TopoRefine: Topology-Aware Correspondence and Residual Refinement for Training-Free Subject-Consistent Generation

Zhanxin Gao[1] **Zexin Ti**[1] **Chen Zhao**[1] **Beier Zhu**[2]
Xiantao Hu[1] **Yanhao Ge**[3] **Jian Yang**[1] **Ying Tai**[1] *

[1]Nanjing University [2]University of Science and Technology of China [3]VIVO
zxgao@smail.nju.edu.cn yingtai@nju.edu.cn



Figure 1: `TopoRefine` **generates subjects in diverse scenes while preserving their distinctive identities.** The results exhibit varied poses, actions, and spatial layouts across different contexts.

Abstract

Subject-consistent generation (SCG) aims to preserve subject identity across diverse text-driven contexts. Recent training-free methods improve consistency through correspondence-aware transfer and extended attention. However, semantic-driven correspondence can route identity features to appearance-similar but structurally incompatible regions, while extended attention may perturb target-driven pose, layout, and local structure. These limitations can cause reference-pose over-copying and structural artifacts, suggesting that training-free SCG requires

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

controlled identity injection: the target generation should retain its prompt-driven structure, and the reference image should provide identity without overriding it. We propose TopoRefine, a training-free framework with Topology-aware Subject Correspondence (TSC) and Residual Identity Refinement (RIR). TSC regularizes semantic matching with subject-internal geodesic relations for structurally compatible identity transfer, while RIR injects reference identity as a foreground-gated, magnitude-aligned residual. Qualitative and quantitative results show that TopoRefine improves subject consistency while better preserving target-driven structure.

1 Introduction

Large-scale generative models, including diffusion models [1, 2, 3, 4, 5, 6, 7, 8] and visual autoregressive models [9], have greatly advanced realistic and creative visual synthesis. In this context, Subject-Consistent Generation (SCG) [10, 11] has emerged as an important task, aiming to generate images that depict the same subject consistently across diverse text-driven contexts. Fig. 1 illustrates this goal through results generated by TopoRefine, which preserve subject identity while varying poses, actions, and scene layouts. This capability is valuable for applications such as visual storytelling, character design, digital avatars, and personalized content creation.

To avoid costly per-subject training or optimization, recent work has increasingly explored training-free methods [12, 13] for consistent text-to-image generation. These methods directly modulate the internal representations of pretrained diffusion models, enabling zero-shot generalization to unseen subjects and plug-and-play adaptation across model architectures. Early approaches often rely on extended attention, which extends self-attention by allowing tokens from different generated images to interact in a shared attention space. While extended attention improves subject consistency, it can also couple the generated images, causing the reference subject to over-constrain the target structure, from global pose and layout to local structural details. This coupling may lead to reference-pose over-copying and local structural artifacts [14, 12, 15]. Recent methods therefore move beyond global sharing toward correspondence-aware identity transfer.

CharaConsist [16] uses point-tracking attention to establish fine-grained cross-image correspondences, while CoDi [14] performs correspondence-guided identity transport followed by late identity refinement. These methods move training-free SCG from global feature sharing toward more controlled identity injection. However, current training-free SCG methods remain insufficiently controlled in both correspondence estimation and late-stage identity injection. First, at the correspondence level, their matching mechanisms are still primarily driven by point-wise semantic similarity. Such semantic-driven correspondence can be ambiguous under repeated textures, symmetric parts, and non-rigid deformation, where visually similar regions may correspond to structurally different subject parts. As illustrated in Fig. 2(a), misrouted identity features can perturb the target-driven pose and overall structure, producing scattered local structural errors. Second, at the attention-update level, extended attention enhances identity consistency by placing reference-subject tokens and target tokens in a shared attention pool, but it may also make late-stage refinement less stable. As shown in Fig. 2(b), this instability can manifest as shadow-like blotches and local structural distortions.

To address these limitations, we propose TopoRefine, a training-free SCG framework for controlled identity injection. TopoRefine contains two components: Topology-aware Subject Correspondence (TSC) and Residual Identity Refinement (RIR). TSC augments semantic matching with foreground-graph geodesic relations in a fused Gromov–Wasserstein transport formulation, routing reference identity features to structurally compatible target regions. RIR replaces competitive extended attention with a foreground-gated, magnitude-aligned residual, preserving the target-native attention update while reinforcing subject identity.

We instantiate TopoRefine on both UNet/SDXL and DiT/Qwen backbones. Quantitative and qualitative experiments demonstrate that TopoRefine consistently improves the trade-off between subject consistency and target-driven structural fidelity across both backbones. Our contributions are summarized as follows:

- We formulate training-free SCG as controlled identity injection and identify two key sources of structural interference: semantic-driven correspondence and competitive extended attention.

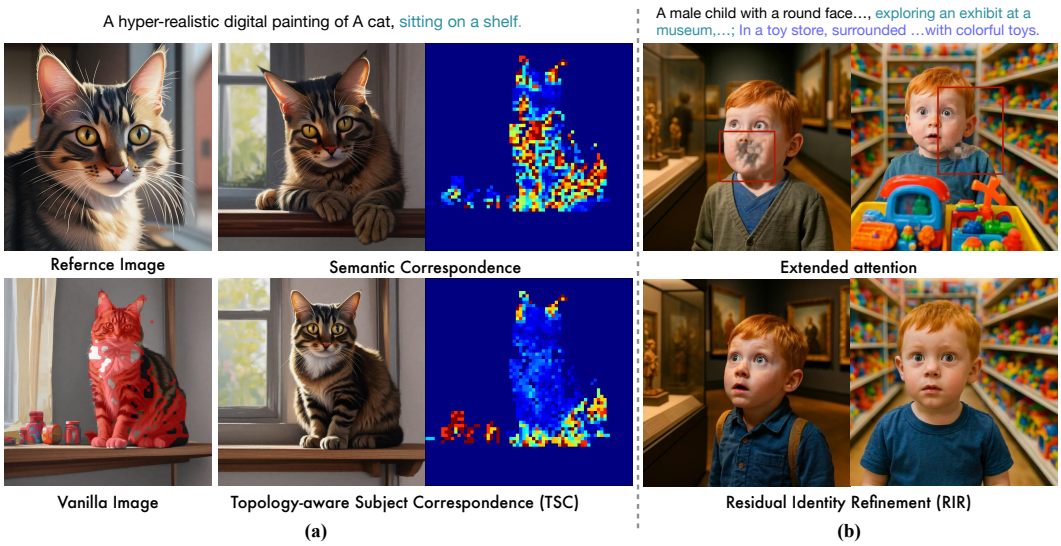


Figure 2: **(a)** Compared with semantic correspondence, TSC better preserves the pose and overall structure of the vanilla image. Local structure error maps show neighborhood-level structural discrepancies, where warmer colors indicate larger errors. TSC yields fewer scattered errors and more structurally compatible identity transfer. **(b)** Extended attention may introduce structural artifacts, such as shadow-like blotches and local corruptions, while RIR preserves cleaner subject structure.

