Setting the Stage: Text-Driven Scene-Consistent Image Generation

Cong Xie*, Che Wang*, Yan Zhang†, Ruiqi Yu, Han Zou, Zheng Pan, and Zhenpeng Zhan†

Baidu Inc., China
{zhangyan97,zhanzhenpeng01}@baidu.com

Fig. 1: Scene Staging: Text-driven scene-consistent image generation. Images labeled with the subscript "scene" are reference scene images used to guide generation; the remaining images are our results, which depict diverse viewpoints and compositions while faithfully following the textual instructions and preserving the reference scene identity, rather than merely copying the reference.

Abstract. We focus on the foundational task of Scene Staging: given a reference scene image and a text condition specifying an actor category to be generated in the scene and its spatial relation to the scene, the goal is to synthesize an output image that preserves the same scene identity as the reference image while correctly generating the actor according to the spatial relation described in the text. Existing methods struggle with this task, largely due to the scarcity of high-quality paired data and unconstrained generation objectives. To overcome the data bottleneck, we propose a novel data construction pipeline that combines real-world photographs, entity removal, and image-to-video diffusion models to generate training pairs with diverse scenes, viewpoints and correct entity-scene relationships. Furthermore, we introduce a novel correspondence-guided

* Equal contribution.
† Corresponding author.

attention loss that leverages cross-view cues to enforce spatial alignment with the reference scene. Experiments on our scene-consistent benchmark show that our approach achieves better scene alignment and text–image alignment than state-of-the-art baselines, according to both automatic metrics and human preference studies. Our method generates images with diverse viewpoints and compositions while faithfully following the textual instructions and preserving the reference scene identity.

Keywords: Scene Consistency · Text to Image Generation · Image Editing

1 Introduction

The demand for visual storytelling, such as comic generation and video storyboarding, has surged in recent years. In the production of visual stories, the creation of narrative imagery can be viewed as a two-step process: first, setting the stage (the background scene), and second, placing the actors (character or object). Existing image editing models [12,20,25,26] have made significant progress in local image editing, such as inserting or replacing entities while largely preserving the overall image composition. Therefore, they are well suited for the second step: given an already established stage image that does not require major compositional changes, the reference actor can be inserted or replaced accordingly.

In this work, we focus on the stage setting step: generating an image conditioned on a reference scene image and textual instructions that specify the actor category (e.g., a man or a book) and its spatial relationship within the scene. In visual storytelling, scenes often need to be depicted from different viewpoints or compositions with respect to the actor, which introduces two challenges: (1) instruction following: the generated image should adapt the reference scene to the desired viewpoint and composition specified by the text rather than simply replicate it; (2) scene consistency: the generated images should preserve the same scene identity as the reference to maintain narrative coherence.

Existing methods [1, 8, 20, 25, 28], struggle to balance scene consistency with instruction following, either replicating the reference scene with high consistency but poor responsiveness to the prompt, or prioritizing prompt compliance at the expense of scene consistency. Such limitations largely stem from both data availability and training objectives. From a data perspective, effective training pairs for this task should satisfy **three key properties**: (1) the input and target images share the same scene identity but differ in viewpoint and composition; (e.g., the text may request a top-down view of the scene even when the reference image is a low-angle view, as in the lighthouse example in Figure 1) (2) the actor is absent from the input image but correctly placed in the target image; (3) the training data contains a large and diverse set of scenes.

It is non-trivial to construct a training data fulfilling all the three properties described above. Existing multi-view scene datasets, such as DL3DV-10K [13] , only satisfy Property (1). One may attempt to place actors into the target-view

image using inpainting or image editing models. However, the generated results often suffer from scale mismatch, lighting inconsistency, and blending artifacts along object boundaries. Such artifacts break the correct entity–scene relationships and degrade the quality of supervision. Another option is to mine frame pairs from web videos. However, shot changes and scene transitions frequently result in frame pairs that do not depict the same scene at all. Restricting the sampling to adjacent frames can preserve scene identity, but it significantly reduces the viewpoint baseline, yielding pairs with nearly identical spatial layouts. As a result, such data fails to teach the model how to maintain scene consistency under meaningful viewpoint or compositional changes.

To overcome the limitations of existing datasets, we introduce a scene-consistent data construction pipeline designed to fulfill all three properties. Specifically, we directly use real-world photographs as training targets to satisfy Properties 2 and 3. We apply Entity Removal to obtain clean background images while preserving realistic scene layouts, further supporting Property 2. Finally, we leverage video generation models to synthesize spatially diverse yet scene-consistent views, which enables us to satisfy Property 1. Building on these data, we further propose a novel correspondence-guided attention loss to explicitly optimize the network for scene preservation. To address the unconstrained nature of standard diffusion models during cross-view generation, we leverage cross-view correspondence cues to directly supervise the intermediate joint-attention layers of the diffusion backbone. By regularizing the attention score matrix, this loss forces tokens that correspond to the same scene location across different views to allocate higher attention weights to each other. Consequently, by encouraging the joint attention to place high mass on corresponding scene regions, this explicit supervision compels the model to preserve the underlying scene structure and appearance, even as the viewpoint and composition undergo significant text-driven modifications.

