsymbol{\theta}}(\boldsymbol{z}_t, t, \boldsymbol{c}, \boldsymbol{z}^{\text{subj}})\big)$$

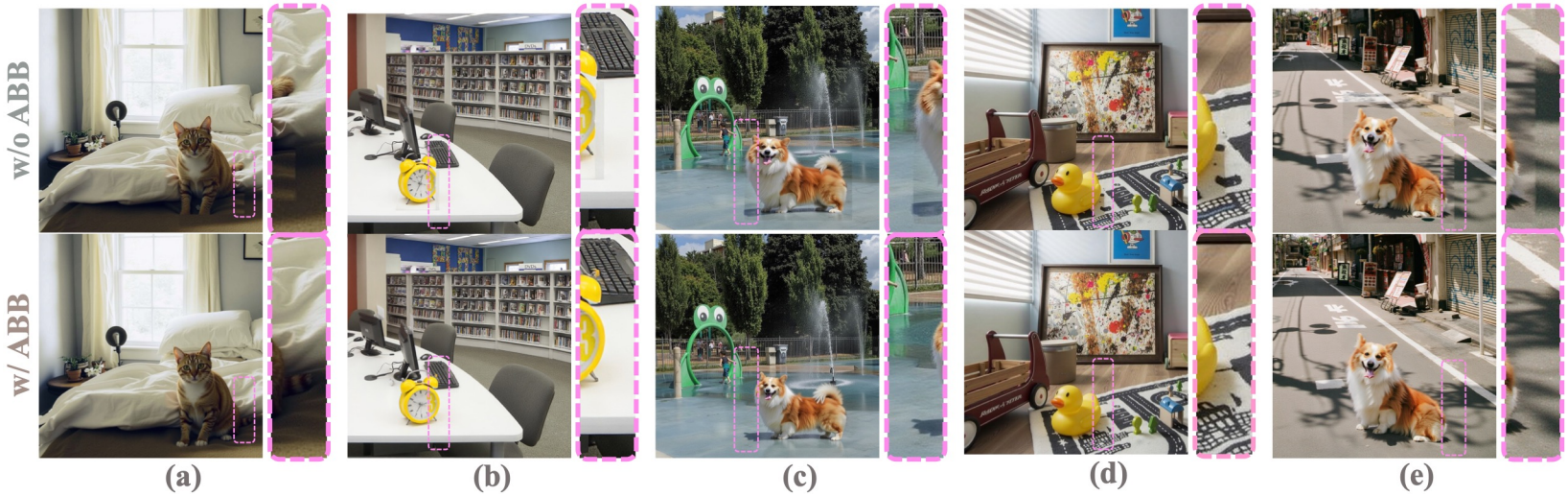
Method Overview

3 Adaptive Background Blending (ABB)

Generate semantic masks from attention maps for seamless blending.

$$z_t' = \hat{M} \odot z_t + \left(1 - \hat{M}\right) \odot z_t^{\text{bg}}, \qquad \hat{M} = \mathbb{1}\{t > \tau\}\ \mathcal{D}(M^{\text{attn}}) + \mathbb{1}\{t \le \tau\}\ M^{\text{user}}$$