- We propose TSC and RIR, which respectively enable topology-aware identity routing and foreground-gated residual refinement for preserving target-native structure.
- We instantiate TopoRefine on both UNet/SDXL and DiT/Qwen backbones, demonstrating improved subject consistency and pose diversity while maintaining competitive prompt fidelity.

2 Related Work

Personalized and story-based generation. Subject-consistent generation is closely related to text-to-image personalization and story visualization [17, 18, 19, 20, 21]. Personalization methods such as Textual Inversion [22], DreamBooth [23], Custom Diffusion [24], and subsequent encoder- or adapter-based methods [25, 26, 27, 28] adapt pretrained diffusion models to synthesize user-specified subjects from a few reference images. While effective, they typically require per-subject optimization, additional training, or specialized encoders, which limits their scalability for open-domain consistent generation. Story visualization methods [19, 29, 30, 17, 31] also aim to maintain coherent characters across visual narratives, but often rely on task-specific training data or additional generative architectures. Related vision–language research explores image captioning and its in-context configurations [32, 33], the distinctiveness and fidelity of generated descriptions [34], comprehensive 3D scene description [35], and grounded visual reasoning [36]. These studies address visual understanding, while our work focuses on preserving subject identity across text-driven generations. In contrast, our work focuses on training-free subject-consistent generation, where identity is injected by modulating the internal representations of pretrained diffusion models at inference time.

Training-free subject-consistent generation. Training-free SCG methods avoid per-subject fine-tuning and exploit the intrinsic representations of pretrained text-to-image models [29]. 1Prompt1Story [37] leverages the contextual consistency of language models by concatenating multiple prompts into a single input, but prompt-level control alone provides limited spatial control over subject identity. ConsiStory [12] and StoryDiffusion [13] improve consistency through attention-based sharing across generated images. Although such global feature sharing reinforces identity, it can also couple different generations, causing the reference subject to over-constrain target poses, layouts, or local structures. Recent methods therefore move toward correspondence-aware identity transfer. CharaConsist [16] introduces point-tracking attention and adaptive token merging for fine-grained character consistency in DiT-based models, while CoDi [14] performs early correspondence-guided

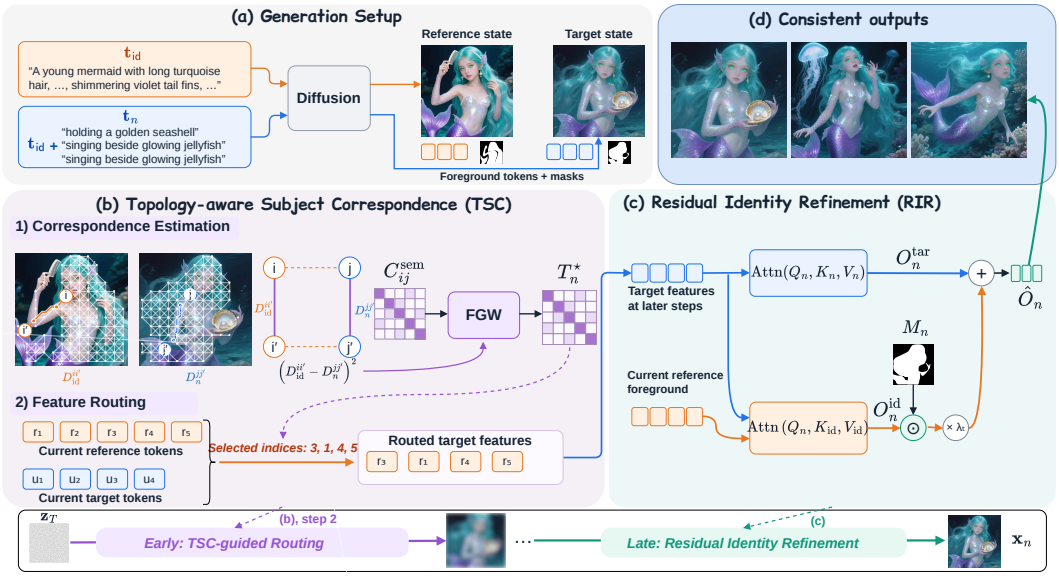


Figure 3: **Overview of** `TopoRefine`**. (a)** Reference and target generations share a subject identity but follow different context prompts. **(b)** Early-stage `TSC` combines semantic similarity and foreground geodesic structure to guide identity feature routing. **(c)** Late-stage `RIR` reinforces identity through a foreground-gated residual while preserving target-native attention. **(d)** Outputs maintain subject identity across diverse poses, actions, and scenes.

identity transport followed by late identity refinement. These methods demonstrate the importance of explicit correspondence for training-free SCG. Related work on multimodal object tracking also explores spatial-temporal modeling [38] and curriculum adaptation [39] to maintain target correspondence under appearance changes. However, existing correspondence mechanisms in training-free SCG are still mainly driven by point-wise semantic similarity, which can be ambiguous under repeated textures, symmetric parts, and non-rigid deformation. Moreover, late-stage extended attention places reference and target tokens in a shared attention pool, where reference identity may compete with the target-native update and perturb the emerging target structure. Our `TopoRefine` addresses these two limitations with `TSC`, which augments semantic matching with subject-internal geodesic relations, and `RIR`, which injects reference identity as a foreground-gated, magnitude-aligned residual.

3 Method

We propose `TopoRefine`, a training-free framework for subject-consistent generation. As illustrated in Fig. 3, the framework consists of Topology-aware Subject Correspondence (`TSC`) for structurally compatible identity routing and Residual Identity Refinement (`RIR`) for controlled late-stage identity injection. `TSC` regularizes correspondence with subject-internal topology, while `RIR` injects reference identity as a localized residual without overriding the target-native generation path.

3.1 Preliminaries

Setup. Subject-consistent generation (SCG) aims to synthesize a set of images that share the same subject identity across diverse contexts. Formally, given N textual prompts $\{\mathbf{t}_n\}_{n=1}^N$, each prompt consists of a shared subject prompt $\mathbf{t}_{\text{id}}$ and a target-specific context prompt $\mathbf{a}_n$, *i.e.*, $\mathbf{t}_n = [\mathbf{t}_{\text{id}}, \mathbf{a}_n]$. A reference image $\mathbf{x}_{\text{id}}$ is generated from $\mathbf{t}_{\text{id}}$, and the goal is to generate target images $\{\mathbf{x}_n\}_{n=1}^N$ that preserve the same subject identity while following their corresponding context prompts $\{\mathbf{a}_n\}_{n=1}^N$.

