InstaPano: Zero-shot Instance Layout Controlled Panorama Generation Via Global Attention Fusion

Zejian Li[1] *, Rui Huang[1], Lefan Hou[2], Pei Chen[2], Heyuan Xu[3], Shengyuan Zhang[2], Kewen Zhu[4], Huanghuang Deng[2], Li Liu[5], and Lingyun Sun[2]

[1] School of Software Technology, Zhejiang University, Ningbo, China
{zejianlee,22551377}@zju.edu.cn
[2] College of Artificial Intelligence, Zhejiang University, Hangzhou, China
[3] College of Software, Northeastern University, Shenyang, China
[4] School of Art and Archaeology, Zhejiang University, Hangzhou, China
[5] School of Big Data & Software Engineering, Chongqing University, Chongqing, China

Abstract. Achieving both global semantic coherence and precise instance level control in wide-aspect-ratio panorama generation is an unresolved challenge. Existing methods that synchronize independent views to generate panoramas often lack semantic coherence and struggle with fine-grained multi-instance placement, resulting in contextual artifacts and fragmented objects. We introduce InstaPano, a training-free framework for zero-shot instance-controlled panorama generation. InstaPano integrates a Global Attention Fusion mechanism into a pre-trained layout-to-image model. Through Sync-Fuse-Dispatch workflow, it periodically aggregates latent features from all local views to construct a unified global context, performs multi-level attention computation over this context to achieve true fusion, and then dispatches the enriched global features back to each view. This enables the coherent rendering of complex panoramas with multiple specified instances. Furthermore, we introduce a conditional positional mask to resolve object repetition artifacts that may arise in large bounding boxes. On a newly constructed yet challenging benchmark for panoramic instance layout control, InstaPano achieves superior performance in both layout fidelity and semantic coherence, faithfully generating complex panoramic scenes.

Keywords: Multi-Instance Generation · Controllable Panorama Generation · Joint Diffusion · Attention Fusion

1 Introduction

In recent years, diffusion models [5,12,21,27,28] have revolutionized image generation, including wide-aspect-ratio text-to-panorama synthesis. Existing methods

* Corresponding author.

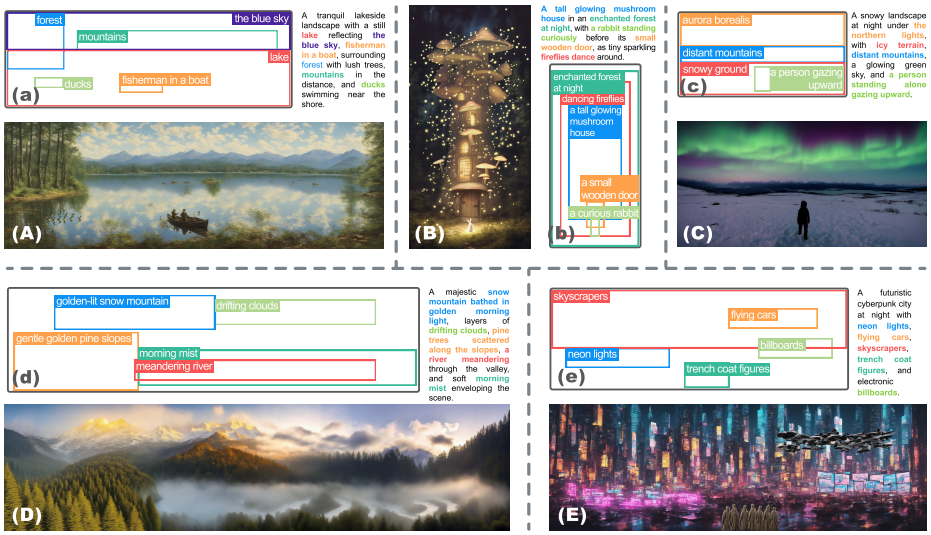

Fig. 1: InstaPano achieves zero-shot, instance-level, fine-grained layout control on panoramic images with different aspect ratios. The model accurately places objects according to the specified bounding boxes (labeled in lower case), even in complex scenes, resulting in generated images (labeled in upper case).

generate panoramic scenes through iterative extension [1, 19, 41] or joint diffusion paradigms, e.g., MultiDiffusion [2]. Meanwhile, specialized architectures directly model 360-degree spherical panoramas globally, such as MVDiffusion [31], PanFusion [38], and DiT360 [8]. For clarity, we use "panorama" to refer to wide-aspect-ratio planar images generated via horizontal extension and view stitching, aligning with our focus on controllable generation over extended canvases.

