PIT: PRESERVING IDENTITY THROUGH TRAJECTORIES FOR LAYOUT-AWARE MULTI-REFERENCE IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Layout-aware multi-reference image generation aims to place multiple reference subjects in prescribed regions while preserving their identities and adapting them to a coherent text-specified scene. Recent methods strengthen reference-to-region binding through layout-conditioned inputs or structured attention, typically relying on a single evolving latent trajectory for both identity preservation and scene adaptation. However, reference-specific evidence may become less recoverable from the evolving scene-adaptive state, even though reference conditioning remains available at every step. Once identity deviations arise, they may persist across subsequent steps, while no alternative identity history is retained for correction. To address this limitation, we propose PIT, a framework that maintains a persistent identity trajectory alongside a primary trajectory. Both trajectories share the same text and reference conditions but evolve under different guidance rules. Accordingly, the primary trajectory adapts the generation to the target scene, whereas the identity trajectory maintains an alternative state history with greater emphasis on reference guidance. To exploit this alternative history, PIT compares the two trajectory states at the same timestep and uses the difference between their clean predictions to correct the primary trajectory within the estimated subject regions. Extensive experiments show that PIT achieves a strong overall balance across identity preservation, spatial accuracy, and text consistency.

1 INTRODUCTION

Image generation offers a flexible way to create visual content from textual descriptions. For applications such as product display and storybook illustration, creators also need control over which specific subjects appear and how they are arranged in the scene (Hu et al., 2026; Xu et al., 2026). Reference images specify the subjects' identities, spatial layouts define their placement and scale, and text describes the intended scene and interactions. Layout-aware multi-reference image generation combines these complementary inputs to compose multiple reference subjects into a new scene, as shown in Fig. 1(a). The task requires preserving each subject's distinctive identity and prescribed placement while adapting its pose, geometry, lighting, and interactions to form a coherent, prompt-aligned scene.

Recent methods have made substantial progress in reference-to-region binding, namely, associating each reference subject with its intended spatial location. PositionIC (Hu et al., 2026) introduces position-annotated data synthesis and visibility-aware attention to improve spatial control and identity consistency in multi-subject customization. ContextGen (Xu et al., 2026) incorporates a composite layout image for contextual layout anchoring and introduces identity consistency attention to associate target regions with their corresponding reference subjects. These advances strengthen spatial correspondence, but identity preservation and scene adaptation still rely on a single-trajectory process. In a single-trajectory process, one evolving latent state must accommodate both reference-specific appearance and scene-dependent changes in pose, geometry, lighting, and interactions. Although reference conditioning remains active throughout sampling, we observe that a generated subject can remain spatially and semantically plausible while becoming less faithful to its reference, as illustrated in Fig. 1(b). Once such a deviation occurs, subsequent predictions still proceed from the same evolving state, while no alternative identity history is retained to support correction. We refer

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.

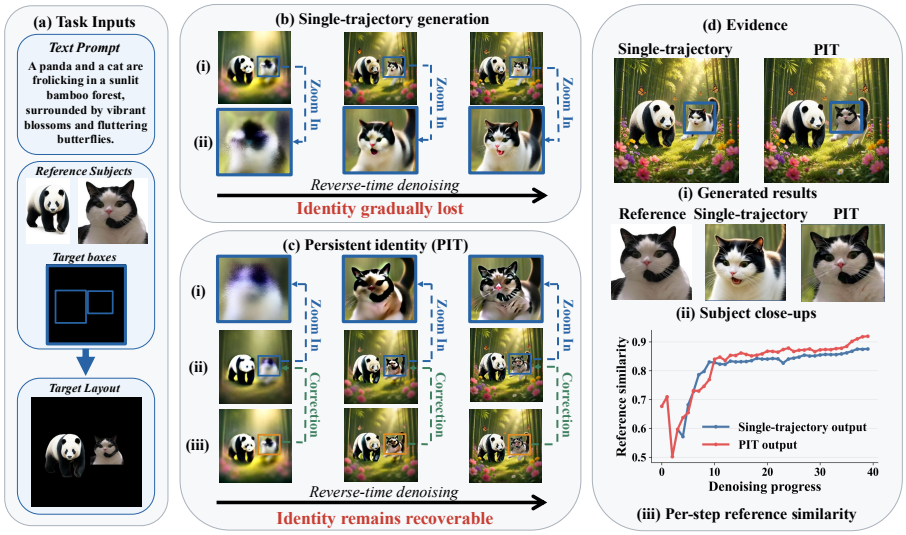

Figure 1: **Motivation and overview of PIT.** (a) Task inputs. (b) Single-trajectory generation: (i) intermediate trajectory snapshots over time, and (ii) zoomed-in evidence showing identity degradation. (c) PIT: (i) retained reference details, (ii) the primary trajectory, and (iii) the persistent identity trajectory. (d) Evidence: (i) final generated results, (ii) subject close-ups, and (iii) per-step reference similarity curves between the reference subject and the generated subject region for single-trajectory generation and PIT. The single-trajectory process gradually loses reference identity, whereas PIT preserves an alternative identity history and uses it to recover subject identity during generation.

to the challenge of keeping reference-specific information accessible throughout scene adaptation as identity persistence.

Motivated by this analysis, we propose PIT (Preserving Identity through Trajectories), which maintains a persistent identity trajectory alongside the primary trajectory, as illustrated in Fig. 1(c). Both trajectories share the same prompt and reference conditions but evolve under different guidance rules. The primary trajectory adapts the generation to the target scene, whereas the identity trajectory places greater emphasis on reference guidance, thereby preserving an alternative state history for identity correction. At each denoising step, PIT compares the two trajectory states at the same timestep and uses the difference between their clean predictions to derive a history-dependent correction. The correction is applied only within the estimated subject regions of the primary trajectory, so that identity information can be recovered without unnecessarily altering the surrounding scene. As illustrated in Fig. 1(d), this mechanism allows PIT to preserve reference identity while the primary trajectory continues to adapt to the target scene. Experiments on layout-aware multi-reference benchmarks show that PIT improves reference identity preservation and spatial accuracy while maintaining strong prompt consistency.

Our contributions are threefold:

- We formulate identity persistence as a trajectory-level challenge in layout-aware multi-reference generation: although reference conditions remain available throughout sampling, single-trajectory inference does not retain an alternative identity history that can support correction once the evolving state deviates from the reference.
- We introduce PIT (Preserving Identity through Trajectories), which maintains a persistent identity trajectory alongside a scene-adaptive primary trajectory. By comparing the two trajectory states at the same timestep, PIT derives a history-dependent correction from their clean-prediction difference and applies it within the estimated subject regions.
- We validate PIT through extensive comparisons with external methods and controlled single-trajectory baselines. The results show that maintaining an alternative identity history consistently improves reference identity preservation and spatial accuracy, while retaining strong text consistency and visual quality across different inference settings.

2 Related Work

Personalized and layout-aware multi-subject generation. Personalized generation represents user-provided concepts through optimized embeddings, subject-specific updates, or feed-forward image adapters (Gal et al., 2023; Ruiz et al., 2023; Kumari et al., 2023; Ye et al., 2023). Multi-subject methods reduce identity mixing and attribute leakage through localized attention, concept-specific representations, feature modulation, or dedicated data and objectives (Liu et al., 2023; Xiao et al., 2025; Ma et al., 2024; Jang et al., 2024; Huang et al., 2025; Wang et al., 2025b; Chen et al., 2025; She et al., 2026; Wang et al., 2026; Phung et al., 2026), while spatial-control methods control object placement using boxes, masks, and instance-level conditions through conditioned attention, instance controllers, or consensus sampling (Li et al., 2023; Xie et al., 2023; Zhou et al., 2024; Wang et al., 2024; Bar-Tal et al., 2023). Recent layout-aware multi-reference systems combine these capabilities through explicit layout semantics, region-restricted attention, structured multimodal inputs, and position-supervised training (Peng et al., 2025a; Chen et al., 2026; Hu et al., 2026; Xu et al., 2026; Saha et al., 2026; Shi et al., 2026; Xiang et al., 2025; Yang et al., 2026; Mikaeili et al., 2025). Structured conditioning establishes reference-to-region correspondence and spatial control, whereas PIT addresses the complementary trajectory-level challenge of keeping reference identity accessible after correct binding is established. We refer to this remaining problem as identity persistence.

Prediction composition and parallel diffusion processes. Combining predictions and coordinating multiple diffusion processes have been explored across diverse generation and editing tasks. Compositional generation methods combine condition-specific predictions or guidance signals (Liu et al., 2022; Ho & Salimans, 2022; Brooks et al., 2023; Bansal et al., 2024; Epstein et al., 2023; Mo et al., 2024), while related approaches coordinate spatial regions, multiple views, or different conditions through synchronized sampling and prediction composition (Bar-Tal et al., 2023; Lee et al., 2023; Geng et al., 2024; Huang et al., 2023; Wang et al., 2025a; Wan et al., 2026). Editing and customization methods further exchange features or exploit differences between source and target predictions or trajectories (Tumanyan et al., 2023; Cao et al., 2023; Ju et al., 2024; Xu et al., 2024; Yang et al., 2025; Jiang & Chen, 2025; Hertz et al., 2023; Koo et al., 2024; Kulikov et al., 2025). PIT differs in the role assigned to parallel trajectories: its auxiliary trajectory preserves an alternative identity history for the same generation process, rather than representing a different region, view, condition, or source–target state. By comparing the two trajectory states under matched conditions and timestep, PIT derives a correction that reflects their distinct state histories.

3 Method

3.1 Problem Setup and Guidance Formulation

Given a text prompt p, reference subjects $\mathcal{R} = \{r_n\}_{n=1}^{N}$, and a target layout $\mathcal{L}$, our goal is to synthesize an image that places each subject in its designated region while preserving its identity and adapting it to the prompt-specified scene. We instantiate the image condition $\mathbf{c}_I$ using a CanvasComposer-style pruned layered representation (Qian et al., 2026); implementation details are provided in Appendix A.1. We adapt the backbone to the layered multi-reference input using standard LoRA training under its original flow-matching objective. PIT combines this task adaptation with the dual-trajectory generation procedure described below.