Foreground subject tokens. Following CoDi [14], we obtain subject foreground masks by averaging the image-text attention maps associated with subject-related text tokens and applying Otsu's thresholding [40]. This produces a reference mask M_{id} and target masks $\{M_n\}_{n=1}^N$.

At a given denoising step and network layer, let X_{id} and X_n denote the hidden spatial feature tokens of the reference and the n-th target image, respectively. Using the binary masks to select foreground locations, we extract the corresponding foreground subject tokens:

$$S_{\mathrm{id}} = X_{\mathrm{id}}[M_{\mathrm{id}}] \in \mathbb{R}^{m\times d}, \qquad S_n = X_n[M_n] \in \mathbb{R}^{l\times d}, \tag{1}$$

where $X[M]$ denotes selecting the spatial tokens at locations where the mask value is one, and m and l are the numbers of foreground tokens in the reference and target masks, respectively. We denote the individual subject tokens by $\{\mathbf{s}_{\mathrm{id}}^i\}_{i=1}^m$ and $\{\mathbf{s}_n^j\}_{j=1}^l$.

3.2 Topology-Aware Subject Correspondence

Given the foreground subject tokens $S_{\mathrm{id}} = \{\mathbf{s}_{\mathrm{id}}^i\}_{i=1}^m$ and $S_n = \{\mathbf{s}_n^j\}_{j=1}^l$, our goal is to establish reference-to-target correspondences that match subject regions with similar appearance and compatible structural roles.

We define the semantic cost between reference and target foreground tokens using cosine distance:

$$C_{ij}^{\mathrm{sem}} = 1 - \frac{(\mathbf{s}_{\mathrm{id}}^i)^\top \mathbf{s}_n^j}{\|\mathbf{s}_{\mathrm{id}}^i\|_2 \|\mathbf{s}_n^j\|_2}. \tag{2}$$

While this feature-similarity term favors locally similar tokens, it is point-wise and does not encode pairwise structural relations within the subject. We therefore introduce a topology term based on foreground geodesic distances, as defined below.

Foreground topology. To describe the internal structure of each subject, we build a graph on the foreground support of its mask. For the reference subject, we denote the graph as $\mathcal{G}_{\mathrm{id}} = (\mathcal{V}_{\mathrm{id}}, \mathcal{E}_{\mathrm{id}})$, where each node $v_{\mathrm{id}}^i \in \mathcal{V}_{\mathrm{id}}$ corresponds to a foreground token $\mathbf{s}_{\mathrm{id}}^i \in S_{\mathrm{id}}$. Two nodes are connected if their corresponding spatial locations are adjacent in the mask. We use an 8-neighbor grid, with edge weight 1 for horizontal/vertical neighbors and $\sqrt{2}$ for diagonal neighbors. The target graph $\mathcal{G}_n = (\mathcal{V}_n, \mathcal{E}_n)$ is constructed analogously from M_n, with node v_n^j corresponding to token $\mathbf{s}_n^j$.

We define the foreground geodesic distance between two subject tokens as the shortest-path distance between their corresponding graph nodes. This yields the pairwise distance matrices

$$D_{\mathrm{id}}^{ii'} = \mathrm{dist}_{\mathcal{G}_{\mathrm{id}}}(v_{\mathrm{id}}^i, v_{\mathrm{id}}^{i'}), \qquad D_n^{jj'} = \mathrm{dist}_{\mathcal{G}_n}(v_n^j, v_n^{j'}), \tag{3}$$

where $D_{\mathrm{id}} \in \mathbb{R}^{m\times m}$ and $D_n \in \mathbb{R}^{l\times l}$. For disconnected foreground masks, unreachable node pairs are assigned twice the maximum finite distance before normalization. We then normalize each distance matrix by its maximum finite value, making the topology term depend on relative foreground geometry rather than the absolute scale of the mask. These matrices describe how foreground tokens are arranged within the subject, so the transport objective favors matches that preserve relative within-subject distances and penalizes mappings that connect similar-looking but structurally distant parts.

Foreground token masses. Not all selected foreground tokens are equally reliable: boundary tokens, weakly activated subject regions, and visually ambiguous parts may provide noisy correspondence evidence. We therefore use the subject-response maps, obtained by averaging image-text attention maps over subject-related tokens, as confidence priors for foreground tokens. Let $A_{\mathrm{id}}(i)$ and $A_n(j)$ denote the corresponding response values at foreground tokens $\mathbf{s}_{\mathrm{id}}^i$ and $\mathbf{s}_n^j$, respectively. We normalize these responses over each foreground support:

$$a_i = \frac{A_{\mathrm{id}}(i)}{\sum_{i'=1}^m A_{\mathrm{id}}(i')}, \qquad b_j = \frac{A_n(j)}{\sum_{j'=1}^l A_n(j')}, \tag{4}$$

yielding token masses $\mathbf{a} = (a_1, \ldots, a_m)$ and $\mathbf{b} = (b_1, \ldots, b_l)$ for the reference and target foreground sets. These masses are used as the marginal weights in the topology-aware transport objective defined below.

Fused topology-aware correspondence. We formulate reference-to-target matching by fusing the semantic token cost with foreground geodesic structure. For each target image n, we solve for a correspondence matrix $T_n \in \mathbb{R}_+^{m\times l}$, where T_{ij} measures the matching strength between reference token $\mathbf{s}_{\mathrm{id}}^i$ and target token $\mathbf{s}_n^j$. Let

$$\Pi(\mathbf{a}, \mathbf{b}) = \left\{ T \in \mathbb{R}_+^{m\times l} \mid T\mathbf{1}_l = \mathbf{a},\ T^\top \mathbf{1}_m = \mathbf{b} \right\} \tag{5}$$

be the feasible set of correspondence matrices, where the row and column constraints enforce the reference and target token masses. We obtain $T_n^\star$ by solving an entropically regularized fused Gromov–Wasserstein objective [41]:

$$T_n^\star = \underset{T\in\Pi(\mathbf{a},\mathbf{b})}{\arg\min}\ (1-\gamma)\sum_{i,j} T_{ij}C_{ij}^{\mathrm{sem}} + \gamma \sum_{i,i',j,j'} \left(D_{\mathrm{id}}^{ii'} - D_n^{jj'}\right)^2 T_{ij}T_{i'j'} + \varepsilon_{\mathrm{ot}}\sum_{i,j} T_{ij}(\log T_{ij} - 1). \tag{6}$$

The semantic term favors feature-similar reference-target token pairs, while the topology term favors correspondences whose pairwise relations are consistent under the reference and target foreground geodesic distances. The coefficient $\gamma \in [0,1]$ balances feature similarity and topology preservation, and $\varepsilon_{\mathrm{ot}}$ controls the entropy regularization.