Goal: get smoother object boundaries and better scene integration



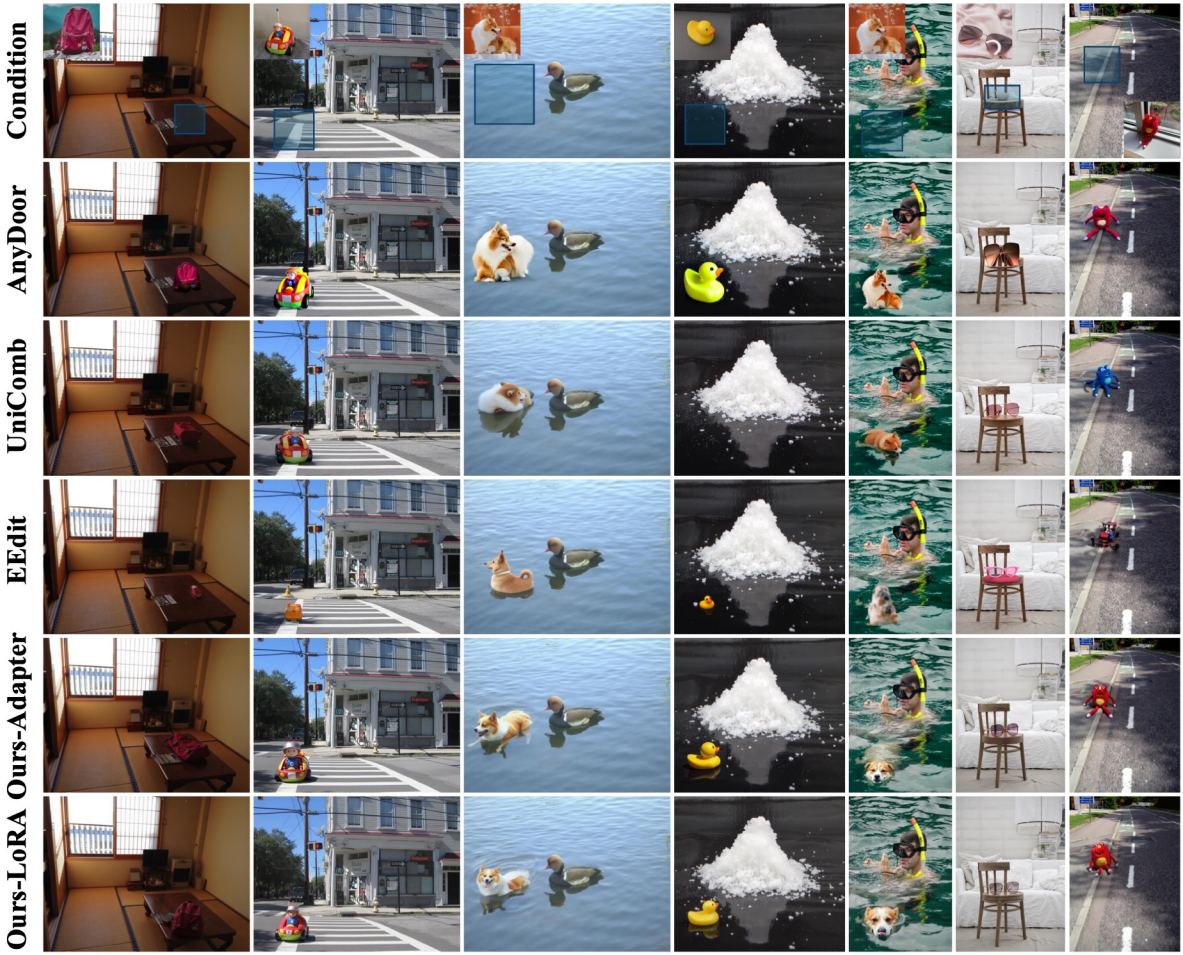
Experiments

Qualitative Results

ComplexCompo Benchmark

Features:

- multiple image resolutions
- challenging lighting conditions



Experiments

Quantitative Results

Table 1: Comparison of composition performance across two benchmarks. The best result in each column is highlighted in **bold**, while the second-best is underlined. Metrics shown in pink are those specifically trained to better align with human preferences. Abbreviations: IRF= Instance Retrieval Features; IR = ImageReward; VR = VisionReward; URE = UnifiedReward-Edit-qwen3vl-8b.