Our contributions are summarized as follows:

- We propose a novel and scalable scene-consistent data generation pipeline. This pipeline overcomes the key obstacles in constructing high-quality training pairs that guarantee large scene diversity, correct entity-scene relationships, and rich viewpoint variations.
- We introduce a novel correspondence-guided attention loss that regularizes the attention score matrix, forcing tokens corresponding to the same scene location across different views to allocate higher attention weights to each other. This mechanism guarantees robust scene preservation even when the narrative introduces significant text-driven viewpoint and composition shifts.
- We demonstrate that our method consistently outperforms existing baselines, achieving a superior balance between accurate text instruction following and scene consistency. Extensive experiments show that the generated images faithfully reflect the textual instructions while preserving the given reference scene identity. This advantage is validated through both automatic quantitative metrics and comprehensive human evaluations.

2 Related Works

Text-Driven Image Editing. Existing image editing models [12, 20, 26] are typically designed as multi-task, instruction-following general frameworks. For example, FLUX.1 Kontext [9] unifies image generation and editing within a single flow matching model via sequence concatenation in the latent space. ACE++ [15] introduces an instruction-based diffusion framework that handles diverse editing tasks through context-aware content filling. However, these versatile frameworks lack explicit training for scene-oriented tasks. Consequently, they frequently encounter a severe trade-off between environment preservation and instruction adherence: they tend to either rigidly reconstruct the original scene while failing to reflect the text prompt, or aggressively execute semantic modifications at the expense of the underlying scene's structural consistency.

Image Inpainting. Local image inpainting [1, 28] primarily targets filling missing regions or adding local content to an existing image. While early deep learning approaches like MAT [10] utilized transformers for large hole filling, they specifically customized the attention mechanism to aggregate non-local information only from valid tokens, enabling efficient high-resolution processing. With the advent of diffusion models, text-guided inpainting introduced much richer semantic controllability. For instance, SmartBrush [27] incorporates both text and shape guidance, predicting object masks during synthesis to explicitly preserve the surrounding background. To further enhance generation quality and applicability, BrushNet [8] proposes a plug-and-play, decomposed dual-branch diffusion architecture. By explicitly separating masked image feature extraction from the noisy latent generation, it achieves dense per-pixel constraints and maintains strong structural fidelity. However, these methods inherently rely on fixed spatial layouts. Consequently, when a narrative demands global shifts in perspective or composition, they cannot reliably transform the overall scene while preserving the underlying environment.

Consistent Image Generation. Consistent image generation has witnessed rapid development in recent years. A multitude of methods, ranging from adapter-based conditional models (e.g., IP-Adapter [29], InstantID [23], PhotoMaker [11]) and attention-tuning mechanisms (e.g., StoryDiffusion [31], StoryMaker [32]) to efficient autoregressive frameworks (e.g., Infinite-Story [17]), have been proposed to maintain fine-grained character attributes and global styles across multiple generated frames. While extensive research has successfully addressed character (actor) consistency, scene (stage) consistency remains a significant bottleneck. SceneDecorator [21] emerges as the most relevant baseline, employing a training-free global-to-local planning strategy. However, such training-free attention sharing relies heavily on raw visual feature similarity. It degrades significantly when the narrative introduces substantial viewpoint or layout changes, as the un-tuned attention mechanisms fail to align drastically altered visual features. In contrast, our method focuses on the foundational task of Scene Staging: seamlessly embedding a text-specified entity into a reference scene under dynamic viewpoint and composition changes, while accurately reflecting spatial relationships and preserving the original structural integrity of the environment.

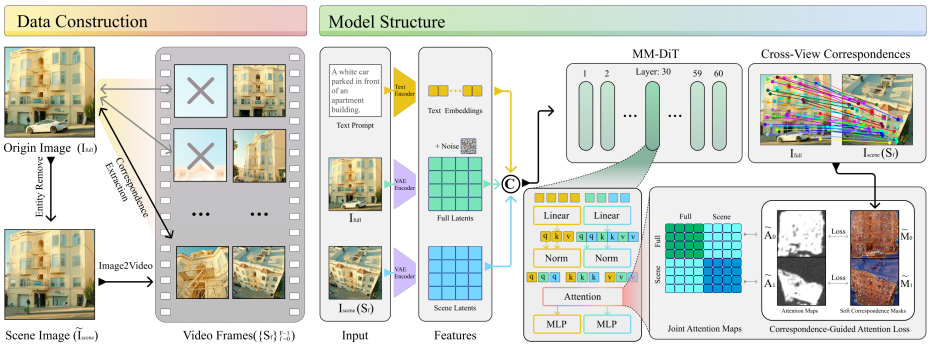

Fig. 2: Overall framework of our method with two contributions: (a) Scene-Consistent Data Construction Pipeline, generating high-quality, scene-consistent paired data with large scene diversity, correct entity-scene relationships, and rich viewpoint variations; (b) Correspondence-Guided Attention Loss, encouraging attention to focus on spatially corresponding regions to better preserve scene structure.

3 Methods

3.1 Overview

Given a reference scene image I_{scene} and a text condition c that specifies both the actor category (e.g., a man or a book) to be placed in the scene and its spatial relationship to the scene, our goal is to generate a scene-consistent output image I_{out} that (i) depicts the same scene identity as I_{scene}, and (ii) correctly generates the actor according to the spatial relation described in the text. We denote the generative model as

$$I_{\text{out}} = f_\theta(I_{\text{scene}}, c), \tag{1}$$

where f_θ is implemented as a diffusion-based image generator conditioned on both the reference scene and the text.