Correspondence-guided feature routing. The optimal plan $T_n^\star$ defines a topology-aware reference-to-target correspondence. For early-stage foreground feature injection, we first column-normalize the transport plan to obtain a target-conditioned correspondence distribution:

$$P_{ij} = \frac{T_{ij}^\star}{\sum_{i'=1}^{m} T_{i'j}^\star}. \tag{7}$$

For each target foreground token, we select the reference token with the highest normalized correspondence probability:

$$i^\star(j) = \arg\max_i P_{ij}, \qquad c_j = \max_i P_{ij}. \tag{8}$$

If the correspondence confidence c_j is above a threshold τ, we route the selected reference feature to the target token; otherwise, we keep the original target feature. Let $\bar{\mathbf{s}}_n^j$ denote the resulting feature for the j-th target foreground token:

$$\bar{\mathbf{s}}_n^j = \begin{cases} \mathbf{s}_{\mathrm{id}}^{i^\star(j)}, & c_j > \tau, \\ \mathbf{s}_n^j, & \text{otherwise}. \end{cases} \tag{9}$$

This routing is applied in the early transport stage, while late-stage identity details are handled by residual identity refinement.

3.3 Residual Identity Refinement

During identity refinement, the target-driven structure continues to evolve and remains sensitive to reference injection. If the reference signal is coupled too strongly with the target update, it can reduce pose diversity or distort local geometry. A common extended-attention refinement injects reference identity by concatenating the target keys and values with the reference-subject keys and values:

$$O_n^{\mathrm{ext}} = \mathrm{Attn}\left(Q_n, [K_n; K_{\mathrm{id}}], [V_n; V_{\mathrm{id}}]\right), \tag{10}$$

where Q_n, K_n, V_n are projected from the target hidden tokens, and $K_{\mathrm{id}}, V_{\mathrm{id}}$ are projected from the reference foreground subject tokens. This formulation strengthens identity details, but target and reference tokens are normalized in the same attention pool. We instead preserve the target-native update and add reference identity through a foreground-gated residual branch.

Residual identity branch. We first compute the target-native self-attention output:

$$O_n^{\mathrm{tar}} = \mathrm{Attn}(Q_n, K_n, V_n). \tag{11}$$

Then, using the same target queries, we compute an identity residual by attending only to the reference foreground keys and values:

$$O_n^{\mathrm{id}} = \mathrm{Attn}\left(Q_n, K_{\mathrm{id}}, V_{\mathrm{id}}\right). \tag{12}$$

We inject this identity branch only into the target foreground region:

$$\tilde{O}_n = O_n^{\mathrm{tar}} + \lambda_t \left(M_n \odot O_n^{\mathrm{id}}\right), \tag{13}$$

where M_n is the target foreground mask broadcast to the feature dimension, and λ_t controls the residual strength at denoising step t. To keep the residual update stable, we align the magnitude of the injected output with the target-native output:

$$\hat{O}_n = \frac{\|O_n^{\mathrm{tar}}\|_2}{\|\tilde{O}_n\|_2 + \epsilon}\tilde{O}_n. \tag{14}$$

Table 1: **Quantitative comparison** of subject consistency, pose diversity, and prompt fidelity on UNet/SDXL and DiT/Qwen backbones. Best results within each backbone are marked in **bold**. ↑ indicates higher is better, and ↓ indicates lower is better.

Backbone	Method	Subject Consistency			Pose Diversity (↑)	Prompt Fidelity (↑)
		CLIP-I (↑)	DINO-v2 (↑)	DreamSim (↓)		
UNet / SDXL	1Prompt1Story [37]	0.8327	0.8049	0.3453	0.0948	0.8381
	ConsiStory [12]	0.8557	0.8218	0.2774	0.0843	**0.8763**
	StoryDiffusion [13]	0.8578	0.8283	0.2793	0.0932	0.8667
	CoDi [14]	0.8647	0.8406	0.2481	0.0986	0.8715
	`TopoRefine` (ours)	**0.8694**	**0.8428**	**0.2332**	**0.1021**	0.8724
DiT / Qwen	CharaConsist [16]	0.7922	0.7406	0.3903	0.1272	**0.8215**
	`TopoRefine` (ours)	**0.8054**	**0.7558**	**0.3581**	**0.1320**	0.8142

Table 2: **Quantitative component analysis** on Topology-aware Subject Correspondence (`TSC`) and Residual Identity Refinement (`RIR`) across UNet/SDXL and DiT/Qwen backbones. `TSC` improves subject consistency and pose diversity over semantic-driven correspondence. `RIR` further preserves pose diversity and prompt fidelity while maintaining strong subject consistency.

Backbone	`TSC`	`RIR`	Pose Diversity (↑)	Subject Consistency			Prompt Fidelity (↑)
				CLIP-I (↑)	DINO-v2 (↑)	DreamSim (↓)	
UNet/SDXL			0.0986	0.8647	0.8406	0.2481	0.8715
	✓		0.1006	**0.8701**	**0.8431**	**0.2310**	0.8723
	✓	✓	**0.1021**	0.8694	0.8428	0.2332	**0.8724**
DiT/Qwen			0.1146	0.7715	0.7310	0.4200	0.7975
	✓		0.1280	0.8016	0.7512	0.3655	0.8129
	✓	✓	**0.1320**	**0.8054**	**0.7558**	**0.3581**	**0.8142**

Unlike Eq. (10), the target-native output O_n^{tar} remains explicit, and the reference identity is introduced only as a foreground-gated, magnitude-aligned correction.

4 Experiments

4.1 Setup

Evaluation benchmark. We evaluate `TopoRefine` on an evaluation benchmark containing nearly 370 prompt groups, yielding nearly 1,900 generated images. Each group contains prompts for the same subject across different scenes, actions, and compositions. Nearly 50% of the groups depict human subjects, supporting pose diversity evaluation.

Baselines and implementation details. We compare `TopoRefine` against representative SOTA training-free SCG methods on two backbone settings. For the UNet/SDXL setting, we include ConsiStory [12], StoryDiffusion [13], 1Prompt1Story [37], and CoDi [14]. All UNet/SDXL methods are implemented with Stable Diffusion XL 1.0 [42] and generate images at 1024×1024 resolution, except for StoryDiffusion, which is evaluated at 768×768 due to its high memory consumption, following its original setting. For the DiT/Qwen setting, we compare with CharaConsist [16], where both CharaConsist and our DiT instantiation are implemented on Qwen-Image [43]. We reproduce all baselines using their official implementations when available, and implement the Qwen-Image version of CharaConsist following its published design. For fair comparison, we enforce identical noise seeds across methods within each backbone setting, ensuring that the same prompt is initialized from the same random noise. All experiments use 50 denoising steps. The residual refinement weight λ_t is initialized to 0.3 at the beginning of the residual refinement stage and linearly decays to 0 by the 45-th denoising step. All UNet/SDXL experiments are run on NVIDIA RTX A6000 GPUs, whereas all DiT/Qwen experiments are run on NVIDIA A100 GPUs.