Beyond pure text conditioning, several works [6, 7, 31, 37, 38, 43] introduce structural cues, such as depth maps or layout distance maps, to guide spatial organization, either through auxiliary conditioning modules, e.g., ControlNet [40], or dedicated architectural designs. However, these methods mainly focus on structural or coarse spatial layouts. Fine-grained, instance-level layout control, where users precisely specify the placement of multiple distinct entities using bounding boxes and text prompts, remains largely unexplored.

Achieving precise multi-instance layout control on an extended canvas poses fundamental challenges. On fixed-size square images, typically with a 1:1 aspect ratio, pre-trained Layout-to-Image (L2I) models [17, 33, 34, 39, 44, 45] have demonstrated strong adherence to bounding box instructions. However, directly retraining such models for panoramic scales is severely limited by training memory costs and the scarcity of panoramic datasets with fine-grained multi-instance layout annotations. Therefore, a training-free, zero-shot method for instance-controlled panorama generation is highly desirable.

To bypass inference-time memory limitations, MultiDiffusion [2] decomposes a panorama into overlapping local views, denoises them independently, and ag-

gregates their predictions after each denoising step. For instance-controllable panorama generation, existing MultiDiffusion-based methods commonly follow a post-hoc local sample fusion strategy, where object- or region-specific conditions are assigned to separate local views and merged only after independent denoising. Although scalable to wide canvases, this strategy limits fine-grained instance-level layout control, as self- and cross-attention remain mostly confined within each local view before aggregation. Consequently, the model lacks a holistic layout-aware representation of all objects and boxes, leading to fragmented, duplicated, or inconsistent entities across view boundaries, especially in complex extended layouts, as shown in Figure 3.

Subsequent methods such as SyncDiffusion [15], GVCFDiffusion [29], and MAD [24] improve cross-view coherence through perceptual loss, guided fusion, or internal attention modifications. However, object regions are still largely generated separately, offering limited global reasoning over all instance boxes and their spatial relationships. To overcome this limitation, we introduce InstaPano, a training-free framework that incorporates a Global Attention Fusion mechanism into a pre-trained L2I model. Instead of simply stitching overlapping views, InstaPano follows a Sync-Fuse-Dispatch (SFD) workflow. It periodically synchronizes latent representations from overlapping local views into a unified global canvas, performs multi-level attention for cross-view information exchange and instance-level layout-guided cross-attention, and then dispatches the enriched global features back to each local view. In this way, InstaPano shifts the generation process from post-hoc sample fusion to holistic layout-aware generation: all local views and instance layouts are jointly represented on a shared global canvas, while the pre-trained L2I model's layout-grounding priors are preserved and extended to panoramic generation. This enables coherent global structure and precise instance placement in zero-shot wide-aspect-ratio generation, as illustrated in Figure 1. Furthermore, we identify an artifact of object duplication within large specified regions and address it with a conditional positional masking strategy that regularizes spatial attention.

To rigorously evaluate this task, we construct a new benchmark for instance layout-controlled panorama generation. Experiments show that InstaPano consistently improves layout fidelity and cross-view coherence over region-based joint diffusion methods. These results demonstrate that holistic layout-aware attention on a shared global canvas is effective for extending pre-trained L2I models to wide-aspect-ratio panoramas while preserving global consistency and accurate instance grounding.

The main contributions of this paper are summarized as follows:

- We formalize the task of instance-level layout-controlled panorama generation and construct a benchmark to evaluate fine-grained spatial grounding on extended canvases.
- We propose InstaPano, a training-free framework that transforms region-based joint diffusion into holistic layout-aware panorama generation through Global Attention Fusion, enabling scalable and globally consistent layout control across panoramic views.

– Compared with region-based generation methods built on MultiDiffusion, InstaPano achieves superior layout fidelity and cross-view coherence.

2 Related Work

Our research lies at the intersection of layout-controlled synthesis and panorama generation. We briefly review representative literature in these domains and highlight the gaps addressed by our work.