We write $F_{ab}(\mathbf{x}_t) = f_\theta(\mathbf{x}_t, t; p^{(a)}, \mathbf{c}_I^{(b)})$, where $a, b \in \{0, 1\}$ indicate whether the text and image conditions are enabled, respectively, and 0 denotes the corresponding null condition. We follow the backbone's linear flow parameterization, where σ_t denotes the noise level at timestep t. For a latent state $\mathbf{x}$ and predicted velocity $\mathbf{v}$, the corresponding clean estimate is

$$\mathcal{C}_t(\mathbf{x}, \mathbf{v}) = \mathbf{x} - \sigma_t \mathbf{v}. \tag{1}$$

We define the primary velocity by combining text guidance with spatially localized image guidance:

$$\mathbf{v}_t^P = F_{11}(\mathbf{x}_t^P) + s_T D_T(\mathbf{x}_t^P) + s_I \bar{D}_{I,t}(\mathbf{x}_t^P), \tag{2}$$

Here, s_T and s_I are the text and image guidance scales. Because the image condition consists of subject-specific layers (Appendix A.1), we use $\bar{D}_{I,t}(\mathbf{x}) = \mathbf{M}_t \odot D_I(\mathbf{x})$ to restrict the additional image-guidance residual to their target regions. The soft mask $\mathbf{M}_t \in [0, 1]^{h \times w}$ is initialized from

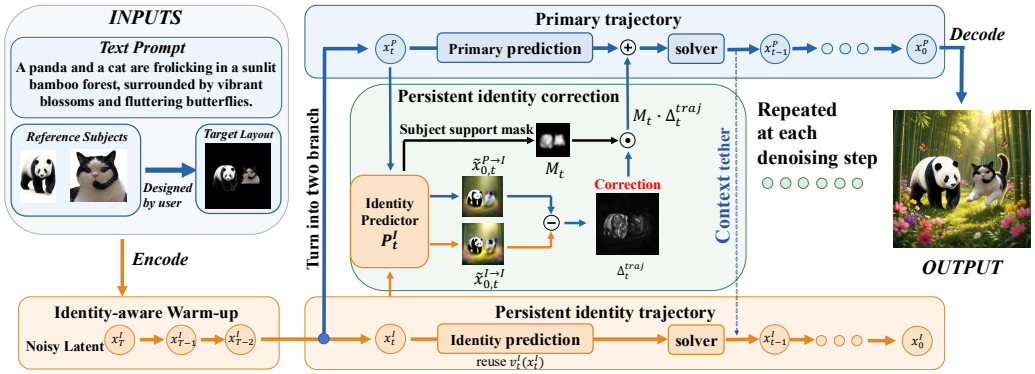

Figure 2: **Overview of PIT.** After an identity-aware warm-up, PIT evolves a scene-adaptive primary trajectory together with a persistent identity trajectory. At each step, the identity predictor $\mathcal{P}_t^I$ is applied to both states, and the difference between their clean-latent predictions defines a history-dependent correction. The masked correction is added to the primary clean prediction and converted to the corrected velocity $\mathbf{v}_t^{\mathrm{PIT}}$. The solver then advances the primary and identity trajectories with $\mathbf{v}_t^{\mathrm{PIT}}$ and $\mathbf{v}_t^I$, respectively. Finally, the dashed tether blends the updated states using the subject mask, preserving the identity trajectory within subject regions while aligning its surrounding context with the primary trajectory.

the user-specified layout and refined during sampling; it is broadcast over latent channels, and $\odot$ denotes element-wise multiplication. Its construction is detailed in Appendix A.2.

In this formulation, both text and reference guidance are evaluated at the same evolving primary state. Although reference guidance remains available at every step, subsequent predictions continue to depend on this single state as it evolves throughout generation. Once identity deviations arise, single-trajectory inference retains no alternative identity history that can support correction. PIT addresses this state-dependent limitation by maintaining a persistent identity trajectory and using its alternative state history to provide regional corrections to the primary trajectory. Figure 2 provides an overview of this process.

3.2 Persistent Subject-Local Identity Trajectory

The persistent identity trajectory is designed to retain an alternative state history that emphasizes reference guidance while remaining compatible with the evolving scene context. We therefore evolve an auxiliary state $\mathbf{x}_t^I$ under the identity velocity:

$$\mathbf{v}_t^I(\mathbf{x}) = F_{11}(\mathbf{x}) + s_I D_I(\mathbf{x}). \tag{3}$$

This velocity retains the jointly conditioned prediction F_{11} and adds the reference-guidance residual, without the additional text-guidance residual used by the primary velocity.

Because the subsequent correction compares two trajectories with different roles, we first establish a shared reference-aware state before separating them. Specifically, we advance a single state under Eq. 3 for the first K sampling steps, allowing both trajectories to inherit the same identity-aware subject structure. The shared state is then split into two trajectories: the primary trajectory uses Eq. 2 for scene adaptation, while the identity trajectory continues under Eq. 3.

After the two trajectories separate, the identity trajectory is retained within the estimated subject regions while sharing the evolving scene context elsewhere. We therefore combine the independently updated identity state with the updated primary state using the soft subject mask. We define the mask-weighted blending operator as

$$\mathcal{B}_{\mathbf{M}}(\mathbf{a}, \mathbf{b}) = (1 - \mathbf{M}) \odot \mathbf{a} + \mathbf{M} \odot \mathbf{b}, \tag{4}$$

where $\mathbf{a}$ and $\mathbf{b}$ denote the primary and identity states, respectively.

Let $\mathcal{S}_t(\mathbf{x}, \mathbf{v})$ denote one step of the underlying flow-matching sampler, which advances the current latent state from timestep t to the next timestep using the predicted velocity $\mathbf{v}$. At each subsequent

step, the primary state is first updated using the trajectory correction described in Sec. 3.3. We then independently advance the auxiliary state under the identity velocity:

$$\widetilde{\mathbf{x}}^I_{t-1} = \mathcal{S}_t\left(\mathbf{x}^I_t, \mathbf{v}^I_t(\mathbf{x}^I_t)\right), \tag{5}$$

and then blend it with the updated primary state using the subject mask:

$$\begin{aligned}\mathbf{x}^I_{t-1} &= \mathcal{B}_{\mathbf{M}_t}\left(\mathbf{x}^P_{t-1}, \widetilde{\mathbf{x}}^I_{t-1}\right)\\ &= (1-\mathbf{M}_t)\odot \mathbf{x}^P_{t-1} + \mathbf{M}_t \odot \widetilde{\mathbf{x}}^I_{t-1}.\end{aligned} \tag{6}$$

This update maintains a subject-focused identity history while synchronizing its surrounding context with the primary trajectory. The resulting persistent state provides the alternative identity history read by the paired trajectory correction in Sec. 3.3.

3.3 Paired Trajectory Correction

Our goal is to use the alternative identity history to supplement the primary trajectory while retaining the primary state as the carrier of scene adaptation. Following this motivation, PIT evaluates both states using the same identity predictor at the same timestep and transfers their matched prediction difference rather than the auxiliary prediction itself.

Using the clean-estimate mapping in Eq. 1, we define the *identity predictor* shown in Fig. 2 as the clean-latent estimate induced by the identity velocity in Eq. 3:

$$\begin{aligned}\mathcal{P}^I_t(\mathbf{x}) &= \mathcal{C}_t\left(\mathbf{x}, \mathbf{v}^I_t(\mathbf{x})\right)\\ &= \mathbf{x} - \sigma_t\left[F_{11}(\mathbf{x}) + s_I D_I(\mathbf{x})\right].\end{aligned} \tag{7}$$

At each timestep, we compute the primary clean estimate and two matched identity predictions:

$$\widehat{\mathbf{x}}^P_{0,t} = \mathcal{C}_t(\mathbf{x}^P_t, \mathbf{v}^P_t), \quad \widehat{\mathbf{x}}^{I\to I}_{0,t} = \mathcal{P}^I_t(\mathbf{x}^I_t), \quad \widehat{\mathbf{x}}^{P\to I}_{0,t} = \mathcal{P}^I_t(\mathbf{x}^P_t). \tag{8}$$

The latter two predictions form a matched pair: they use the same predictor, conditions, and timestep, differing only in the trajectory state being evaluated. We therefore define their difference as the history-dependent trajectory correction:

$$\Delta^{\mathrm{traj}}_t = \widehat{\mathbf{x}}^{I\to I}_{0,t} - \widehat{\mathbf{x}}^{P\to I}_{0,t}. \tag{9}$$

Expanding Eq. 9 gives

$$\Delta^{\mathrm{traj}}_t = \left(\mathbf{x}^I_t - \mathbf{x}^P_t\right) - \sigma_t\left[\mathbf{v}^I_t(\mathbf{x}^I_t) - \mathbf{v}^I_t(\mathbf{x}^P_t)\right]. \tag{10}$$

Thus, the correction is not a direct latent difference, but also reflects how the identity predictor responds to the two trajectory states.

We restrict the trajectory correction to the estimated subject support:

$$\widehat{\mathbf{x}}^{\mathrm{PIT}}_{0,t} = \widehat{\mathbf{x}}^P_{0,t} + \mathbf{M}_t \odot \Delta^{\mathrm{traj}}_t. \tag{11}$$

Here, $\mathbf{M}_t$ is obtained by aggregating per-subject attention maps; its construction is detailed in Appendix A.2. This operation preserves the primary prediction outside the resulting support.

Because the solver operates on velocity predictions, we express the corrected clean estimate in the corresponding velocity parameterization:

$$\mathbf{v}^{\mathrm{PIT}}_t = \mathbf{v}^P_t - \frac{1}{\sigma_t}\mathbf{M}_t \odot \Delta^{\mathrm{traj}}_t, \tag{12}$$

and update the primary state as

$$\mathbf{x}^P_{t-1} = \mathcal{S}_t\left(\mathbf{x}^P_t, \mathbf{v}^{\mathrm{PIT}}_t\right). \tag{13}$$

The updated primary state then provides the scene context for the auxiliary-state synchronization in Eq. 6. In this way, the persistent identity history supplies subject-local correction within the estimated subject regions, while the primary trajectory remains the carrier of scene adaptation.