Evaluation metrics. Following prior work [12, 13, 37], we evaluate generated images from three complementary perspectives: subject consistency, pose diversity, and prompt fidelity. For subject consistency, we compute the average pairwise image similarity within each prompt group using

CLIP-I [44] and DINO-v2 [45], where higher values indicate stronger identity preservation. We additionally report DreamSim distance [46], where lower values indicate higher perceptual similarity. For pose diversity, we focus on the human-subject subset of the evaluation benchmark and extract 2D body keypoints using ViTPose [47]. To remove global translation, rotation, and scale effects, we align poses by Procrustes analysis [48], and measure diversity as the average Euclidean distance between corresponding keypoints across aligned image pairs. For prompt fidelity, we use CLIP-Score [44] to measure image-text alignment between each generated image and its target prompt.

4.2 Experimental Results

Qualitative comparison. Fig. 4 shows qualitative comparisons on both UNet/SDXL and DiT/Qwen backbones. On UNet/SDXL, existing methods often struggle to jointly preserve subject consistency and target-driven pose diversity. For example, ConsiStory tends to generate highly similar poses in the puppy cases, and its subject consistency degrades when the target pose changes more substantially. StoryDiffusion also exhibits pose over-copying in the boy examples, where the body and head remain in nearly the same position and orientation across prompts. 1Prompt1Story allows more diverse poses, but often weakens subject identity. CoDi better preserves the vanilla target structure in some cases, but can still fail under complex poses, such as the second-row boy example. In contrast, `TopoRefine` more faithfully follows the pose and layout of the vanilla target generation while maintaining strong subject consistency. On DiT/Qwen, CharaConsist shows weaker identity preservation across generated samples, whereas `TopoRefine` achieves more consistent subject identity and better target-driven structure preservation.

Quantitative comparison. Table 1 reports the main quantitative comparison on UNet/SDXL and DiT/Qwen backbones. On UNet/SDXL, `TopoRefine` achieves the best subject consistency, the lowest DreamSim score (0.2332), and the highest pose diversity (0.1021). This indicates that `TopoRefine` improves identity preservation without sacrificing target-driven pose variation. Its prompt fidelity remains competitive, with only a small gap to ConsiStory (0.8724 vs. 0.8763). The same trend holds under the DiT/Qwen setting. Compared with CharaConsist, `TopoRefine` improves DreamSim from 0.3903 to 0.3581. It also increases pose diversity from 0.1272 to 0.1320, while maintaining comparable prompt fidelity. These results indicate that topology-aware correspondence and residual identity refinement are not tied to a specific diffusion architecture, but generalize from UNet/SDXL to DiT/Qwen generation.

4.3 Ablation Studies

Quantitative component analysis. Table 2 evaluates `TSC` and `RIR` on both UNet/SDXL and DiT/Qwen. The unchecked baseline uses semantic correspondence with extended attention. Replacing semantic correspondence with `TSC` consistently improves subject consistency and pose diversity, showing that subject-internal geodesic relations provide more structurally compatible identity transfer. For example, `TSC` improves pose diversity from 0.0986 to 0.1006 on UNet/SDXL and from 0.1146 to 0.1280 on DiT/Qwen, while also improving CLIP-I and DreamSim on both backbones. Adding `RIR` further improves the overall trade-off. It achieves the best pose diversity and prompt fidelity on UNet/SDXL, and improves all metrics over `TSC` alone on DiT/Qwen. These results indicate that foreground-gated residual refinement provides a more stable late-stage identity injection than competitive extended attention, preserving target-driven pose and prompt alignment while maintaining strong subject consistency.

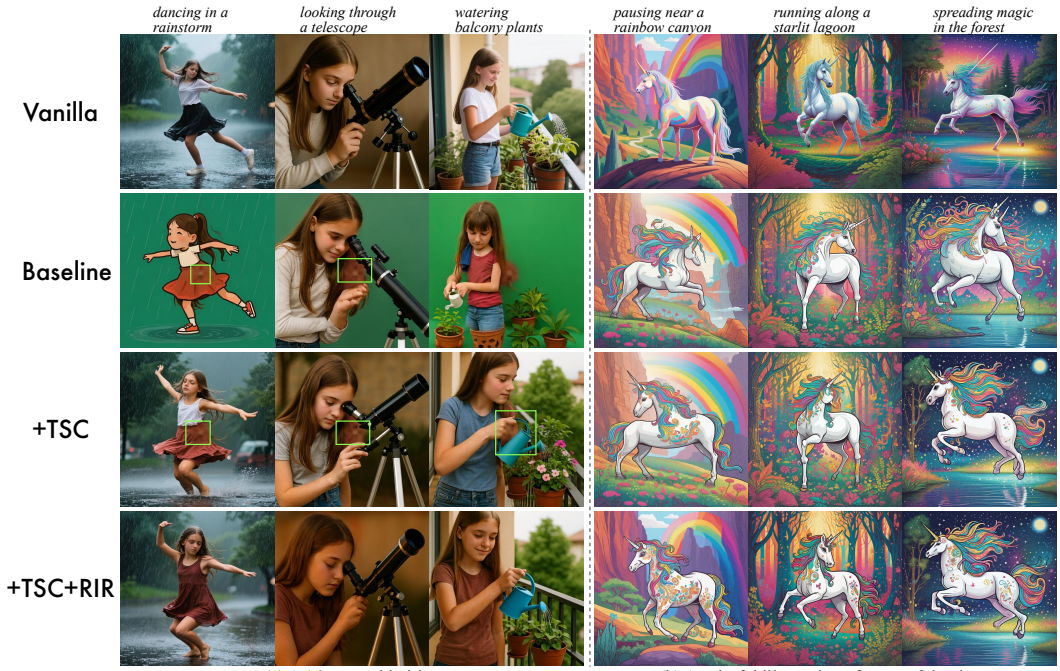
Qualitative component analysis. As shown in Fig. 5, the baseline with semantic correspondence and extended attention often fails to preserve the target structure produced by the vanilla model, indicating that identity injection can perturb the target-driven pose and layout on both UNet/SDXL and DiT/Qwen. On DiT/Qwen, this structural disruption is more severe, causing corrupted backgrounds and shadow-like artifacts. Replacing semantic correspondence with `TSC` substantially improves alignment with the vanilla structure: the generated subject better follows the target-driven pose, and the background corruption on DiT/Qwen is largely alleviated. This confirms that topology-aware subject correspondence enables identity transfer between structurally compatible regions. Further replacing extended attention with `RIR` improves structural alignment with the vanilla generation and removes the shadow artifacts introduced by extended attention, especially on DiT/Qwen. Overall, `TSC` and `RIR` jointly improve pose diversity and prompt fidelity while maintaining strong subject consistency.