As illustrated in Figure 2, our method has two components. First, we build a scene-consistent data construction pipeline that converts real world images into multi-view tuples comprising entity-removed scenes, synthesized scene views, and soft correspondence masks. Second, during training we apply a correspondence-guided attention loss that uses these correspondence cues to regularize intermediate attention maps, encouraging attention to focus on spatially corresponding regions across views and thereby better preserving scene structure.

3.2 Scene-Consistent Data Construction

As described in the Introduction, ideal training data should fulfill three properties. Next, we introduce a data construction pipeline that satisfies all these properties.

Real-world Photographs as Supervision to Fulfill Properties 2 and 3. We start by collecting real-world photographs containing humans and generic

objects in diverse scenes. These images I_{full} do not need to be paired multi-view observations; single-view photographs are sufficient. Such data is easy to collect at scale, enabling a large-scale and diverse set of scenes to satisfy Property 3. Moreover, since these are real photographs, the objects are naturally and correctly placed in the scene, which partially satisfies Property 2.

Entity Removal to Fulfill Property 2. For each image I_{full}, we automatically erase a target entity region (person or object) to obtain an entity-removed scene $\tilde{I}_{\text{scene}}$. For human-centric photos, we apply FLUX.1 Kontext Dev [9] for person removal, and additionally apply filtering techniques such as face detection [6] and human segmentation [14, 18] to discard low-quality results and maintain high-quality entity-removed scenes. For non-human objects, we use open-source image-editing datasets OmniEdit [24] that provide object-removal image pairs. The resulting $\tilde{I}_{\text{scene}}$ images serve as realistic scene-only inputs for the subsequent multi-view synthesis stage.

Multi-View Scene Generation via Image-to-Video to Fulfill Property 1. Given $\tilde{I}_{\text{scene}}$, we employ a pre-trained image-to-video diffusion model $\mathcal{G}_{\text{i2v}}$ (Wan 2.2 [22]) to synthesize a short video sequence

$$\{S_f\}_{f=0}^{F-1} = \mathcal{G}_{\text{i2v}}(\tilde{I}_{\text{scene}}), \tag{2}$$

where F is the total number of frames, and each frame S_f corresponds to the same underlying scene under a slightly different camera motion.

Benefiting from large-scale training on real-world videos, the image-to-video diffusion model $\mathcal{G}_{\text{i2v}}$ can transform a single high-quality photograph into a short clip with spatial coherence and smooth camera motion. In our pipeline, this allows any still image to be converted into a compact multi-view sequence, greatly expanding the pool of training scenes and naturally enriching scene viewpoints.

Instead of pairing with all synthesized frames, we first filter out those exhibiting severe video hallucinations or structural drift. By evaluating the robust inliers obtained from the feature matching network [19], we discard low-quality outputs and pair the source image I_{full} only with one of the surviving high-quality frames S_f to form a scene pair (I_{full}, S_f). In our training formulation, S_f plays the role of I_{scene}, while I_{full} serves as a proxy for I_{out}.

Cross-View Correspondence Extraction. To obtain explicit scene-level correspondence cues between I_{full} and the selected scene view S_f, we apply a pre-trained feature matching network [19] that estimates sparse correspondences between two views of the same scene under viewpoint changes. Given the pair (I_{full}, S_f), the matcher outputs

$$\{(x_i^0, y_i^0)\}_{i=1}^{N}, \quad \{(x_i^1, y_i^1)\}_{i=1}^{N}, \tag{3}$$

where (x_i^0, y_i^0) and (x_i^1, y_i^1) denote coordinates of matched points in I_{full} and S_f, respectively.

We convert these sparse matches into dense, soft correspondence masks. Let $H \times W$ be the image resolution and $K \in \mathbb{R}^{(2r+1)\times(2r+1)}$ be a radial kernel (e.g., Gaussian) whose values decay with distance from the center. For each match $((x_i^0, y_i^0), (x_i^1, y_i^1))$, we add a copy of K centered at (x_i^0, y_i^0) and (x_i^1, y_i^1) onto

masks $M_0, M_1 \in \mathbb{R}^{H \times W}$, respectively, accumulating overlapping contributions and clipping them to a fixed range. We then crop and downsample M_0 and M_1 to the spatial resolution of the latent feature maps used for attention supervision, obtaining

$$\tilde{M}_0, \tilde{M}_1 \in \mathbb{R}^{H_s \times W_s}. \tag{4}$$

High responses in $\tilde{M}_0$ and $\tilde{M}_1$ highlight regions that are spatially corresponding between the two views.

3.3 Correspondence-Guided Attention Loss

We build on the state-of-the-art image editing model Qwen-Image-Edit [25] and augment it with correspondence-guided attention supervision.

MMDiT Backbone. Our model adopts the double-stream Multimodal Diffusion Transformer (MMDiT) architecture of Qwen-Image-Edit, jointly modeling noise and image latents under text conditioning. Given a noisy latent of the target image I_{full} (proxy for I_{out}), a reference scene image I_{scene}, and a text prompt c describing the inserted entity, the model predicts the added noise by using two streams: one for image tokens, and another for conditioning tokens (which jointly encode the scene and the text).