Figure 3: **Qualitative comparisons.** (a) Results on LAMICBENCH++ using the same references and prompts. (b) Results on COCO-MIG with the same fixed text-derived references released by ContextGen. The displayed layouts are visualizations only, while each method receives spatial control through its native input format. PIT better preserves reference-specific identities while following the requested scene and spatial composition.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Implementation details. We instantiate PIT on Qwen-Image-Edit-2511 (Wu et al., 2025) and adapt its diffusion transformer to the layered multi-reference input using LoRA. We construct the LoRA training set from MICo-150K (Wei et al., 2026) by converting its generated reference images into the same CanvasComposer-style layered input representation; details are provided in Appendix A.1. We train a rank-128 LoRA for 20K optimization steps on 8 H100 GPUs, using a learning rate of 10^{-5} and a maximum input area of 1,048,576 pixels. Inference uses 40 sampling steps and an identity-aware warm-up of $K = 3$ steps. We retain the default Qwen-Image-Edit-2511 text CFG scale of 4.0, which corresponds to $s_T = 3$ under our residual parameterization. We set the image-guidance coefficient to the aligned value $s_I = 3$ without tuning it on the evaluation benchmark. Appendix C.4 analyzes sensitivity to s_I while keeping the model-default text guidance fixed. All inference experiments are conducted on a single H100 GPU.

Benchmarks and Metrics. We employ two distinct benchmark datasets for evaluation.

LAMICBENCH++. Following ContextGen (Xu et al., 2026), we evaluate on LAMICBENCH++, a multi-reference benchmark for assessing subject preservation and composition quality in subject-driven generation. The benchmark contains 160 cases with reference subjects spanning humans, animals, and objects, aggregated from XVerseBench (Chen et al., 2025), DreamBench++ (Peng et al., 2025b), and MS-Bench (Wang et al., 2025b). It includes 50 cases with two references, 40 with three, 30 with four, 20 with five, and 20 with more than five, which are grouped into Fewer Subjects (two or three references) and More Subjects (four or more references). We retain all evaluation cases and reference subjects from LAMICBENCH++, and construct the layout conditions based

Table 1: Comparison with external methods on LAMICBENCH++ using the original prompts. The final AVG column is computed on the All split. All metrics are higher-is-better; the best and second-best results in each column are bold and underlined, respectively.

Method	Fewer Subjects (2–3)						More Subjects (4+)						AVG
	ITC	AES	IDS	IPS	mIoU	AVG	ITC	AES	IDS	IPS	mIoU	AVG	
LAMIC	82.41	57.00	17.50	70.92	15.48	48.66	72.07	59.74	7.91	60.46	6.29	41.29	44.47
MS-Diffusion	87.11	58.96	13.06	79.02	34.56	54.54	80.05	**61.16**	11.23	73.03	34.03	51.90	53.09
MUSE	67.25	50.89	14.54	83.42	36.76	50.57	49.71	49.59	11.34	76.06	23.19	41.98	45.97
ContextGen	92.82	**59.52**	45.17	82.85	52.37	66.55	83.18	60.43	42.06	79.21	54.76	63.93	65.27
SIGMA-GEN	92.98	59.20	38.64	81.21	46.91	63.79	82.66	60.47	26.27	76.07	46.22	58.34	60.86
ConsistCompose	95.07	55.68	27.92	84.10	**67.85**	66.12	86.54	56.17	17.99	77.60	62.36	60.13	62.80
PositionIC	65.83	53.71	35.56	81.12	51.00	57.44	33.68	54.10	19.74	69.72	42.92	44.03	50.44
PIT (ours)	**96.00**	56.54	**55.50**	**84.97**	62.25	**71.05**	**91.32**	55.54	**48.42**	**84.10**	**66.65**	**69.21**	**70.18**

on its released bounding-box annotations. Followng the original evaluation protocol, we report four complementary metrics: AES for aesthetic quality, ITC for global text–image consistency evaluated using approximately 2K VQA questions, IDS for facial identity retention, and IPS for reference-subject preservation based on visual features. Since our setting explicitly requires subjects to appear in designated regions, we additionally report mIoU, which measures the localization overlap between matched generated subjects and their target bounding boxes. All metrics are mapped to a 0–100 scale, with higher values indicating better performance. We additionally report the balanced aggregate $\mathrm{AVG} = \frac{1}{5}\left(\mathrm{AES} + \mathrm{ITC} + \mathrm{IDS} + \mathrm{IPS} + \mathrm{mIoU}\right)$.

Table 2: Evaluation on COCO-MIG with text-derived references. All metrics are higher-is-better. The best and second-best results in each column are bold and underlined, respectively.

Model	SR	I-SR	mIoU	G-C	L-C
LAMIC	0.0175	0.1300	0.2366	29.8003	24.2109
MS-Diffusion	0.0450	0.2900	0.3485	30.3893	25.6633
MUSE	**0.3150**	**0.7078**	0.5630	30.8982	26.1318
ContextGen	0.1562	0.5616	0.5114	30.5701	26.3235
SIGMA-GEN	0.0475	0.3281	0.3739	30.4819	25.8185
ConsistCompose	0.2350	0.6034	0.5595	30.1106	25.8750
PositionIC	0.0800	0.3547	0.3826	29.7561	24.7346
PIT (ours)	0.2787	0.6528	**0.5955**	**31.0944**	**26.4938**

COCO-MIG with text-derived references. As a complementary evaluation of spatial and attribute control, we follow ContextGen (Xu et al., 2026) and adapt COCO-MIG (Zhou et al., 2024) to a reference-conditioned setting. Specifically, textual instance descriptions are first converted into subject reference images using FLUX.1-DEV (Black Forest Labs, 2024). We directly use the fixed, pre-generated references released by ContextGen, together with the prescribed layouts, for all evaluated methods. COCO-MIG comprises 800 COCO images (Lin et al., 2014) with spatial and color annotations. Under the official protocol, global and instance-level success rates (SR and I-SR) jointly measure spatial and color accuracy, while global and local CLIP scores (G-C and L-C) measure semantic consistency. All metrics are higher-is-better.

4.2 BASELINES

We compare PIT with representative layout-aware and multi-reference generation systems: LAMIC (Chen et al., 2026), MS-Diffusion (Wang et al., 2025b), PositionIC (Hu et al., 2026), MUSE (Peng et al., 2025a), ContextGen (Xu et al., 2026), SIGMA-GEN (Saha et al., 2026), and ConsistCompose (Shi et al., 2026). SIGMA-GEN is evaluated using its 2D spatial-control setting, matching the reference-subject and prescribed-region inputs of our benchmarks. We also use the unadapted Qwen-Image-Edit-2511 (Wu et al., 2025) as a zero-shot reference in supplementary experiments. For each external method, we use the authors' released implementation and checkpoint and provide the same references, prompts, and target layouts whenever supported by its interface.

Table 3: Controlled ablations on the All split of LAMICBENCH++. Only the indicated component changes; all other settings are fixed.

Variant	ITC	AES	IDS	IPS	mIoU	AVG
Single-trajectory	<u>93.66</u>	**56.36**	49.56	84.18	64.45	69.64
Static mask	92.27	55.55	50.99	83.96	**66.09**	69.77
Direct identity prediction	92.26	55.87	<u>51.56</u>	83.82	65.61	69.82
w/o context tethering	91.60	55.44	**52.18**	<u>84.35</u>	<u>65.96</u>	<u>69.90</u>
PIT	**93.95**	<u>56.10</u>	51.46	**84.43**	64.96	**70.18**

Table 4: Effect of identity-aware warm-up length K on the All split. $K = 0$ removes the warm-up; all other settings are fixed.

K	ITC	AES	IDS	IPS	mIoU	AVG
0	<u>94.43</u>	**57.03**	47.93	82.14	60.61	68.43
1	**94.91**	<u>56.85</u>	48.90	82.79	61.72	69.03
3	93.95	56.10	51.46	84.43	64.96	**70.18**
5	90.54	55.79	<u>51.74</u>	<u>84.47</u>	<u>65.65</u>	<u>69.64</u>
7	88.51	55.65	**52.01**	**84.78**	**66.74**	69.54

4.3 COMPARISON WITH EXISTING METHODS

Qualitative comparison. Figure 3 shows representative results under the same reference subjects, prompts, and layouts. PIT places the subjects in the specified regions while preserving their reference-specific identities. Meanwhile, it appropriately adapts their poses, interactions, appearance, and illumination to the textual description, producing a more coherent scene without simply copying the reference subjects.

Quantitative comparison. Table 1 shows that PIT achieves the best overall AVG of 70.18 on LAMICBENCH++, outperforming ContextGen by 4.91 points and demonstrating a stronger balance across identity, text, quality, and localization metrics. PIT also leads on both subject-count splits, with its larger 5.28-point gain on More Subjects indicating sustained effectiveness with four or more references. On COCO-MIG with text-derived references, PIT ranks first in mIoU, G-C, and L-C, indicating strong spatial alignment and global/local semantic consistency, while MUSE leads on the threshold-based SR and I-SR metrics. Together, these results demonstrate strong reference-specific generation and complementary spatial-semantic control.

4.4 ABLATION STUDY

Single-trajectory counterpart. The single-trajectory counterpart retains the complete primary inference process but removes the persistent identity state, providing the most direct controlled test of PIT's central design. The primary gain is identity preservation: PIT improves IDS from 49.56 to 51.46 and IPS from 84.18 to 84.43, while ITC and AES change by only $+0.29$ and -0.26, respectively. This targeted improvement is consistent with the persistent trajectory recovering reference-specific information without materially changing text consistency or visual quality. The overall AVG increases from 69.64 to 70.18, with a larger 0.80-point gain on the more challenging More Subjects split. Across three matched seeds, PIT consistently improves the mean AVG on both splits (Appendix C.3). In Fig. 4(a), the single-trajectory results retain the correct subject category but resemble different instances, whereas PIT preserves distinctive reference characteristics while adapting the subjects to the requested scene. Because the two variants share the same trained LoRA, layered inputs, support masks, guidance settings, warm-up, solver, and random seeds, these results provide controlled evidence for the benefit of maintaining and reading a persistent identity history.