(a) A puppy

(b) A teenage boy with short, spiky black hair, a slight build, and dark brown eyes

Figure 4: **Qualitative comparison with current SCG methods.** Existing methods often struggle to jointly preserve subject identity and target-driven pose diversity. In contrast, `TopoRefine` maintains consistent subject identity across prompts while better preserving diverse poses, actions, and layouts.



(a) A 16 years old girl (b) A colorful illustration of a graceful unicorn

Figure 5: **Qualitative component analysis** across UNet/SDXL and DiT/Qwen backbones. Baseline uses semantic correspondence with extended attention. `TSC` replaces semantic correspondence with topology-aware subject correspondence, while `RIR` replaces extended attention with residual identity refinement. Together, they improve pose diversity and prompt fidelity while maintaining strong subject consistency.

5 Conclusion

We presented `TopoRefine`, a training-free SCG framework that improves identity injection through topology-aware correspondence and residual refinement. `TSC` augments semantic matching with subject-internal geodesic relations, enabling identity transfer between structurally compatible regions. `RIR` preserves the target-native attention update and injects reference identity as a foreground-gated, magnitude-aligned residual, reducing interference with the emerging target structure. Across UNet/SDXL and DiT/Qwen backbones, `TopoRefine` consistently improves subject consistency, pose diversity, and prompt fidelity. These results highlight topology-aware routing and residual identity injection as effective design principles for training-free subject-consistent generation.

6 Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62406135), the Natural Science Foundation of Jiangsu Province (No. BK20241198), the Gusu Innovation and Entrepreneur Leading Talents Program (No. ZXL2024362), the Jiangsu Province Graduate Research and Practice Innovation Program (No. 26CXAI25), and the vivo Research Fund.

References

[1] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[2] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022.

[3] Zhennan Chen, Yajie Li, Haofan Wang, Zhibo Chen, Zhengkai Jiang, Jun Li, Qian Wang, Jian Yang, and Ying Tai. Ragd: Regional-aware diffusion model for text-to-image generation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19331–19341. IEEE, 2025.

[4] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[5] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, 2023.

[6] Zhennan Chen, Junwei Zhu, Xu Chen, Jiangning Zhang, Jiawei Chen, Zhuoqi Zeng, Wei Zhang, Chengjie Wang, Jian Yang, and Ying Tai. L2p: Unlocking latent potential for pixel generation. *arXiv preprint arXiv:2605.12013*, 2026.

[7] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *AAAI*, 2024.

[8] Zhennan Chen, Junwei Zhu, Xu Chen, Jiangning Zhang, Xiaobin Hu, Hanzhen Zhao, Chengjie Wang, Jian Yang, and Ying Tai. Dip: Taming diffusion models in pixel space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 36136–36146, 2026.

[9] Zhennan Chen, Tianxing Shi, Pengcheng Xu, Kepan Nan, Qian Wang, Zili Yi, Jian Yang, and Ying Tai. Synvar: Synergizing spatial and semantic alignment in visual autoregressive model. In *European Conference on Computer Vision*, pages 201–218. Springer, 2026.

[10] Dawid J Kopiczko, Tijmen Blankevoort, and Yuki M Asano. Vera: Vector-based random matrix adaptation. In *ICLR*, 2024.

[11] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

[12] Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. Training-free consistent text-to-image generation. *SIGGRAPH*, 2024.

[13] Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. In *NeurIPS*, 2024.

[14] Zhanxin Gao, Beier Zhu, Jian Yang, Ying Tai, et al. Codi: Subject-consistent and pose-diverse text-to-image generation. In *ICLR*, 2026.

[15] Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention. In *CVPR*, 2024.

[16] Mengyu Wang, Henghui Ding, Jianing Peng, Yao Zhao, Yunpeng Chen, and Yunchao Wei. Characonsist: Fine-grained consistent character generation. In *ICCV*, 2025.

[17] Shuai Yang, Yuying Ge, Yang Li, Yukang Chen, Yixiao Ge, Ying Shan, and Yingcong Chen. Seed-story: Multimodal long story generation with large language model. *arXiv preprint arXiv:2407.08683*, 2024.

[18] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *CVPR*, 2024.

[19] Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. Storygan: A sequential conditional gan for story visualization. In *CVPR*, 2019.

[20] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

[21] Chang Liu, Haoning Wu, Yujie Zhong, Xiaoyun Zhang, Yanfeng Wang, and Weidi Xie. Intelligent grimm-open-ended visual storytelling via latent diffusion models. In *CVPR*, 2024.

[22] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2023.

[23] 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 *CVPR*, 2023.

[24] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023.

[25] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on graphics (TOG)*, 42(1):1–13, 2022.

[26] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *IJCV*, 2024.

[27] Yu Zeng, Vishal M Patel, Haochen Wang, Xun Huang, Ting-Chun Wang, Ming-Yu Liu, and Yogesh Balaji. Jedi: Joint-image diffusion models for finetuning-free personalized text-to-image generation. In *CVPR*, 2024.

[28] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *IJCV*, 2024.

[29] Adyasha Maharana, Darryl Hannan, and Mohit Bansal. Storydall-e: Adapting pretrained text-to-image transformers for story continuation. In *ECCV*. Springer, 2022.

[30] Tanzila Rahman, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, Shweta Mahajan, and Leonid Sigal. Make-a-story: Visual memory conditioned consistent story generation. In *CVPR*, 2023.

[31] Zilyu Ye, Jinxiu Liu, JinJin Cao, Zhiyang Chen, Ziwei Xuan, Mingyuan Zhou, Qi Liu, and Guo-Jun Qi. Openstory: A large-scale open-domain dataset for subject-driven visual storytelling. In *CVPR*, 2024.

[32] Kun-Yu Lin, Hongjun Wang, Weining Ren, and Kai Han. Panoptic captioning: An equivalence bridge for image and text. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems*, 2025.

[33] Xu Yang, Yongliang Wu, Mingzhuo Yang, Haokun Chen, and Xin Geng. Exploring diverse in-context configurations for image captioning. *Advances in Neural Information Processing Systems*, 36:40924–40943, 2023.

[34] Yuhang Huang, Zihan Wu, Chongyang Gao, Jiawei Peng, and Xu Yang. Exploring the distinctiveness and fidelity of the descriptions generated by large vision-language models. *arXiv preprint arXiv:2404.17534*, 2024.