We keep the variational autoencoder (VAE), text encoder, and backbone weights of Qwen Image Edit frozen, and only introduce a small number of trainable parameters via lightweight adapters. This parameter-efficient fine-tuning preserves the strong priors of the base model while enabling specialization to our scene staging objective and the proposed correspondence-guided attention regularization.

Correspondence-Guided Attention Loss. Given the source image I_{full} (denoted as view 0) and the scene condition S_f (denoted as view 1), we extract their visual tokens from an intermediate attention layer of MMDiT, $z_0 \in \mathbb{R}^{L_0 \times d}$, $z_1 \in \mathbb{R}^{L_1 \times d}$, where L_0 and L_1 denote the number of tokens for view 0 and view 1, respectively, and d is the token feature dimension. We then form a joint token sequence $X = [z_0; z_1] \in \mathbb{R}^{(L_0+L_1) \times d}$, and use it as both queries and keys, $Q = K = X$. The attention score matrix is computed as

$$A = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right) \in \mathbb{R}^{(L_0+L_1)\times(L_0+L_1)}. \tag{5}$$

We denote by $\mathcal{I}_0 = \{1, \ldots, L_0\}$ the index set of tokens from view 0 and by $\mathcal{I}_1 = \{L_0 + 1, \ldots, L_0 + L_1\}$ the index set of tokens from view 1. Each entry A_{ij} represents the attention weight assigned by token i to token j. In particular, for $i \in \mathcal{I}_0$ and $j \in \mathcal{I}_1$, A_{ij} encodes the attention from a token in view 0 to a token in view 1, while A_{ji} with $j \in \mathcal{I}_1$ and $i \in \mathcal{I}_0$ encodes the reverse direction. In our setting, we expect tokens that correspond to the same scene location across views to allocate higher attention to each other, so that spatially matching regions are emphasized in the cross-view attention map. If a token $p \in \mathcal{I}_0$ participates in a successful keypoint correspondence, its counterpart in $\mathcal{I}_1$ will exhibit a similar visual feature. As a result, both the attention that p allocates to tokens

$q \in \mathcal{I}1$ and the attention it receives from them tend to be higher. We therefore summarize its cross-view interaction by symmetrizing and averaging these two directions as

$$a_p^{(0)} = \frac{1}{2}\big(\frac{1}{L_1}\sum_{q\in\mathcal{I}_1} A_{pq} + \frac{1}{L_1}\sum_{q\in\mathcal{I}_1} A_{qp}\big), \quad q \in \mathcal{I}_1. \tag{6}$$

Since the keypoint-based correspondence map $\tilde{M}_0$ assigns a high value to the location of token p, we use it as supervision in our loss, encouraging $a_p^{(0)}$ to take a similarly high value:

$$\mathcal{L}_{\text{attn0}} = \frac{1}{N}\left(\left\|\tilde{A}_0 - \tilde{M}_0\right\|_2^2\right), \quad (\tilde{A}_0)_p = a_p^{(0)}, \tag{7}$$

where $N = H_s W_s$ denotes the number of spatial locations per view.

Similarly, we apply the same supervision in the reverse direction from q to p, leading to the following attention loss:

$$\mathcal{L}_{\text{attn1}} = \frac{1}{N}\left(\left\|\tilde{A}_1 - \tilde{M}_1\right\|_2^2\right). \tag{8}$$

The correspondence-guided attention loss is defined as

$$\mathcal{L}_{\text{attn}} = \mathcal{L}_{\text{attn0}} + \mathcal{L}_{\text{attn1}}. \tag{9}$$

This loss encourages the joint attention to place high mass on spatially corresponding scene regions in both views, guiding the model to preserve underlying scene structure while editing according to the text and to produce viewpoint-consistent content. In practice, we apply this supervision only to a single mid-level joint-attention block, which empirically balances strong spatial guidance with sufficient flexibility in the remaining layers.

3.4 Training Objective

We train the model using a combination of the flow matching loss and the correspondence-guided attention loss. We construct the intermediate noisy latent of the target image I_{full} at a randomly sampled timestep $t \in [0, 1]$, and train the network to predict the corresponding vector field given the noisy latent, the reference scene I_{scene}, and the text c. Let u_0 denote the clean latent representation of I_{full}, and $u_1 \sim \mathcal{N}(0, I)$ denote the initial Gaussian noise. The intermediate noisy state u_t is constructed via linear interpolation: $u_t = tu_0 + (1-t)u_1$. The model prediction v_θ is trained to match the ground-truth velocity $v_t = u_0 - u_1$. The flow matching loss is formulated as:

$$\mathcal{L}_{\text{flow}} = \mathbb{E}_{t,u_0,u_1}\left\|v_\theta(u_t, t, I_{\text{scene}}, c) - v_t\right\|^2. \tag{10}$$

The total training objective is

$$\mathcal{L} = \mathcal{L}_{\text{flow}} + \lambda\,\mathcal{L}_{\text{attn}}, \tag{11}$$

where λ controls the strength of the correspondence-guided attention supervision. In our experiments, we use a fixed λ (e.g., $\lambda = 3.0$), and apply $\mathcal{L}_{\text{attn}}$ only to selected attention blocks.