Static-mask sensitivity. We replace the estimated subject support with the fixed metadata mask while keeping all other settings unchanged. The static mask increases mIoU from 64.96 to 66.09, consistent with restricting the trajectory correction to the user-specified spatial support throughout sampling. However, unlike the estimated support, it cannot adapt as the generated subject extent evolves with the prompt and scene. This setting yields lower ITC, AES, IDS, and IPS, reducing the overall AVG from 70.18 to 69.77. This comparison evaluates the sensitivity of PIT to fixed versus evolving correction support.

Correction formulation. We compare PIT's paired trajectory correction with direct transfer of the identity-state prediction. The direct-transfer variant is defined as $\hat{\mathbf{x}}_{0,t}^{\text{out}} = (1 - \mathbf{M}_t) \odot \hat{\mathbf{x}}_{0,t}^{P} + \mathbf{M}_t \odot \hat{\mathbf{x}}_{0,t}^{I \to I}$. Although direct transfer obtains comparable IDS and higher mIoU, it reduces ITC, AES, IPS, and the overall AVG from 70.18 to 69.82. This result is consistent with using the persistent trajectory to supplement, rather than directly replace, the scene-adaptive primary prediction.

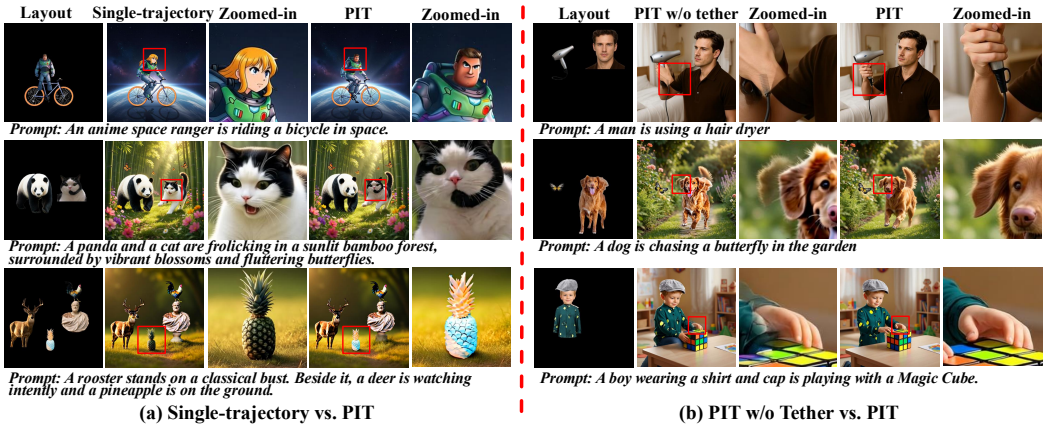

Figure 4: **Single-trajectory and context-tethering ablations.** (a) Compared with the single-trajectory counterpart, PIT better preserves reference-specific subject characteristics while maintaining scene adaptation. (b) Without context tethering, unconstrained evolution of the identity trajectory can introduce structural distortions or duplicated content, whereas PIT maintains a more coherent scene. Red boxes indicate the enlarged regions.

Context tethering. Context tethering keeps the persistent trajectory compatible with the primary scene context outside the estimated subject regions. Removing it increases IDS from 51.46 to 52.18 and mIoU from 64.96 to 65.96, but lowers ITC, AES, IPS, and the overall AVG from 70.18 to 69.90. It also produces structural distortions and duplicated content in the examples shown in Fig. 4(b). These results support using tethering to maintain scene coherence and structural stability rather than to directly strengthen identity.

Identity-aware warm-up length. Table 4 evaluates both the presence of identity-aware warm-up and its length K, with all other settings fixed. Compared with $K = 0$, the default $K = 3$ improves IDS, IPS, and mIoU by 3.53, 2.29, and 4.35 points, respectively, increasing AVG from 68.43 to 70.18. This result supports establishing shared identity-aware subject structure before the two trajectories assume different roles. Increasing K further continues to improve IDS, IPS, and mIoU, but delays trajectory separation and leaves fewer steps for primary scene adaptation, accompanied by lower ITC and AES. We therefore use $K = 3$, which achieves the highest overall AVG and provides the best balance between identity-aware initialization and subsequent scene adaptation.

5 Conclusion

We introduced PIT for layout-aware multi-reference image generation, motivated by identity persistence: correct reference-to-region binding does not ensure that reference-specific evidence remains recoverable as a single latent state adapts to a target scene. PIT maintains a scene-adaptive primary trajectory together with a subject-local persistent identity trajectory that preserves an alternative identity history. It uses the difference between matched clean predictions from the two states to supplement the primary trajectory within the estimated subject regions, while retaining the primary state as the carrier of scene adaptation. Experiments on LAMICBench++ demonstrate strong identity preservation, spatial accuracy, prompt consistency, and visual quality, while controlled comparisons with the matched single-trajectory counterpart isolate the benefit of maintaining the persistent identity history. Complementary results on COCO-MIG with text-derived references further demonstrate strong spatial and semantic controllability. Overall, controlled comparisons show that maintaining an alternative identity history improves reference-specific identity preservation during scene-adaptive multi-reference generation.

Limitations. Maintaining and reading the persistent identity trajectory requires additional model evaluations; runtime and memory measurements are provided in Appendix A.3. When a prompt calls for substantial changes to a reference subject's appearance or geometry, PIT may not fully satisfy both reference fidelity and prompt compliance.

AI USE STATEMENT

Generative AI tools assisted with writing, literature discovery, research and experimental discussions, coding, and the presentation of mathematical derivations and results. The authors reviewed and verified all AI-assisted content, made all final decisions, and take responsibility for the paper and accompanying artifacts.

REFERENCES

Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Roni Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 51304–51323, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/dfbb3d1807b21dadee735eb75069ada4-Paper-Conference.pdf.

Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. MultiDiffusion: Fusing diffusion paths for controlled image generation. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 1737–1752. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/bar-tal23a.html.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18392–18402, 2023. doi: 10.1109/CVPR52729.2023.01764.

Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22503–22513, 2023. doi: 10.1109/ICCV51070.2023.02062.

Bowen Chen, Brynn zhao, Haomiao Sun, Li Chen, Xu Wang, Daniel Du, and Xinglong Wu. Xverse: Consistent multi-subject control of identity and semantic attributes via dit modulation. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 7437–7459. Curran Associates, Inc., 2025. doi: 10.52202/085713-0254. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/0b0153a91f827b14e8bfea4e211362f3-Paper-Conference.pdf.

Yuzhuo Chen, Zehua Ma, Jianhua Wang, Kai Kang, Shunyu Yao, and Weiming Zhang. Lamic: Layout-aware multi-image composition via scalability of multimodal diffusion transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 3174–3182, 2026.

Dave Epstein, Allan Jabri, Ben Poole, Alexei Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 16222–16239. Curran Associates, Inc., 2023. doi: 10.52202/075280-0714. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/3469b211b829b39d2b0cfd3b880a869c-Paper-Conference.pdf.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NAQvF08TcyG.

Daniel Geng, Inbum Park, and Andrew Owens. Visual anagrams: Generating multi-view optical illusions with diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24154–24163, 2024. doi: 10.1109/CVPR52733.2024.02280.

Amir Hertz, Kfir Aberman, and Daniel Cohen-Or. Delta denoising score. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2328–2337, 2023. doi: 10.1109/ICCV51070.2023.00221.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Junjie Hu, Tianyang Han, Kai Ma, Jialin Gao, Yang Song, Xianhua He, Junfeng Luo, Xiaoming Wei, and Wenqiang Zhang. Positionic: Unified position and identity consistency for image customization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9139–9148, June 2026.

Qihan Huang, Siming Fu, Jinlong Liu, Hao Jiang, Yipeng Yu, and Jie Song. Resolving multi-condition confusion for finetuning-free personalized image generation. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 39, pp. 3707–3714, 2025.

Ziqi Huang, Kelvin C.K. Chan, Yuming Jiang, and Ziwei Liu. Collaborative diffusion for multi-modal face generation and editing. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6080–6090, 2023. doi: 10.1109/CVPR52729.2023.00589.

Sangwon Jang, Jaehyeong Jo, Kimin Lee, and Sung Ju Hwang. Identity decoupling for multi-subject personalization of text-to-image models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 100895–100937. Curran Associates, Inc., 2024. doi: 10.52202/079017-3200. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/b6e67ae290635d0874c4cb43ba2a2cfb-Paper-Conference.pdf.

Ruixiang Jiang and Chang Wen Chen. Diffartist: Towards structure and appearance controllable image stylization. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 9598–9607, 2025.

Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. Pnp inversion: Boosting diffusion-based editing with 3 lines of code. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=FoMZ4ljhVw.

Juil Koo, Chanho Park, and Minhyuk Sung. Posterior distillation sampling. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13352–13361, 2024. doi: 10.1109/CVPR52733.2024.01268.

Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Flowedit: Inversion-free text-based editing using pre-trained flow models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19721–19730, 2025. doi: 10.1109/ICCV51701.2025.01834.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1931–1941, 2023. doi: 10.1109/CVPR52729.2023.00192.

Yuseung Lee, Kunho Kim, Hyunjin Kim, and Minhyuk Sung. Syncdiffusion: Coherent montage via synchronized joint diffusions. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 50648–50660. Curran Associates, Inc., 2023. doi: 10.52202/075280-2203. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/9ee3a664ccfeabc0da16ac6f1f1cfe59-Paper-Conference.pdf.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22511–22521, 2023. doi: 10.1109/CVPR52729.2023.02156.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1.