[35] Kun-Yu Lin, Chengke Bu, Zhenguo Li, and Kai Han. Holo-captioning: A comprehensive textual view of 3d scenes. In *European Conference on Computer Vision*, 2026.

[36] An-Lan Wang, Guozhi Tang, Lei Liao, Hanshen Zhu, Kai Huang, Jingqun Tang, Jiaming Zhou, and Kun-Yu Lin. Reason with Thumbnails, Answer with Focus: An efficient and effective paradigm for multimodal grounded visual reasoning. In *International Conference on Machine Learning*, 2026.

[37] Tao Liu, Kai Wang, Senmao Li, Joost van de Weijer, Fahad Shahbaz Khan, Shiqi Yang, Yaxing Wang, Jian Yang, and Ming-Ming Cheng. One-prompt-one-story: Free-lunch consistent text-to-image generation using a single prompt. In *ICLR*, 2025.

[38] Xiantao Hu, Ying Tai, Xu Zhao, Chen Zhao, Zhenyu Zhang, Jun Li, Bineng Zhong, and Jian Yang. Exploiting multimodal spatial-temporal patterns for video object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 3581–3589, 2025.

[39] Xiantao Hu, Fansheng Zeng, Bineng Zhong, Zhangyong Tang, Wenxuan Fang, Jun Li, Ying Tai, and Jian Yang. Curriculum adaptation for one-stream rgb–t tracking. *Pattern Recognition*, page 113494, 2026.

[40] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979.

[41] Titouan Vayer, Laetitia Chapel, Rémi Flamary, Romain Tavenard, and Nicolas Courty. Fused gromov-wasserstein distance for structured objects. *Algorithms*, 13(9):212, 2020.

[42] 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. *arXiv preprint arXiv:2307.01952*, 2023.

[43] 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.

[44] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *EMNLP*, 2021.

[45] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[46] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In *NeurIPS*, 2023.

[47] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. In *NeurIPS*, 2022.

[48] Peter H Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966.

A Human Preference Study

Table 3 reports the human preference results across seven methods. `TopoRefine`-Qwen and `TopoRefine`-SDXL receive the highest and second-highest preference shares, respectively. Compared with CharaConsist, the strongest baseline in this study, their preference shares are higher by 9.41 and 1.47 percentage points. These results complement the automated evaluations in Section 4.2.

Table 3: **Human preference results.** Preference shares across the seven evaluated methods.

Method	Preference (%) ↑
1Prompt1Story	5.15
ConsiStory	8.97
StoryDiffusion	7.94
CoDi	14.56
CharaConsist	17.50
`TopoRefine`-SDXL	18.97
`TopoRefine`-Qwen	**26.91**

B Computational Cost

B.1 Inference Time and GPU Memory

We measure wall-clock inference time and peak GPU memory for five samples generated from one prompt group at 1024×1024 resolution. Settings are identical within each backbone: UNet/SDXL uses one NVIDIA RTX A6000 GPU, while DiT/Qwen uses one NVIDIA A100 GPU. Table 4 reports the results. On UNet/SDXL, `TopoRefine` is faster than CoDi (151.19 s versus 169.56 s), with similar peak GPU memory. On DiT/Qwen, it takes slightly longer than CharaConsist (551.36 s versus 513.92 s), again with similar memory usage. Times across the two backbone groups should be interpreted in the context of their different models and hardware.

Following CoDi and CharaConsist, a vanilla pass extracts correspondence features. FGW is then solved once per target to obtain $T_n^\star$, which is reused throughout `TSC`, rather than recomputed at every denoising step.

Table 4: **Inference cost.** Wall-clock time and peak GPU memory for five samples from one prompt group at 1024×1024.

Backbone	Method	Time (s)	Peak GPU memory (GB)
UNet/SDXL	Vanilla SDXL	61.19	36.782
	1Prompt1Story	110.73	17.231
	ConsiStory	120.76	46.675
	CoDi	169.56	45.208
	`TopoRefine`	151.19	45.561
DiT/Qwen	Vanilla Qwen	437.55	58.543
	CharaConsist	513.92	60.873
	`TopoRefine`	551.36	60.848

C Comparison with FLUX.1 Kontext and Nano Banana Pro

We compare `TopoRefine` with FLUX.1 Kontext [dev] and Nano Banana Pro using a shared reference image. We first generate the reference image with Qwen-Image from the identity prompt and then provide this image as input to both editing models. For these models, each target prompt is rewritten as an editing instruction that specifies the new action and scene while explicitly requesting preservation of the reference subject's identity. For example, the target prompt

> "A young woman with shoulder-length auburn hair, delicate facial features, an athletic figure, warm olive skin, and deep green eyes playing the violin on a candlelit stage."

is reformulated as

"The same woman from the reference image is now playing the violin on a candlelit stage. Preserve her facial identity and distinctive appearance while naturally adapting her pose, expression, and surroundings to the new scene."

Fig. 6 presents the qualitative comparison. `TopoRefine` achieves strong subject consistency, scene diversity, and prompt adherence, with results comparable to Nano Banana Pro in these examples. Compared with FLUX.1 Kontext [dev], our method better preserves the reference subject's appearance, whereas FLUX.1 Kontext [dev] exhibits less stable identity preservation, including shifts in skin tone and sharper facial features than those in the reference.