4 Experiments

4.1 Implementation Details

The training set is constructed from Places365 [30] and OmniEdit [24] by filtering out low-quality samples using cues such as face, DINOv2 [16] semantics, and contours to reject issues like incomplete entity removal, resulting in 96,040 distinct entity-centric scene images. For each image, we synthesize a short video clip using Wan 2.2 [22], where the motion prompt fed to Wan 2.2 is refined by a vision–language model conditioned on each scene image, yielding diverse yet plausible camera motions. We then filter the frames using MINIMA [19], a RoMa-based [3] dense matcher, by enforcing a default threshold of over 2000 robust inliers to mitigate video hallucination and structural drift. Text descriptions for all samples are annotated using the GLM-4.1V-Thinking vision–language model [7]. Crucially, during data loading, to prevent the model from learning trivial identity mappings, we employ a temporally-weighted sampling strategy. Based on the prior that later frames inherently exhibit larger cumulative camera motions, they are dynamically assigned higher selection probabilities within each video clip.

Training is conducted on 8 NVIDIA H100 GPUs with a per-GPU batch size of 1 and gradient accumulation over 6 steps, resulting in an effective batch size of 48. We adopt multi-resolution training with arbitrary aspect ratios, constraining the longer image side to at most 1024 pixels (maximum resolution 1024×1024). Optimization proceeds for 4,000 steps. At inference time, we use 28 denoising steps.

4.2 Evaluation Setup

Benchmark. The test set comprises 425 images collected from a variety of scene types. For each image, five distinct textual prompts are created using Gemini-2.5-Pro [2], yielding a total of 2,125 scene-prompt pairs. All evaluated models are tested on this identical set of inputs to ensure comparability under consistent conditions.

Evaluation Metrics. We assess the quality of the generated image using Gemini 2.5 Flash [2], focusing on two aspects: (i) Gemini 2.5 Flash scene alignment (G2.5F-SA), which measures the scene consistency between the reference image and the generated frame. It evaluates whether the model preserves the spatial structural layout of the scene while ignoring acceptable modifications. Variations such as object modifications, depth-of-field effects, zoom or rotation, and moderate viewpoint changes are considered acceptable modifications and are excluded from negative judgment, with the score focusing solely on scene background

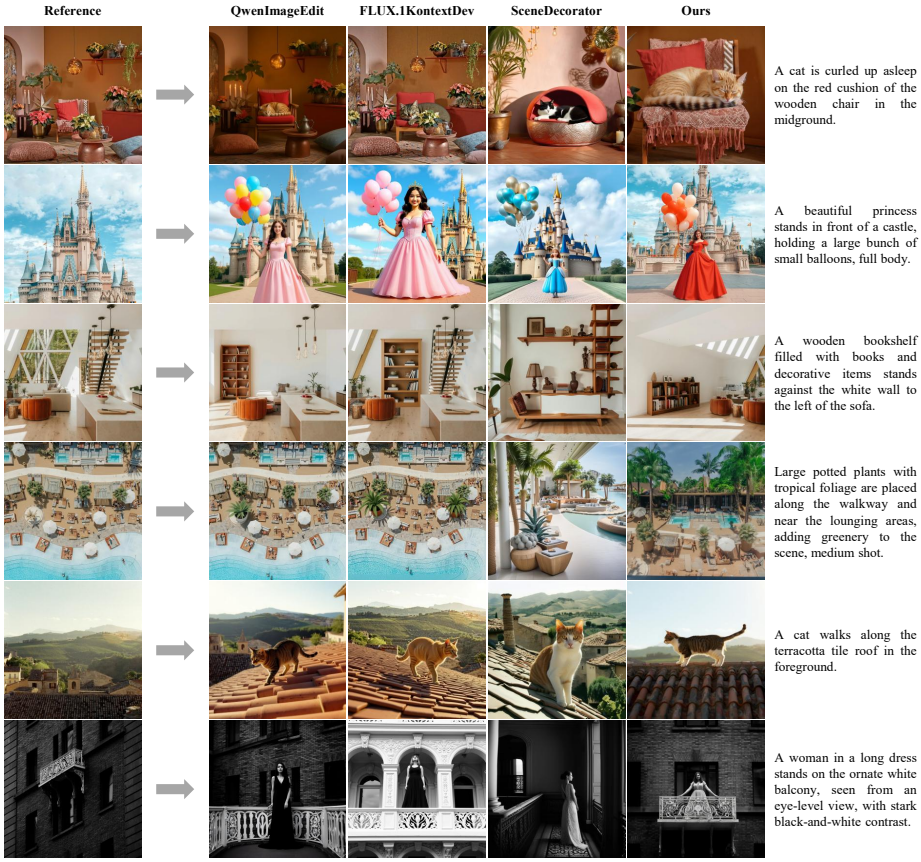

Fig. 3: Qualitative comparison of our method with other baselines. Our method demonstrates superior scene consistency and text-image consistency compared to other baselines.

Table 1: Quantitative comparison of automatic metrics and user study across baselines. The best result is highlighted in **bold**.