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B. Tenenbaum. Compositional visual generation with composable diffusion models. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 423–439, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-19790-1.

Zhiheng Liu, Yifei Zhang, Yujun Shen, Kecheng Zheng, Kai Zhu, Ruili Feng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones 2: customizable image synthesis with multiple subjects. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. In *ACM SIGGRAPH 2024 Conference Papers*, SIGGRAPH '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400705250. doi: 10.1145/3641519.3657469. URL https://doi.org/10.1145/3641519.3657469.

Aryan Mikaeili, Amirhossein Alimohammadi, Negar Hassanpour, Ali Mahdavi-Amiri, and Andrea Tagliasacchi. Griffin: Generative reference and layout guided image composition, 2025. URL https://arxiv.org/abs/2509.23643.

Sicheng Mo, Fangzhou Mu, Kuan Heng Lin, Yanli Liu, Bochen Guan, Yin Li, and Bolei Zhou. Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7465–7475, 2024. doi: 10.1109/CVPR52733.2024.00713.

Fei Peng, Junqiang Wu, Yan Li, Tingting Gao, Di Zhang, and Huiyuan Fu. Muse: Multi-subject unified synthesis via explicit layout semantic expansion. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15885–15895, 2025a. doi: 10.1109/ICCV51701.2025.01474.

Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. Dreambench++: A human-aligned benchmark for personalized image generation. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=4GSOESJrk6.

Quynh Phung, Sandesh Ghimire, Minsi Hu, Chung-Chi Tsai, and Jia-Bin Huang. Universe: A unified modulation framework for segmentation-free, disentangled multi-concept personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22007–22016, June 2026.

Gordon Guocheng Qian, Ruihang Zhang, Tsai-Shien Chen, Yusuf Dalva, Anujraaj Argo Goyal, Willi Menapace, Ivan Skorokhodov, Meng Dong, Arpit Sahni, Daniil Ostashev, Ju Hu, Mukesh Singhal, Sergey Tulyakov, and Kuan-Chieh Jackson Wang. Canvascomposer: Personalized group photo generation via a multi-reference canvas, 2026. URL https://arxiv.org/abs/2510.20820.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22500–22510, 2023. doi: 10.1109/CVPR52729.2023.02155.

Oindrila Saha, Vojtech Krs, Radomir Mech, Subhransu Maji, Kevin James Blackburn-Matzen, and Matheus Gadelha. SIGMA-gen: Structure and identity guided multi-subject assembly for image generation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=x2DWTywZ1i.

Dong She, Siming Fu, Mushui Liu, Qiaoqiao Jin, Hualiang Wang, weihao zhou, and Jidong Jiang. Mosaic: Multi-subject personalized generation via correspondence-aware alignment and disentanglement. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 63128–63156, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/67182d38f943d51ff5f1896841751243-Paper-Conference.pdf.

Xuanke Shi, Boxuan Li, Xiaoyang Han, Zhongang Cai, Lei Yang, Quan Wang, and Dahua Lin. Consistcompose: Unified multimodal layout control for image composition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 495–505, June 2026.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1921–1930, 2023. doi: 10.1109/CVPR52729.2023.00191.

Pengcheng Wan, Liang Han, Lin Xu, Bowen Xiao, and Liqiang Nie. Mixdiffusion: Mixing diffusion-based uni-condition text-to-image generation models for multi-condition image synthesis, 2026. URL https://arxiv.org/abs/2607.17634.

Luozhou Wang, Guibao Shen, Wenhang Ge, Guangyong Chen, Yijun Li, and Yingcong Chen. Text-anchored score composition: Tackling condition misalignment in text-to-image diffusion models. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 21–37, Cham, 2025a. Springer Nature Switzerland. ISBN 978-3-031-72970-6.

Shulei Wang, Longhui Wei, Xin He, Jianbo Ouyang, Hui Lu, Zhou Zhao, and Qi Tian. Psr: Scaling multi-subject personalized image generation with pairwise subject-consistency rewards. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14756–14766, June 2026.

Xierui Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. Ms-diffusion: Multi-subject zero-shot image personalization with layout guidance. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 95118–95146, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/ed4df1609bf7d8602435341c9ce2ab5f-Paper-Conference.pdf.

Xudong Wang, Trevor Darrell, Sai Saketh Rambhatla, Rohit Girdhar, and Ishan Misra. Instancediffusion: Instance-level control for image generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6232–6242, 2024. doi: 10.1109/CVPR52733.2024.00596.

Xinyu Wei, Kangrui Cen, Hongyang Wei, Zhen Guo, Bairui Li, Zeqing Wang, Jinrui Zhang, and Lei Zhang. Mico-150k: A comprehensive dataset advancing multi-image composition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29695–29706, June 2026.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. URL https://arxiv.org/abs/2508.02324.

Qiang Xiang, Shuang Sun, Binglei Li, Dejia Song, Huaxia Li, Yibo Chen, Xu Tang, Yao Hu, and Junping Zhang. Instanceassemble: Layout-aware image generation via instance assembling attention. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 115424–115451. Curran Associates, Inc., 2025. doi: 10.52202/085713-3849. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/a793ef048356f4249df0a997758d9f33-Paper-Conference.pdf.

Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *International Journal of Computer Vision*, 133(3):1175–1194, 2025.

Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7418–7427, 2023. doi: 10.1109/ICCV51070.2023.00685.

Ruihang Xu, Dewei Zhou, Fan Ma, and Yi Yang. Contextgen: Contextual layout anchoring for identity-consistent multi-instance generation. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 95978–96007, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/9b01b79449b54c24ead5725e14affa79-Paper-Conference.pdf.

Sihan Xu, Yidong Huang, Jiayi Pan, Ziqiao Ma, and Joyce Chai. Inversion-free image editing with language-guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9452–9461, June 2024.

Han Yang, Chuanguang Yang, Qiuli Wang, Zhulin An, Weilun Feng, Libo Huang, and Yongjun Xu. Multi-party collaborative attention control for image customization. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7942–7951, 2025. doi: 10.1109/CVPR52734.2025.00744.

Yunkai Yang, Yudong Zhang, Xinying Chen, Haoyuan Liang, Yizhuo Niu, Jinshuai Cheng, Kunquan Zhang, Liziyue Fang, Weitao Wan, and Runmin Dong. Controlref: Efficient layout-guided multi-instance generation via anchored 4d-rope, 2026. URL https://arxiv.org/abs/2608.06878.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models, 2023. URL https://arxiv.org/abs/2308.06721.

Hui Zhang, Dexiang Hong, Yitong Wang, Jie Shao, Xinglong Wu, Zuxuan Wu, and Yu-Gang Jiang. Creatilayout: Siamese multimodal diffusion transformer for creative layout-to-image generation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18487–18497, 2025. doi: 10.1109/ICCV51701.2025.01718.

Dewei Zhou, You Li, Fan Ma, Xiaoting Zhang, and Yi Yang. Migc: Multi-instance generation controller for text-to-image synthesis. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6818–6828, 2024. doi: 10.1109/CVPR52733.2024.00651.

A Implementation Details

A.1 Layered Input Preprocessing and LoRA Adaptation

Our conditioning representation follows the transparent latent-pruning principle of CanvasComposer (Qian et al., 2026): after encoding the layered inputs, we retain tokens from mask-derived subject regions and discard tokens from empty canvas regions. As an implementation-level adaptation, subject layers with non-overlapping masks are combined into shared conditioning canvases before encoding, reducing redundant canvas representations. The input subject masks also provide the initial spatial support for restricting the additional image-guidance residual in Eq. 2; this support is subsequently refined as described in Appendix A.2. The same preprocessing is used by PIT and all controlled single-trajectory variants, while external baselines retain their native input interfaces.

For LoRA training, we convert MICo-150K into the same layered input format. We extract foreground subjects from its generated reference images, match them to the corresponding instances in the target images, and retain verified reference–target correspondences to construct the spatially aligned subject layers. During LoRA adaptation, we retain both text and image conditions for every training sample, with all conditioning-dropout probabilities set to zero.

A.2 Subject-Support Mask Construction

Scope. We derive a subject-support mask by aggregating the backbone's existing image-attention maps. The mask restricts the additional image-guidance residual and trajectory correction to the estimated subject regions, and provides the spatial support for context tethering. It requires neither a separate mask predictor nor segmentation of the generated output. The same construction is used by PIT and its controlled single-trajectory counterpart.

Attention aggregation. For each subject n, we record the attention scores between the current generated-image tokens and the conditioning tokens belonging to that subject during the $F_{11}(\mathbf{x}_t^I)$ evaluation of the persistent identity state. At zero-indexed transformer block ℓ, the score at generated-image token p is

$$A_{t,\ell}^n(p) = \frac{1}{N_h} \sum_{h=1}^{N_h} \max_{q \in \mathcal{K}_n} \frac{\langle Q_{t,\ell,h}(p), K_{t,\ell,h}(q) \rangle}{\sqrt{d}}, \tag{14}$$

where $\mathcal{K}_n$ indexes the conditioning tokens of subject n, and N_h denotes the number of attention heads. We average scores from transformer blocks $\mathcal{L} = \{24, 30, 36\}$ and apply per-subject min–max normalization:

$$\bar{A}_t^n = \mathrm{Norm}\left(\frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} A_{t,\ell}^n \right). \tag{15}$$

To reduce temporal fluctuations, we maintain an exponential moving average,

$$E_t^n = 0.8 E_{t-1}^n + 0.2 \bar{A}_t^n, \tag{16}$$

with $E_t^n = \bar{A}_t^n$ at its first observation. We retain the top 15% spatial tokens for each subject and apply a 5×5 average-pooling filter to obtain a soft support map S_t^n. The subject maps are combined and resized to the latent resolution:

$$\mathbf{M}_t = \mathrm{Resize}\left(\max_n S_t^n \right), \tag{17}$$

where resizing uses bilinear interpolation and the result is clipped to $[0, 1]$. The spatial maximum produces a single union support covering the attended subject regions, rather than assigning a separate correction to each subject.