Bench	Method	Training -Free	Base Model	External Model	Subject Identity Consistency				Background		Image Quality			
					CLIP-I ↑	DINOv2 ↑	IRF ↑	DreamSim ↓	LPIPS ↓	SSIM ↑	IR ↑	VR ↑	HPS ↑	URE ↑
Dream-Edit-Bench (220)	Flux.1 Fill (Black Forest Labs, 2024b)	✘	FLUX	-	0.7328	0.6745	0.5754	0.5233	0.0166	0.9076	0.5577	3.5997	8.6432	21.5812
	MADD (He et al., 2024)	✘	SD	DINO	0.7118	0.6279	0.4333	0.5810	0.0604	0.8182	-0.2545	2.7011	1.2443	13.8148
	ObjectStitch (Song et al., 2023)	✘	SD	VIT	0.7567	0.6930	0.5525	0.5093	0.0190	0.8316	0.0791	3.2416	7.4529	19.1886
	DreamCom (Lu et al., 2023c)	✘	SD	LoRA	0.7414	0.6749	0.5597	0.5626	0.0200	0.8283	0.1873	3.5053	5.9324	19.9296
	AnyDoor (Chen et al., 2024c)	✘	SD	DINO	**0.8183**	0.7283	0.7714	0.3764	0.0251	0.8894	0.4511	3.3946	8.4867	19.0989
	UniCombine (Wang et al., 2025a)	✘	FLUX	LoRA	0.8058	0.7332	0.7579	0.3984	0.0050	0.9397	0.4565	3.6108	8.8415	21.7080
	PBE (Yang et al., 2023)	✘	SD	-	0.7742	0.7040	0.5845	0.4985	0.0197	0.8287	0.2083	3.3482	8.3789	20.2137
	TIGIC (Li et al., 2024b)	✔	SD	-	0.7226	0.6718	0.4711	0.6108	0.0584	0.8153	-0.1332	2.9873	5.2676	17.1000
	TALE (Pham et al., 2024)	✔	SD	-	0.7329	0.6604	0.5007	0.6176	0.0392	0.8251	-0.1502	3.1349	6.3773	18.0784
	TF-ICON (Lu et al., 2023d)	✔	SD	-	0.7479	0.6865	0.5179	0.5441	0.0582	0.8111	0.0816	3.2823	7.2643	18.2716
	DreamEdit (Li et al., 2023b)	✔	SD	LoRA, VIT	0.7703	0.7151	0.6147	0.5047	0.0140	**0.9775**	0.1744	3.1775	6.0250	15.7636
	EEdit (Yan et al., 2025)	✔	FLUX	-	0.6998	0.6590	0.4438	0.6160	**0.0039**	0.9475	0.0216	3.3606	6.6689	19.5603
	Ours-Adapter	✔	FLUX	Adapter	0.8086	0.7415	0.7702	0.3730	0.0236	0.8959	0.5709	**3.6234**	**8.8861**	**22.0182**
	Ours-LoRA	✔	FLUX	LoRA	0.8125	**0.7452**	**0.7900**	**0.3577**	0.0271	0.8847	**0.5906**	3.6161	8.8688	21.9421
Complex-Compo (300)	Flux.1 Fill (Black Forest Labs, 2024b)	✘	FLUX	-	0.7108	0.6475	0.5466	0.6018	0.0232	0.7442	0.4088	3.5737	8.7376	19.7712
	MADD (He et al., 2024)	✘	SD	DINO	0.6780	0.5993	0.3638	0.5979	0.0781	0.5658	-0.0088	2.6582	5.9673	13.0567
	ObjectStitch (Song et al., 2023)	✘	SD	VIT	0.7608	0.7077	0.5513	0.4717	0.0388	0.6357	0.2482	3.4411	8.8389	18.8283
	DreamCom (Lu et al., 2023c)	✘	SD	LoRA	0.648	0.5692	0.2788	0.8192	0.0389	0.6342	-0.0778	3.4409	7.9884	18.6143
	AnyDoor (Chen et al., 2024c)	✘	SD	DINO	0.7982	0.7052	0.7319	0.4493	0.0299	0.7262	0.3804	3.3787	8.9760	18.3550
	UniCombine (Wang et al., 2025a)	✘	FLUX	LoRA	0.7361	0.6552	0.5380	0.5682	0.0237	0.7077	0.2470	3.5454	8.8999	19.8529
	PBE (Yang et al., 2023)	✘	SD	-	0.7537	0.6802	0.5189	0.5187	0.0397	0.6321	0.2139	3.4310	8.5923	18.9507
	TIGIC (Li et al., 2024b)	✔	SD	-	0.6913	0.6329	0.3848	0.6549	0.0929	0.6228	-0.131	2.8898	7.6630	16.4301
	TALE (Pham et al., 2024)	✔	SD	-	0.6816	0.6151	0.3799	0.6773	0.059	0.6334	0.0783	3.4498	8.7351	18.7567
	TF-ICON (Lu et al., 2023d)	✔	SD	-	0.6987	0.6435	0.4167	0.6030	0.0815	0.6216	0.1798	3.4323	9.3258	18.2366
	DreamEdit (Li et al., 2023b)	✔	SD	LoRA, VIT	0.7314	0.6722	0.5069	0.5670	0.0468	0.7201	0.1212	3.2531	8.0434	15.8934
	EEdit (Yan et al., 2025)	✔	FLUX	-	0.6713	0.6153	0.3797	0.6821	**0.0226**	0.7107	0.1433	3.5009	8.7835	19.7348
	Ours-Adapter	✔	FLUX	Adapter	0.7721	0.7107	0.6764	0.4294	0.0404	**0.7789**	0.4090	**3.6020**	9.6485	20.7349
	Ours-LoRA	✔	FLUX	LoRA	**0.7999**	**0.7384**	**0.7659**	**0.3542**	0.0430	0.7634	**0.4246**	3.5951	**9.8418**	**21.0326**

Conclusion

We propose **SHINE**, a training-free framework for realistic image composition.

Key contributions:

- Manifold-Steered Anchor Loss -> **model-agnostic** bridging between pretrained diffusion models and adapter models (image customization)
- Degradation-Suppression Guidance -> a **replacement for negative CFG** (image generation)
- Adaptive Background Blending -> applicable to all mask-based inpainting

Our results show that modern diffusion models like **FLUX already contain strong composition priors** and proper guidance can effectively unlock these capabilities.