Layout-Controlled Image Generation. Layout-controlled image synthesis guides image generation using specified spatial arrangements and attributes. Existing diffusion-based approaches typically fall into two paradigms: *training-free methods* [2,3,16,35] that manipulate attention or energy functions during inference; *Training-based methods*, including GLIGEN [17], InstanceDiffusion [33], MIGC [45], and others [4, 34, 44, 46], introduce learnable modules or specialized architectures (e.g., adapters or separated conditioning branches) to enhance layout adherence. Additionally, *LLM-assisted methods* like LayoutGPT [9] and LMD [18] leverage Large Language Models to deduce intermediate bounding box layouts for compositional visual planning. Despite their effectiveness on standard-resolution images, most of these methods are not inherently designed for wide-aspect-ratio inputs. While MultiDiffusion [2] has demonstrated initial potential for region-based control on extended canvases, achieving fine-grained layout fidelity and global semantic coherence remains an unresolved challenge, underscoring the need for dedicated panoramic layout generation methods.

Panorama Generation. Existing panorama generation methods generally follow two distinct lines: 360° *spherical panoramas* and *wide-aspect-ratio planar panoramas*. Spherical methods, such as PanFusion [38], StitchDiffusion [32], PanoFree [19], PanoDecouple [43] and DiT360 [8], utilize specialized architectures to handle equirectangular distortions and ensure rotational continuity, strictly oriented toward immersive full-view scenes. In contrast, wide-aspect-ratio planar panoramas extend standard diffusion models to long horizontal canvases. Current approaches typically follow either an *iterative extension* paradigm [1, 19, 41] or a *joint diffusion* paradigm that fuses overlapping views during sampling [2]. Within the joint diffusion framework, MultiDiffusion [2] establishes the foundational view-fusion formulation. Subsequent methods have enhanced this pipeline to improve boundary seamlessness: SyncDiffusion [15] and GVCFDiffusion [29] introduce perceptual losses and guided/variance-corrected fusion, while SpotDiffusion [10] employs temporal window-shifting strategies. MAD [24] incorporated attention fusion to promote cross-view information sharing. However, these methods primarily target cross-view coherence. When extended to layout-guided generation, they still inherently process each view as an independent crop and lack a mechanism for *global spatial reasoning*. Existing methods often employ coarse structural controls such as depth maps, edge maps, or sketches (e.g., MultiDiffusion with ControlNet [40] or synchronized diffusion frameworks [13]), while spherical models like PanFusion use distance map cues for global geometry—none provide fine-grained instance-level placement via bounding boxes.

To address this gap, we incorporate a pre-trained Layout-to-Image (L2I) model into the joint diffusion framework. By leveraging global attention fusion, our approach facilitates coherent spatial reasoning and precise layout control for wide planar panoramas.

3 Preliminaries

Latent Diffusion Models (LDMs) [23, 25] (e.g., Stable Diffusion) operate in a latent space obtained via a pretrained VAE $(\mathcal{E}, \mathcal{D})$ [14]. A noise prediction network ϵ_θ is trained to denoise $z_0 = \mathcal{E}(x_0)$, with objective

$$\mathcal{L} = \mathbb{E}_{t, z_0, \epsilon}\Big[\|\epsilon - \epsilon_\theta(z_t, t, c)\|^2\Big] \tag{1}$$

where c is the conditioning input. To incorporate external conditions like text, LDMs employ cross-attention operation. Layout-to-image models like MIGC [45] and GroundingBooth [36] use **masked cross-attention**:

$$\text{Attention}(Q, K, V, M) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}} + M\right)V, \tag{2}$$

where mask M enforces specific regions to attend only to corresponding textual tokens. Our method, InstaPano, extends this mechanism for panoramic generation.

Joint diffusion. MultiDiffusion [2] generates panoramas by applying a pretrained model to overlapping crops and fusing the outputs at each denoising step through this optimization:

$$\Psi(J_t \mid z) = \arg\min_{J \in \mathcal{J}} \sum_i^n \left\| W_i \odot \left[F_i(J) - \Phi(I_i^t \mid y_i)\right]\right\|^2. \tag{3}$$

Here W_i is a blending mask and y_i the text condition for region i.

4 Pre-Experiment

Generating wide-aspect-ratio panoramas faces a fundamental memory limitation. Typical pipelines like MultiDiffusion address this through joint diffusion paths, decomposing panoramas into overlapping square images processed independently, with latent aggregation performed after each denoising step.