Initialization and static alternative. The initial and fallback support is obtained by nearest-neighbor resizing the input subject masks to the token grid and expanding them by one token. When attention scores are unavailable, PIT uses this input support directly. In the static-mask ablation, the same input-derived support is kept fixed throughout sampling. All mask settings are fixed across samples and are not tuned per instance. Table 3 reports the corresponding sensitivity analysis.

A.3 Runtime and Memory Cost

PIT uses the same backbone and 40-step sampling schedule as its single-trajectory counterpart, while requiring additional model evaluations to maintain and read the persistent identity trajectory. We measure CUDA-synchronized generation latency and peak allocated memory on a single H100 over all 160 benchmark inputs, after one unmeasured warm-up execution. Model loading, image saving, visualization, and metric computation are excluded.

Table 5: Runtime and memory cost on a single H100 over 160 benchmark inputs. Both methods use the same backbone and 40-step sampling schedule. Latency is reported as mean $\pm$ standard deviation.

Method	Steps	Latency (s/image)	Relative	Peak VRAM (GB)
Single-trajectory	40	41.61 ± 4.77	$1.00\times$	59.05
PIT	40	65.68 ± 6.96	$1.58\times$	59.05

PIT incurs $1.58\times$ latency relative to the matched single-trajectory counterpart, while the measured peak memory remains unchanged at the reported precision. In the current implementation, the additional trajectory evaluations are fixed per denoising step rather than instantiated separately for each subject.

B Theoretical Supplement

This appendix derives the spatial operator used for both auxiliary-state tethering and trajectory correction. Given a spatial ownership map, the operator is the unique pointwise least-squares consensus of the two latent targets.

B.1 Definition and Unique Closed-Form Solution

Let $\mathbf{a}, \mathbf{b} \in \mathbb{R}^{h\times w\times d}$ be two latent-valued spatial targets and let $\mathbf{M} \in [0,1]^{h\times w}$ be an ownership map. For every spatial position p, define

$$\ell_p(\mathbf{z}_p) = \frac{1}{2}(1-M_p)\|\mathbf{z}_p - \mathbf{a}_p\|_2^2 + \frac{1}{2}M_p\|\mathbf{z}_p - \mathbf{b}_p\|_2^2. \tag{18}$$

The gradient with respect to the channel vector $\mathbf{z}_p$ is

$$\begin{aligned}\nabla_{\mathbf{z}_p}\ell_p &= (1-M_p)(\mathbf{z}_p - \mathbf{a}_p) + M_p(\mathbf{z}_p - \mathbf{b}_p)\\ &= \mathbf{z}_p - \big[(1-M_p)\mathbf{a}_p + M_p\mathbf{b}_p\big].\end{aligned} \tag{19}$$

The Hessian is the identity matrix because $(1-M_p)+M_p = 1$. Hence Eq. 18 is strictly convex and has the unique minimizer

$$\mathbf{z}_p^\star = (1-M_p)\mathbf{a}_p + M_p\mathbf{b}_p. \tag{20}$$

Stacking the independent positions yields Eq. 4. An equivalent weighted-norm notation could place $(1-M_p)^{1/2}$ and $M_p^{1/2}$ inside the norms because $\|\sqrt{w}\,\mathbf{u}\|_2^2 = w\|\mathbf{u}\|_2^2$. We use the sum-of-weighted-squares form in Eq. 18 because it exposes the role of the mask directly and avoids unnecessary square-root notation.

B.2 Binary Selection and Soft Ownership

If $M_p \in \{0,1\}$, Eq. 20 is exact region selection: it returns $\mathbf{a}_p$ for $M_p = 0$ and $\mathbf{b}_p$ for $M_p = 1$. If $M_p \in (0,1)$, it is a convex barycenter between the two targets. The same result has a simple probabilistic interpretation. Let a random target $\mathbf{Y}_p$ equal $\mathbf{a}_p$ with probability $1-M_p$ and $\mathbf{b}_p$ with probability M_p. Then

$$\ell_p(\mathbf{z}_p) = \frac{1}{2}\mathbb{E}\left[\|\mathbf{z}_p - \mathbf{Y}_p\|_2^2\right], \tag{21}$$

whose minimizer is the conditional mean $\mathbb{E}[\mathbf{Y}_p] = (1-M_p)\mathbf{a}_p + M_p\mathbf{b}_p$. Thus, the soft-mask formula is the canonical squared-error relaxation of binary spatial ownership and introduces no confidence parameter beyond the mask itself.

B.3 Application to Subject-Local Tethering

Before tethering, the auxiliary solver produces $\widetilde{\mathbf{x}}_{t-1}^I$, while the scene-adaptive trajectory produces $\mathbf{x}_{t-1}^P$. We assign the primary state as the non-subject target and the independently updated auxiliary state as the subject target:

$$\mathbf{x}_{t-1}^I = \mathcal{B}_{\mathbf{M}_t}\left(\mathbf{x}_{t-1}^P, \widetilde{\mathbf{x}}_{t-1}^I\right). \tag{22}$$

Equivalently, subtracting $\mathbf{x}_{t-1}^P$ yields $\mathbf{x}_{t-1}^I - \mathbf{x}_{t-1}^P = \mathbf{M}_t \odot (\widetilde{\mathbf{x}}_{t-1}^I - \mathbf{x}_{t-1}^P)$. For a binary mask, tethering eliminates direct state differences outside the subject regions while preserving the independently updated auxiliary state inside them. For a soft mask, it attenuates the difference according to spatial ownership. This guarantee applies to the current latent update; because the denoiser has a global receptive field, surrounding context may still influence later predictions indirectly.

B.4 Application to Regional Trajectory Correction

We use the primary clean estimate $\widehat{\mathbf{x}}^P_{0,t}$ as the base prediction and incorporate the matched trajectory correction through the subject mask. Substituting $\widehat{\mathbf{x}}^P_{0,t}$ and $\widehat{\mathbf{x}}^P_{0,t} + \Delta^{\mathrm{traj}}_t$ into Eq. 4 gives

$$\begin{aligned}
&\mathcal{B}_{\mathbf{M}_t}\left(\widehat{\mathbf{x}}^P_{0,t}, \widehat{\mathbf{x}}^P_{0,t} + \Delta^{\mathrm{traj}}_t\right) \\
&\quad = (1 - \mathbf{M}_t) \odot \widehat{\mathbf{x}}^P_{0,t} + \mathbf{M}_t \odot \left(\widehat{\mathbf{x}}^P_{0,t} + \Delta^{\mathrm{traj}}_t\right) \\
&\quad = \widehat{\mathbf{x}}^P_{0,t} + \mathbf{M}_t \odot \Delta^{\mathrm{traj}}_t,
\end{aligned} \tag{23}$$

which recovers Eq. 11. Using $\mathcal{C}_t(\mathbf{x}, \mathbf{v}) = \mathbf{x} - \sigma_t \mathbf{v}$, the corresponding velocity for $\sigma_t > 0$ is

$$\begin{aligned}
\mathbf{v}^{\mathrm{PIT}}_t &= \frac{\mathbf{x}^P_t - \widehat{\mathbf{x}}^{\mathrm{PIT}}_{0,t}}{\sigma_t} \\
&= \mathbf{v}^P_t - \frac{1}{\sigma_t}\mathbf{M}_t \odot \Delta^{\mathrm{traj}}_t,
\end{aligned} \tag{24}$$

recovering Eq. 12. Thus, the masked correction in clean-latent space is converted exactly into a velocity correction for the original solver. This conversion preserves the number of solver steps, while maintaining and reading the auxiliary trajectory requires additional denoiser evaluations.

B.5 How the Correction Reflects Persistent History

Let $\Phi^I_t(\mathbf{x}) = \mathcal{S}_t(\mathbf{x}, \mathbf{v}^I_t(\mathbf{x}))$ denote the identity update before context tethering, and let $\boldsymbol{\delta}_t = \mathbf{x}^I_t - \mathbf{x}^P_t$. Using the updated primary state $\mathbf{x}^P_{t-1}$ from Eq. 13, the tethered state difference satisfies

$$\boldsymbol{\delta}_{t-1} = \mathbf{M}_t \odot \left[\Phi^I_t(\mathbf{x}^I_t) - \mathbf{x}^P_{t-1}\right]. \tag{25}$$

Adding and subtracting $\Phi^I_t(\mathbf{x}^P_t)$ separates this recurrence into

$$\begin{aligned}
\boldsymbol{\delta}_{t-1} = \mathbf{M}_t \odot \Big[&\Phi^I_t(\mathbf{x}^I_t) - \Phi^I_t(\mathbf{x}^P_t) \\
&+ \Phi^I_t(\mathbf{x}^P_t) - \mathbf{x}^P_{t-1}\Big].
\end{aligned} \tag{26}$$

When Φ^I_t is differentiable and the state difference is sufficiently small, a first-order expansion around $\mathbf{x}^P_t$ gives

$$\boldsymbol{\delta}_{t-1} \approx \mathbf{M}_t \odot \left[\mathbf{J}^I_t \boldsymbol{\delta}_t + \Phi^I_t(\mathbf{x}^P_t) - \mathbf{x}^P_{t-1}\right], \tag{27}$$

where $\mathbf{J}^I_t$ is the Jacobian of Φ^I_t evaluated at $\mathbf{x}^P_t$. The first term propagates the existing state difference through the identity update, while the remaining term compares an identity-oriented update from the primary state with the actual corrected primary update. Through this recurrence, the state difference reflects preceding updates and context tethering across denoising steps.

The identity-oriented predictor defined in Eq. 7 maps the two states to matched clean-latent estimates. Under the same local differentiability and small-gap assumptions, their difference satisfies

$$\begin{aligned}
\Delta^{\mathrm{traj}}_t &= \mathcal{P}^I_t(\mathbf{x}^I_t) - \mathcal{P}^I_t(\mathbf{x}^P_t) \\
&\approx \mathbf{J}_{\mathcal{P}^I_t}(\mathbf{x}^P_t)\boldsymbol{\delta}_t.
\end{aligned} \tag{28}$$

Both predictions use the same conditions, timestep, and prediction rule, so their difference arises from the two evolving input states. The correction therefore reflects the accumulated state difference through a common identity-oriented predictor. When the two states coincide, the correction vanishes exactly at that step.