Methods	Automatic Metrics		User Study	
	G2.5F-SA↑	G2.5F-TIA↑	Scene Align↑	Text Align↑
Qwen-Image-Edit [25]	9.434	9.306	34.89%	28.43%
FLUX.1 Kontext Dev [9]	8.395	8.841	26.29%	25.46%
SceneDecorator [21]	4.104	8.176	1.69%	17.67%
Ours	**9.811**	**9.639**	**37.13%**	**28.45%**

consistency. (ii) Gemini 2.5 Flash text-image alignment (G2.5F-TIA), evaluates how faithfully the visuals follow the given textual prompt. Both metrics are scored on a scale from 0 (poor) to 10 (excellent). Previous work mainly relies on CLIP-T [5] and DreamSim [4] for evaluation, but we find that these metrics do not accurately capture our task; a detailed analysis and the specific evaluation prompts used for Gemini 2.5 Flash are provided in the supplementary material.

4.3 Comparison with State-of-the-Art Methods

Quantitative Comparisons. We conduct a comparative evaluation of our method with three representative baselines: Qwen-Image-Edit [25], FLUX.1 Kontext Dev [9], and SceneDecorator [21]. The performance is measured using objective automatic metrics, with detailed results summarized in Table 1. As shown in the table, our method achieves the highest performance across all automatic metrics. While general-purpose editing models like Qwen-Image-Edit and FLUX.1 Kontext Dev demonstrate strong foundational capabilities, they yield lower scores in both scene alignment and text-image alignment compared to our method. We attribute these performance gaps primarily to their reliance on generic training objectives and the absence of fine-tuning on datasets explicitly curated for scene consistency. More remarkably, SceneDecorator suffers a degradation in scene alignment, which empirically validates our earlier hypothesis: training-free attention sharing relies heavily on raw visual similarity, causing the underlying structural integrity to collapse when dynamic camera movements alter the visual features. In contrast, by explicitly regularizing spatial alignment through correspondence guidance, our method successfully maintains the background stage while accurately executing the prompt.

Qualitative Comparisons. The visualization results are presented in Figure 3, with additional examples available in the supplementary material. While Qwen-Image-Edit [25] and FLUX.1 Kontext Dev [9] effectively follow textual instructions, they often suffer from "contextual hallucination"—introducing unintended changes to background textures, scene layouts, or other elements during entity insertion—and demonstrate limited spatial alignment capability. SceneDecorator [21], while maintaining a coherent global atmosphere, often exhibits degradation in fine-grained visual details. As observed in Figure 3, specific background elements (e.g., material textures, wall patterns, and intricate objects) tend to lose fidelity or undergo semantic drift compared to the reference scene. This leads to a noticeable divergence from the original scene identity, particularly in complex environments. In contrast, our method seamlessly integrates new entities while preserving the fine-grained textural richness and structural integrity of the reference "stage".

User Study. We carried out a user study on the 2,125 scene–prompt pairs using the corresponding outputs from all compared methods. The evaluation was conducted on an in-house crowdsourcing platform with 50 participants, yielding a total of 6,375 votes, where each scene–prompt pair was evaluated by three distinct annotators. For each scene–prompt pair, annotators were shown the reference scene and all candidate outputs, and were asked to assess each method

Table 2: Ablation study of correspondence-guided attention loss.

Methods	G2.5F-SA↑	G2.5F-TIA↑
w/o CG-Attn Loss	9.532	8.569
w/ CG-Attn Loss	**9.811**	**9.639**

Table 3: Ablation study of data construction strategies.

Data	G2.5F-SA↑	G2.5F-TIA↑
DL3DV-10K [13]	9.048	8.960
ours w/o filter	9.648	9.079
ours	**9.811**	**9.639**

according to two criteria: (i) text alignment and (ii) scene alignment. The scores in Table 1 indicate the percentage of times each method was selected in the user study. The votes were normalized using a temperature parameter adjustment to ensure fair comparability among different methods.

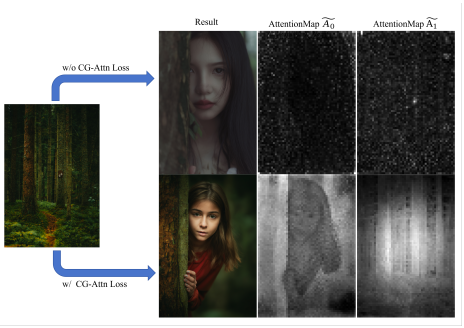

Fig. 4: Generated image and attention map comparison with and without correspondence-guided attention loss. (text prompt:a girl hiding behind a tree, close up)

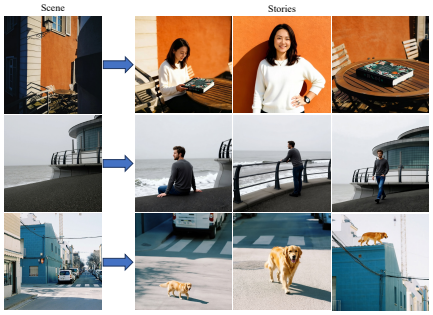

Fig. 5: Coherent visual storyboarding. Multi-frame storylines (right) generated from reference scenes (left). Combining scene staging with identity replacement maintains both scene and actor consistency.

As shown in Table 1, our method achieves the highest preference under both text alignment and scene alignment. Compared to Qwen-Image-Edit [25], our approach is more frequently judged to produce images that both follow the textual instructions and preserve the underlying scene environment. We attribute these gains primarily to the combined effect of our scene-consistent data construction pipeline and correspondence-guided attention loss. The former supplies training pairs with spatially plausible entity–scene relationships and cross-view variations, while the latter explicitly regularizes the spatial layout, thereby validating the effectiveness of our joint design for improving scene consistency.