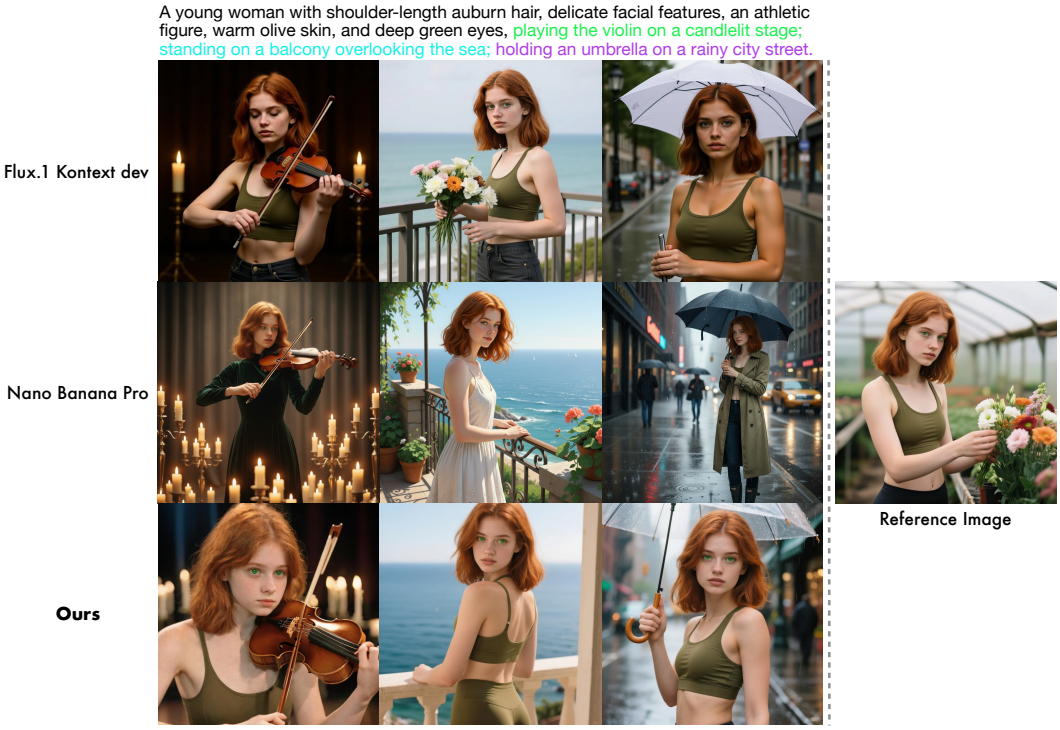


Figure 6: **Qualitative comparison with FLUX.1 Kontext [dev] and Nano Banana Pro.** `TopoRefine` achieves strong subject consistency, scene diversity, and prompt adherence, yielding results competitive with Nano Banana Pro and superior to FLUX.1 Kontext [dev] in this qualitative comparison.

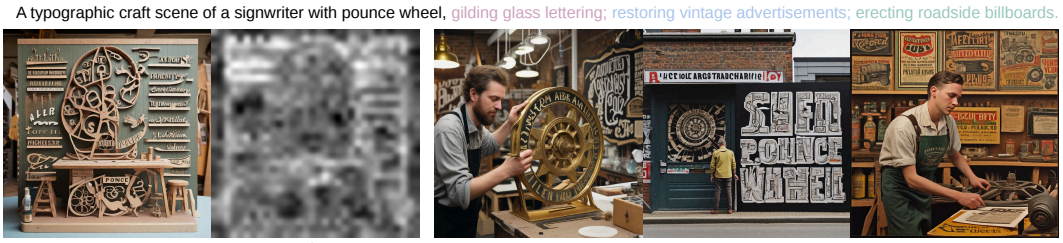


Figure 7: **Failure case: missing subject in the reference.** The reference prompt requests a signwriter, but the generated reference contains no clearly visible craftsperson. We show the reference image, its subject-response map, and three target images generated by `TopoRefine`. The response map is normalized for visualization. Without valid reference-subject identity cues, subject consistency cannot be guaranteed.

D Limitations

Mask failures and missing reference subjects. `TopoRefine` relies on attention-derived foreground masks to select subject tokens, estimate correspondence weights, and gate residual identity injection. When these masks select non-subject regions, errors can affect both `TSC` correspondence and `RIR` refinement because they share the same attention prior. Fig. 7 illustrates an extreme case: the reference omits the requested signwriter, leaving no valid subject identity to extract despite a nonempty foreground mask. Without reliable reference identity cues, subject consistency cannot be guaranteed.

Overlapping masks in multi-subject scenes. Subject-specific attention can help distinguish subjects even when their foreground masks overlap. Moderate overlap need not cause obvious subject mixing (Fig. 8(a)). However, extensive overlap and inaccurate subject selection can make subject-specific correspondence and



Figure 8: **Multi-subject mask overlap.** Each row shows the vanilla target, the girl and dog masks, an overlap visualization, and the TopoRefine output. Green and orange indicate regions selected only by the girl and dog masks, respectively; magenta indicates their intersection. (a) Moderate overlap does not cause obvious subject mixing. (b) Extensive overlap and inaccurate subject selection accompany changes in pose and relative subject scale, with locally entangled body structures.

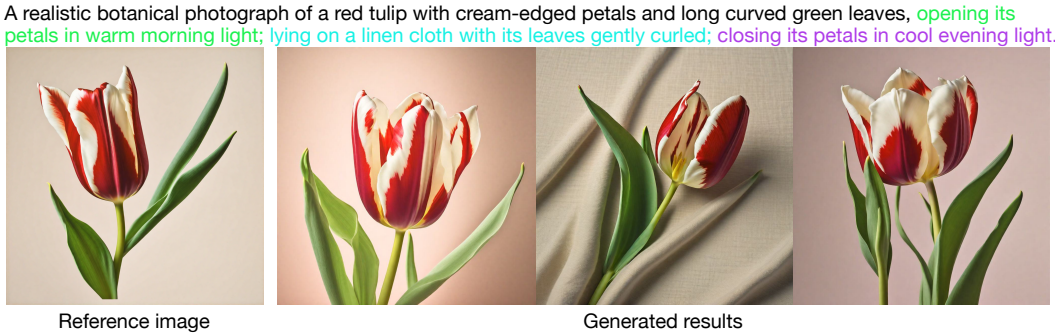


Figure 9: **Non-rigid subject generation.** We show a reference tulip and three target images generated by `TopoRefine`. The results retain the characteristic red-and-cream petal appearance across changes in petal configuration and stem curvature, illustrating applicability to non-rigid plant subjects.

identity refinement less reliable. Fig. 8(b) illustrates changes in pose and relative subject scale, together with locally entangled body structures.

Applicability to non-rigid subjects. `TSC` combines semantic features with relative foreground structure, allowing correspondence estimation under moderate non-rigid deformation. Fig. 9 illustrates this applicability beyond human and animal subjects: the generated tulips retain the reference's red-and-cream petal appearance despite changes in petal configuration and stem curvature.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: abstract and introduction clearly present the core contributions.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper explicitly discusses the limitations, application scope and underlying assumptions of the proposed method.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: This paper focuses on empirical experiments and does not contain theoretical results or mathematical proofs.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.

- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All experimental settingsare clearly described.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The source code are provided in the supplementary material.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All hyperparameters are fully specified in the paper and supplementary material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: This work focuses on performance comparison under standard experimental settings.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The basic computational resources are described in the paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research strictly follows the NeurIPS Code of Ethics with no ethical concerns or violations.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This paper focuses on fundamental algorithm design and theoretical improvement, and has no direct obvious societal impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The proposed model and datasets in this work have no high risk of misuse, and no additional safeguards are required.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All adopted datasets, baseline codes and pre-trained models are properly cited and their original licenses are respected.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We will open-source our implementation with accompanying documentation describing its usage and experimental settings.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This work does not contain crowdsourcing experiments or any research involving human participants.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [No]

Justification: This paper does not involve human subjects or user study, so IRB approval is not required.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not adopted as important or original components of the core method in this work.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.