B.6 Properties and Scope

The resulting operator has four useful properties. First, it is parameter-free once $\mathbf{M}$ is specified. Second, it is spatially local: changing the subject target cannot directly change an output position with $M_p = 0$. Third, it is endpoint-consistent: $\mathcal{B}_{\mathbf{0}}(\mathbf{a}, \mathbf{b}) = \mathbf{a}$ and $\mathcal{B}_{\mathbf{1}}(\mathbf{a}, \mathbf{b}) = \mathbf{b}$. Fourth, PIT is collapse-consistent because $\Delta^{\mathrm{traj}}_t = \mathbf{0}$ when the two states coincide.

Under the pointwise squared-Euclidean objective, Eq. 4 is the unique minimizer. We use this form because it directly realizes mask-defined spatial ownership without additional learned or tuned parameters and applies consistently to both tethering and trajectory transfer.



Figure 5: **Qualitative comparison on LayoutSAM-Eval with text-derived references.** The displayed layouts visualize the prescribed spatial conditions; each method receives the same reference subjects and prompt through its native input format. From top to bottom, the prompts are: "This is a photograph showcasing the night view of a city, capturing the dazzling lights and architectural outlines of the city skyline. The image captures the lights of the city reflected on the water surface, adding a touch of tranquility and beauty to the scene. The foreground features a series of modern buildings, including a prominent dome structure and a tower with a unique design. The midground is a row of skyscrapers, their lights twinkling in the night sky. The background is the city skyline, with the lights of the buildings dotting the night sky, creating a vibrant urban atmosphere."; "This is a photo of an indoor environment in a realistic style, showcasing a beautifully decorated round table. The table is covered with a white tablecloth, on which a bouquet of orange flowers and a card are placed. The table is surrounded by several wooden chairs, with a wooden bookshelf filled with various bottles and jars in the background. The entire scene is illuminated by soft light, creating a warm and comfortable atmosphere."; and "This is a photograph showcasing a corner of a city street, with the focus on two bright blue chairs and a white table. The chairs are placed on the sidewalk, with a simple white table between them. The wall behind the chairs is adorned with two wooden windows, each with a red flower pot containing green plants. On the left side of the image, there is a decorative stone, next to which is a large flower pot filled with red flowers. On the right side of the image, there is a sign with text, but the content is not clear. The entire scene is illuminated by soft natural light, creating a tranquil and pleasant atmosphere."

C Supplementary Experiments

C.1 Complementary Evaluation on LayoutSAM-Eval with Text-Derived References

Dataset and protocol. LayoutSAM-Eval was introduced with CreatiLayout (Zhang et al., 2025) as an open-set layout-to-image benchmark containing 5K prompts with detailed entity-level boxes and descriptions. Following ContextGen (Xu et al., 2026), we use the 999 cases available in the shared reference-conditioned metadata from its filtered 1K subset with sufficiently large boxes, together with the FLUX.1-Dev-generated entity references released by ContextGen. Spatial measures whether the described entity is present in its designated crop; Color, Texture, and Shape use MLLM-based visual question answering to assess the corresponding entity attributes. CLIP measures global image–caption semantic alignment, and Pick denotes the learned text–image preference score. All metrics are higher-is-better. For completeness, we additionally report AVG as the unweighted arithmetic mean of the six displayed metrics, which are all expressed on a 0–100 scale. This aggregate is provided only as a compact summary and is not part of the official LayoutSAM-Eval protocol.

Table 6: Evaluation on the common 999-case LayoutSAM-Eval subset with text-derived references. AVG is the unweighted mean of the six reported metrics. Best and second-best results in each column are bold and underlined, respectively.

Method	Spatial	Color	Texture	Shape	CLIP	Pick	AVG
LAMIC	86.99	77.65	79.06	77.71	27.33	21.97	61.78
MS-Diffusion	86.45	76.84	79.24	77.51	27.03	<u>22.73</u>	61.63
MUSE	92.97	85.18	**88.31**	87.29	26.13	21.54	66.90
ContextGen	<u>93.62</u>	**85.67**	87.26	86.61	27.59	**22.80**	67.26
SIGMA-GEN	92.38	84.73	86.05	85.48	27.27	22.54	66.41
ConsistCompose	93.16	84.51	88.04	**88.23**	**28.09**	22.24	<u>67.38</u>
PositionIC	68.84	55.18	55.72	53.70	22.25	20.81	46.08
PIT (ours)	**93.65**	<u>85.27</u>	<u>88.18</u>	<u>87.99</u>	<u>27.98</u>	22.35	**67.57**

Table 7: Robustness across three matched random seeds on LAMICBENCH++. We report the raw result of every run and their mean ± sample standard deviation. Seed 42 is the main-paper run, while seeds 1 and 7 are additional reruns. Higher is better.

Seed	Method	Fewer AVG	More AVG	All AVG
1	Single-trajectory	70.79	68.83	69.89
	PIT	**71.32**	**69.25**	**70.41**
7	Single-trajectory	70.45	68.88	69.65
	PIT	**71.10**	**69.24**	**70.19**
42	Single-trajectory	70.81	68.41	69.64
	PIT	**71.05**	**69.21**	**70.18**
Mean ± std.	Single-trajectory	70.68 ± 0.20	68.71 ± 0.26	69.73 ± 0.14
	PIT	$\mathbf{71.16 \pm 0.14}$	$\mathbf{69.23 \pm 0.02}$	$\mathbf{70.26 \pm 0.13}$

Scope of the evaluation. LayoutSAM-Eval complements LAMICBENCH++ by evaluating layout following, entity attributes, and global text–image quality. Because its text-derived references may already encode the requested attributes and its metrics do not directly compare generated subjects with reference images, we use it primarily to assess spatial and attribute controllability.

Figure 5 provides representative qualitative results across urban, indoor, and street scenes. PIT follows the prescribed spatial composition while integrating the supplied subject layers into coherent scenes that match the requested attributes and context.

C.2 LAMICBENCH++ EVALUATION DETAILS

LAMICBENCH++ contains 160 multi-reference cases spanning human, animal, and object subjects: 50 cases with two references, 40 with three, 30 with four, 20 with five, and 20 with more than five. We retain all evaluation cases and reference subjects and construct the layout conditions based on the released bounding-box annotations. The Fewer Subjects split contains the 90 cases with two or three references, while the More Subjects split contains the remaining 70 cases.

All quantitative experiments use the complete original prompts; prompts are abbreviated only in figures for readability. Every method is evaluated on the same case list. IDS is computed on valid human faces, whereas IPS and mIoU are computed per reference subject and aggregated at the case level. The All result is evaluated over all 160 cases rather than obtained by averaging the two subject-count splits.

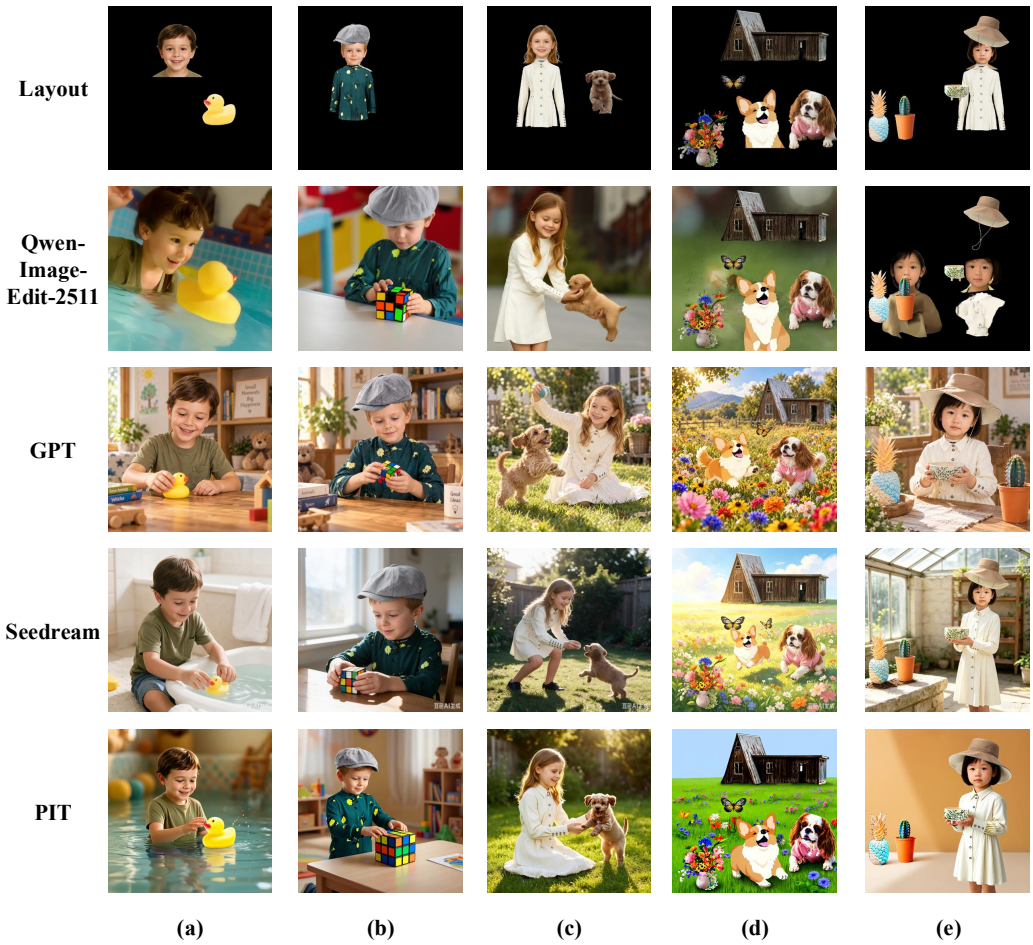

Figure 6: **Qualitative comparison with zero-shot general models.** Rows show the target layout, unadapted Qwen-Image-Edit-2511, GPT, Seedream, and PIT. We prepend a fixed task description for the closed-source models to clarify reference-identity preservation and target-region placement. The unadapted Qwen model receives only the original benchmark prompt, keeping its comparison with the LoRA-adapted model prompt-matched. The prompts are: (a) A boy is playing with a rubber duck. (b) A boy wearing a shirt and cap is playing with a Magic Cube. (c) A girl wearing a dress is playing with a puppy. (d) A dog and a Corgi are playing with a butterfly in a field of flowers, with a wooden house in the background. (e) A girl is wearing a dress and a hat, holding a ceramic bowl, with a pineapple and a cactus nearby. Even with explicit instructions, the general models often alter subject identities or deviate from the prescribed regions, whereas PIT more consistently satisfies both constraints while adapting the subjects to the requested scene.