4.4 Ablation Study

In this section, we conduct ablation studies to evaluate the effectiveness of the two proposed components: correspondence-guided attention loss and scene-consistent data construction pipeline.

Correspondence-Guided Attention Loss. Furthermore, the impact of the proposed correspondence-guided attention loss can be observed in Table 2, which highlights its role in enhancing the coherence of generated scenes. With the attention loss, the model achieves improvements on both scene alignment and text–image alignment metrics, with G2.5F-SA increasing by 0.279 and G2.5F-TIA increasing by 1.07, which demonstrates its ability to maintain stable spatial relationships within the scene. In addition, we provide a visual comparison of generations with and without the proposed correspondence-guided attention loss and analyze the attention maps from the designated mid-level block used to compute the loss, as shown in Figure 4. $\tilde{A}_0$ denotes the aggregated cross-view attention map for the target image (view 0), while $\tilde{A}_1$ denotes the aggregated cross-view attention map for the scene image (view 1); brighter regions indicate higher correlation at the corresponding positions. With the attention loss, the attention maps highlight spatially corresponding regions in both images, indicating that the network has learned to focus on areas crucial for maintaining scene consistency. These observations collectively demonstrate that the attention loss not only improves visual scene consistency but also guides the model to attend to structurally relevant background regions.

Scene-Consistent Data Construction. We compare models trained with our Scene-Consistent Data Construction pipeline (Section 3.2) against a baseline built from real-world multi-view datasets with post-hoc entity insertion. For the baseline, we start from multi-view scene sequences in DL3DV-10K [13] and insert foreground humans or objects into the target views using FLUX.1 Kontext Dev [9], while keeping the network architecture, optimization hyperparameters, and training schedule identical. As reported in Table 3, training on our scene-consistent data yields consistently higher G2.5F-SA and G2.5F-TIA scores than the DL3DV-10K–based baseline, demonstrating the clear advantage of our pipeline: while the DL3DV-10K baseline provides multi-view coverage, the post-hoc insertions lack the organic interaction between entity and scene. The resulting compositing artifacts and instance-scale imbalance likely degrade supervision quality, as further discussed in the supplementary material.

Furthermore, to validate the necessity of our filtering strategy in mitigating structural drift, we ablated training on MINIMA-filtered data against raw, unfiltered data. As shown in Table 3, training on unfiltered data contaminated with hallucinations or layout shifts results in a decrease of G2.5F-SA from 9.811 to 9.648 and a more pronounced drop in G2.5F-TIA from 9.639 to 9.079. These results suggest that our filtering step effectively mitigates the impact of generated artifacts, encouraging the model to learn from more consistent samples and thereby suppressing structural drift.

4.5 Application: Coherent Visual Storyboarding

As discussed in Section 1, continuous visual storytelling necessitates coherence across two primary dimensions: the "stage" and the "actor". While our primary contribution successfully addresses the critical bottleneck of scene consistency, we also explore a practical workaround to achieve comprehensive narrative coherence by incorporating existing identity-preserving techniques.

To demonstrate this, we introduce a cascaded pipeline for comprehensive visual storyboarding. First, we employ our proposed Scene Staging method to synthesize structurally accurate frames, embedding a text-specified entity into the reference scene. Subsequently, we adopt Qwen Image Edit as a post-processing step to replace the generated entity, thereby enforcing strict identity preservation of the "actor".

As illustrated in Figure 5, this cascaded approach demonstrates the feasibility of generating multi-frame storylines. The left column shows the reference scenes, while the right columns present the generated storyboards. The "stage" (e.g., the building facades, the seaside, and the street) is robustly preserved by our method, and the specific identities of the "actors" (e.g., the woman, the book, the man, and the dog) are maintained via the subsequent replacement step.

While this decoupled solution serves as a viable strategy to bridge the gap between scene staging and full visual storytelling, it inherently relies on a multi-step process. This observation highlights the necessity for future research into end-to-end frameworks that jointly optimize for both actor and scene consistency.

5 Conclusion

This paper presents a novel framework explicitly optimized for structural scene preservation, generating images that maintain the underlying spatial layout of a reference scene while synthesizing foreground entities aligned with textual descriptions. Our contributions include: (i) a scene-consistent data construction pipeline that produces high-quality, scene-consistent paired data with precise text–image alignment, enabling effective training with explicit scene consistency constraints; and (ii) a correspondence-guided attention loss that leverages cross-view spatial correspondences to regularize intermediate attention maps, constraining the model to attend to structurally relevant background regions. Extensive experiments across diverse scenarios demonstrate that our method achieves superior performance over existing approaches in both automatic metrics and user studies, with clear improvements in scene alignment and text–image consistency. Qualitative visualizations further confirm that our correspondence-guided design effectively captures fine-grained scene semantics and maintains high scene consistency even under significant viewpoint changes.

Despite these advances, a current limitation of our approach is that it does not explicitly integrate foreground identity preservation into the core model architecture. Achieving comprehensive visual storyboarding currently relies on a

decoupled, two-stage pipeline to maintain the "actor". Nevertheless, in principle, the proposed formulation could be extended to jointly model scene and character consistency by augmenting the data construction pipeline and training objectives with additional identity-aware constraints, which we leave as an important direction for future work.

References

1. Black Forest Labs: Flux.1 tools: Fill, Depth, Canny. `https://blackforestlabs.ai/flux-1-tools/` (2024), accessed: 2024-12-10
2. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
3. Edstedt, J., Sun, Q., Bökman, G., Wadenbäck, M., Felsberg, M.: Roma: Robust dense feature matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19790–19800 (2024)
4. Fu, S., Tamir, N., Sundaram, S., Chai, L., Zhang, R., Dekel, T., Isola, P.: Dreamsim: Learning new dimensions of human visual similarity using synthetic data. arXiv preprint arXiv:2306.09344 (2023)
5. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)
6. Guo, J., Deng, J., Lattas, A., Zafeiriou, S.: Sample and computation redistribution for efficient face detection. arXiv preprint arXiv:2105.04714 (2021)
7. Hong, W., Yu, W., Gu, X., Wang, G., Gan, G., Tang, H., Cheng, J., Qi, J., Ji, J., Pan, L., et al.: Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. arXiv preprint arXiv:2507.01006 (2025)
8. Ju, X., Liu, X., Wang, X., Bian, Y., Shan, Y., Xu, Q.: Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. In: European Conference on Computer Vision. pp. 150–168. Springer (2024)
9. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., et al.: Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. arXiv preprint arXiv:2506.15742 (2025)
10. Li, W., Lin, Z., Zhou, K., Qi, L., Wang, Y., Jia, J.: Mat: Mask-aware transformer for large hole image inpainting. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10758–10768 (2022)
11. Li, Z., Cao, M., Wang, X., Qi, Z., Cheng, M.M., Shan, Y.: Photomaker: Customizing realistic human photos via stacked id embedding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8640–8650 (2024)
12. Li, Z., Liu, Z., Zhang, Q., Lin, B., Yuan, S., Yan, Z., Ye, Y., Yu, W., Niu, Y., Yuan, L.: Uniworld-v2: Reinforce image editing with diffusion negative-aware finetuning and mllm implicit feedback. arXiv preprint arXiv:2510.16888 (2025)
13. Ling, L., Sheng, Y., Tu, Z., Zhao, W., Xin, C., Wan, K., Yu, L., Guo, Q., Yu, Z., Lu, Y., et al.: Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22160–22169 (2024)

14. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: European conference on computer vision. pp. 38–55. Springer (2024)
15. Mao, C., Zhang, J., Pan, Y., Jiang, Z., Han, Z., Liu, Y., Zhou, J.: Ace++: Instruction-based image creation and editing via context-aware content filling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1958–1966 (2025)
16. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
17. Park, J., Lee, K., Gim, J., Jo, H., Oh, M., Choi, W., Hwang, K., Kim, J., Choi, M., Im, S.: Infinite-story: A training-free consistent text-to-image generation. arXiv preprint arXiv:2511.13002 (2025)
18. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
19. Ren, J., Jiang, X., Li, Z., Liang, D., Zhou, X., Bai, X.: Minima: Modality invariant image matching. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 23059–23068 (2025)
20. Seedream, T., Chen, Y., Gao, Y., Gong, L., Guo, M., Guo, Q., Guo, Z., Hou, X., Huang, W., Huang, Y., et al.: Seedream 4.0: Toward next-generation multimodal image generation. arXiv preprint arXiv:2509.20427 (2025)
21. Song, Q., Zhou, D., Lin, J., Shen, F., Wang, J., Hu, X., Chen, C., Heng, P.A.: Scenedecorator: Towards scene-oriented story generation with scene planning and scene consistency. arXiv preprint arXiv:2510.22994 (2025)
22. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
23. Wang, Q., Bai, X., Wang, H., Qin, Z., Chen, A., Li, H., Tang, X., Hu, Y.: Instantid: Zero-shot identity-preserving generation in seconds. arXiv preprint arXiv:2401.07519 (2024)
24. Wei, C., Xiong, Z., Ren, W., Du, X., Zhang, G., Chen, W.: Omniedit: Building image editing generalist models through specialist supervision. In: The Thirteenth International Conference on Learning Representations (2024)
25. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
26. Xia, B., Peng, B., Zhang, Y., Huang, J., Liu, J., Li, J., Tan, H., Wu, S., Wang, C., Wang, Y., et al.: Dreamomni2: Multimodal instruction-based editing and generation. arXiv preprint arXiv:2510.06679 (2025)
27. Xie, S., Zhang, Z., Lin, Z., Hinz, T., Zhang, K.: Smartbrush: Text and shape guided object inpainting with diffusion model. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22428–22437 (2023)
28. Yang, B., Gu, S., Zhang, B., Zhang, T., Chen, X., Sun, X., Chen, D., Wen, F.: Paint by example: Exemplar-based image editing with diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18381–18391 (2023)
29. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)

30. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017)
31. Zhou, Y., Zhou, D., Cheng, M.M., Feng, J., Hou, Q.: Storydiffusion: Consistent self-attention for long-range image and video generation. Advances in Neural Information Processing Systems **37**, 110315–110340 (2024)
32. Zhou, Z., Li, J., Li, H., Chen, N., Tang, X.: Storymaker: Towards holistic consistent characters in text-to-image generation. arXiv preprint arXiv:2409.12576 (2024)