Table 8: Image-guidance scale analysis on the All split of LAMICBENCH++. We fix $s_T = 3$ and vary s_I while keeping all other settings identical. Bold values indicate the better result within each matched single-trajectory/PIT pair. The dagger marks the aligned setting used in the main experiments. All metrics are higher-is-better.

s_I	Method	ITC	AES	IDS	IPS	mIoU	AVG
2	Single-trajectory	**95.93**	**57.15**	46.86	83.91	61.44	69.05
	PIT	95.84	56.97	**48.49**	**84.20**	**62.71**	**69.64**
3†	Single-trajectory	93.66	**56.36**	49.56	84.18	64.45	69.64
	PIT	**93.95**	56.10	**51.46**	**84.43**	**64.96**	**70.18**
4	Single-trajectory	**94.10**	**55.50**	50.28	83.84	66.04	69.95
	PIT	93.07	55.10	**51.25**	**84.51**	**67.28**	**70.24**

C.3 ROBUSTNESS ACROSS RANDOM SEEDS

We evaluate the controlled single-trajectory counterpart and PIT across three matched random seeds, while keeping the trained LoRA, packed inputs, dynamic masks, guidance scales, warm-up, sampler, and all other inference settings fixed. Seed 42 is the main-paper setting, and seeds 1 and 7 are additional reruns. Table 7 reports the raw split-level and overall AVG for every seed, followed by the mean and sample standard deviation across the three seeds. PIT obtains a higher All AVG for every seed and improves the three-seed mean from 69.73 ± 0.14 to 70.26 ± 0.13.

C.4 IMAGE-GUIDANCE SCALE ANALYSIS

We retain the default Qwen-Image-Edit-2511 text CFG scale of 4.0, corresponding to $s_T = 3$ under our residual parameterization, and evaluate PIT and its matched single-trajectory counterpart at $s_I \in \{2, 3, 4\}$. The main setting uses $s_I = 3$.

As shown in Table 8, increasing s_I already strengthens identity preservation in the single-trajectory counterpart: its IDS rises from 46.86 at $s_I = 2$ to 50.28 at $s_I = 4$. Nevertheless, at every matched scale, PIT further improves IDS, IPS, mIoU, and AVG. The incremental IDS gain is $+1.63$, $+1.90$, and $+0.96$ for $s_I = 2$, 3, and 4, respectively. The smaller improvement at $s_I = 4$ is consistent with diminishing gains once single-trajectory identity preservation is already stronger.

The sweep also exposes a trade-off across metrics. Relative to the default $s_I = 3$, PIT at $s_I = 4$ obtains a nearly unchanged AVG (70.24 vs. 70.18), but lower ITC (93.07 vs. 93.95), AES (55.10 vs. 56.10), and IDS (51.25 vs. 51.46); its small aggregate increase is driven primarily by improved localization. We therefore use $s_I = 3$ as a balanced default. Across all matched image-guidance scales, PIT consistently improves the single-trajectory counterpart.

C.5 ZERO-SHOT EVALUATION OF GENERAL MODELS

We further examine whether general image generation and editing models can solve our task without task-specific adaptation. As shown in Figure 6, we prepend a fixed task description for the closed-source models, clarifying that the inputs specify reference subjects whose identities and prescribed regions should be preserved. For the unadapted Qwen-Image-Edit-2511, we retain the original benchmark prompt used by our adapted model so that this row provides a prompt-matched comparison before task-specific LoRA training.

Despite receiving explicit task instructions, the closed-source models do not reliably satisfy identity preservation and spatial placement simultaneously. They can generate plausible scenes, but subjects may drift from the prescribed regions, disappear, or lose reference-specific characteristics. The unadapted Qwen model exhibits similar difficulty under the benchmark prompt. In contrast, PIT more consistently preserves the intended identities and layouts while adapting the subjects to the requested scene. These examples illustrate PIT's advantage over the evaluated zero-shot settings in jointly preserving subject identities and spatial placement.

Table 9: User-study best-choice counts and rates on 20 selected LAMICBENCH++ cases, evaluated by 20 participants. Each column contains 400 choices and sums to 100%.

Method	Reference fidelity	Scene adaptation	Spatial compliance	Overall preference
LAMIC	36 (9.00%)	50 (12.50%)	34 (8.50%)	21 (5.25%)
MS-Diffusion	46 (11.50%)	61 (15.25%)	31 (7.75%)	55 (13.75%)
MUSE	9 (2.25%)	15 (3.75%)	20 (5.00%)	9 (2.25%)
ContextGen	75 (18.75%)	83 (20.75%)	83 (20.75%)	85 (21.25%)
SIGMA-GEN	34 (8.50%)	32 (8.00%)	39 (9.75%)	30 (7.50%)
ConsistCompose	33 (8.25%)	36 (9.00%)	38 (9.50%)	35 (8.75%)
PositionIC	15 (3.75%)	13 (3.25%)	25 (6.25%)	13 (3.25%)
PIT (ours)	**152 (38.00%)**	**110 (27.50%)**	**130 (32.50%)**	**152 (38.00%)**

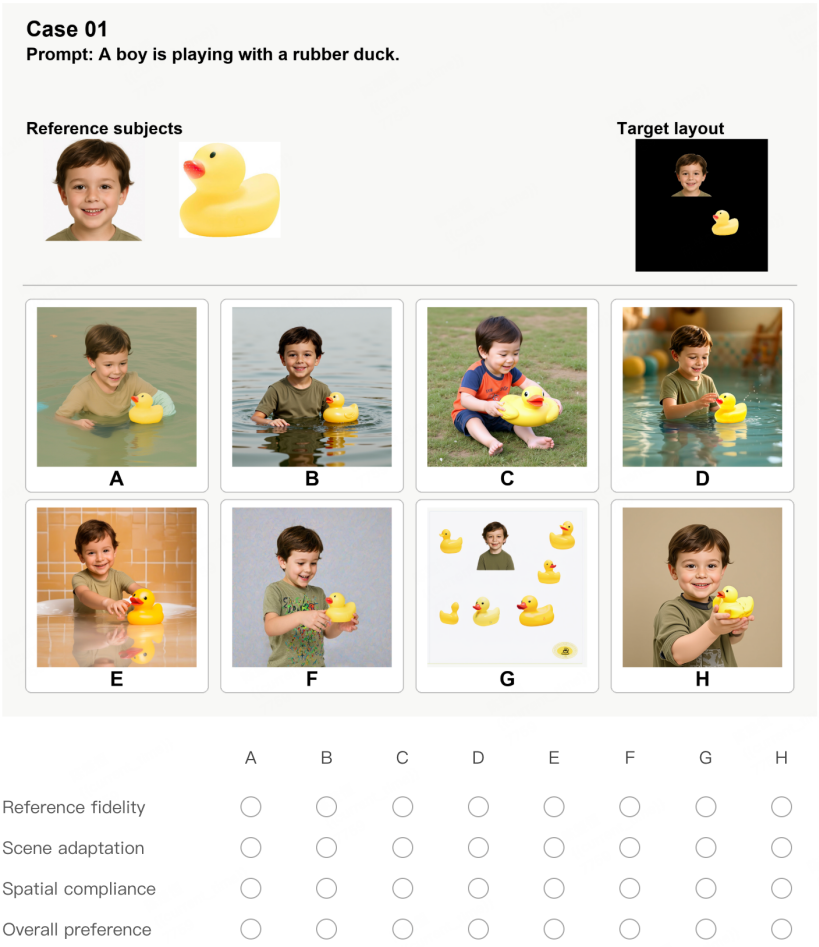

Figure 7: **User-study interface.** Participants viewed the references, prompt, target layout, and eight anonymized outputs, and selected one output under each evaluation criterion. Method identities were hidden, and the mapping to labels A–H was randomized independently for every case.

C.6 USER STUDY

Protocol. Twenty participants evaluated 20 selected cases from LAMICBENCH++. For each case, they were shown the reference subjects, prompt, target layout, and outputs from eight methods. The method names were hidden, and the assignment of methods to labels A–H was randomized

independently for each case. Participants selected one best output for each criterion and could select the same output for multiple criteria.

Reference fidelity asks whether all reference subjects retain their distinctive appearance, including facial characteristics, patterns, and shape; reasonable changes in pose or illumination are not considered identity loss. *Scene adaptation* asks whether the output follows the prompt and integrates the subjects naturally into the requested scene, including specified poses, interactions, and lighting. *Spatial compliance* asks whether each subject appears in its corresponding target region; the black background in the layout is only a visualization and is not a generation requirement. *Overall preference* asks participants to jointly consider reference fidelity, prompt following, spatial accuracy, and visual naturalness.

For method m and criterion q, we compute its best-choice rate as

$$P_{m,q} = \frac{N_{m,q}}{20 \times 20} \times 100\%,$$

where $N_{m,q}$ is the number of times method m is selected under criterion q. There are 400 selections per criterion, and the rates in each column sum to 100%.

PIT receives the highest best-choice rate on all four criteria.