We define these independently processed local images as views, with the complete scene formed by spatial integration termed the global canvas. While this view-based approach resolves memory limitations, it introduces a critical challenge: ensuring seamless transitions and semantic coherence across overlapped views.

Prior work [24] has shown that fusing attention layers across diffusion paths improves the semantic consistency of panoramic generation, the underlying mechanism remains unexplained. To address this, we conduct a visual and conceptual

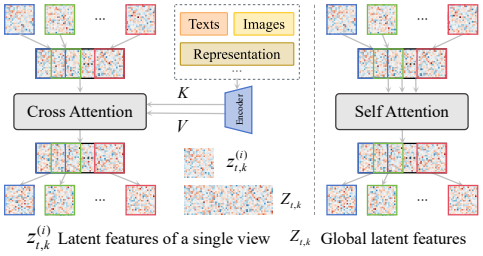

Fig. 2: The Attention Fusion Mechanism: by aggregating latent features from multiple views into a global canvas, the model can perform attention on the entire panoramic canvas.

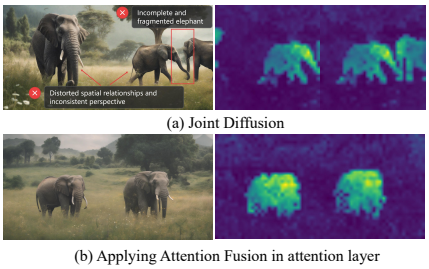

Fig. 3: Attention map for the token "elephant". (a) Without attention fusion: view-specific and fragmented. (b) With attention fusion: coherent global representation.

analysis to uncover how attention fusion enhances consistency, which serves as the empirical motivation for our InstaPano framework.

Specifically, we divide the target panoramic canvas J_T into a set of I overlapping local views $\{v_1, v_2, \ldots, v_I\}$, each of standard resolution $h \times w$, where $\bigcup_{i=1}^{I} v_i = J_T$. At timestep t and layer k, the latent features of each view are denoted as $z_{t,k}^{(i)} \in \mathbb{R}^{C \times h \times w}$. Attention fusion then aggregates $\{z_{t,k}^{(i)}\}_{i=1}^{I}$ into a global latent $Z_{t,k}$ and performs unified attention computation, enabling cross-view semantic alignment and spatial planning at each sampling step (Figure 2).

In standard joint diffusion frameworks (e.g., MultiDiffusion), the self- and cross-attention mechanisms are confined locally within each independent view, preventing cross-view communication. This limitation directly leads to object fragmentation and inconsistent spatial cues. For instance, when generating "a photo of a meadow with an elephant", the attention map for the token "elephant" remains view-specific and fragmented, as visualized in Figure 3(a). This failure to form a holistic plan results in an incoherent final panorama.

In contrast, attention fusion aggregates features from all views into a unified global context, enabling the model to reason over the full canvas. In Figure 3(b), the attention map now accurately localizes the object in the entire panorama.

We refer to this ability as the Global Semantic Modeling Capability, which explains the enhanced consistency and enables us to extend the layout control of pre-trained models to panoramic generation in a scalable and zero-shot manner.

5 Method

5.1 Problem Definition

We formalize Instance Layout Controlled Panorama Generation (ILCPG) task as generating an image from a tuple $\mathcal{T}_{ILCPG} = (P, \{B, D\})$. This task requires placing objects with both spatial, semantic precision and global coherence across a wide-aspect-ratio canvas. The task components are:

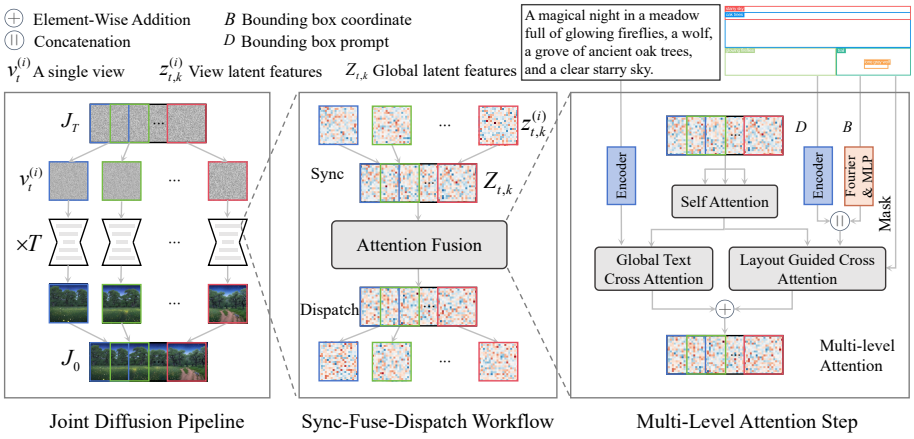

Fig. 4: Overview of InstaPano. The framework alternates Global Context Synchronization, Multi-Level Attention Fusion, and Context Dispatch to enable globally consistent layout-controlled generation.

- P: A prompt describing the overall scene or background.
- $B = \{b_1, \ldots, b_N\}$: A set of N bounding boxes specifying the spatial regions of all objects.
- $D = \{d_1, \ldots, d_N\}$: A set of local descriptions, where d_i is the prompt for the corresponding b_i.

5.2 InstaPano Framework Overview

To address the highly challenging task of instance layout controlled panorama generation, we have designed the InstaPano framework. Instead of training from scratch, InstaPano expands the precise spatial control of a pre-trained layout-to-image (L2I) model to a full panoramic scale via a Global Attention Fusion mechanism. This is achieved through a Sync-Fuse-Dispatch Workflow that specifically operates within the attention layers of the U-Net during the inference of each sampling step in the whole denoising process. As illustrated in Figure 4, this SFD workflow enables effective global attention fusion across the panoramic canvas by replacing standard independent attention computations.

The SFD workflow operates as follows:

1. **Synchronization (Sync)** is to aggregate the latent features of all views into a unified global latent context.
2. **Fusion (Fuse)** is to apply the designed multi-level attention operations on the global context to achieve information fusion across views. Crucially, beyond standard attention fusion, it leverages the pre-trained L2I model's layout-guided masked cross-attention to enforce precise spatial constraints.
3. **Dispatch (Dispatch)** is to split the fused global latent back into local views, injecting global context to guide their subsequent generation.

At each diffusion step, we periodically perform three key operations during attention computation as follows. After each denoising step, we apply MultiDiffusion's synchronization step to aggregate the updated latent features, ensuring consistent and coherent panoramic generation.

5.3 Sync-Fuse-Dispatch Workflow

This process forms the technical core of our method, designed to replace the conventional workflow where U-Net Attention layers process information independently.

Cross-View Latent Synchronization (Sync) The objective of this stage is to aggregate the current state information from all independent views into a single, continuous global context. At a given denoising timestep t and U-Net layer k, we first obtain the set of latent features from all I views, $\{z_{t,k}^{(i)}\}_{i=1}^{I}$. We define a Sync operation that maps this set to a global latent $Z_{t,k}$:

$$Z_{t,k} = \text{Sync}(\{z_{t,k}^{(i)}\}_{i=1}^{I}), \quad \forall i \in [1, I] \tag{4}$$

For the latent feature of the overlapping views, we adopt an averaging fusion strategy:

$$Z_{t,k}(p) = \frac{1}{|\mathcal{I}_p|} \sum_{i \in \mathcal{I}_p} z_{t,k}^{(i)}(p) \tag{5}$$

Here $\mathcal{I}_p$ denotes the set of indices of views that contain spatial position p.

Multi-Level Attention in Global Context (Fuse) The fusion step processes the global latent $Z_{t,k}$ via a multi-level attention block Φ_{attn}, which inherits its architecture from a pretrained layout-to-image model. It sequentially applies Global Self-Attention (SA) and a Combined Cross-Attention (CA). The update of the global latent representation can be formulated as:

$$Z_{t,k+1} = \Phi_{\text{attn}}(Z_{t,k}, P, (B, D)) \tag{6}$$

The internal attention operations follow the standard formulation defined in Eq. (2). The complete computation consists of two stages.

Stage 1: Global Self-Attention (SA). To capture long-range spatial dependencies and promote structural coherence, SA is applied where the query, key, and value matrices (Q, K, V) are all derived from the global latent $Z_{t,k}$, and no mask is used.

$$Q = ZW_Q, \quad K = ZW_K, \quad V = ZW_V, \quad M = \text{None}. \tag{7}$$

Stage 2: Combined Cross-Attention (CA). Following self-attention, a combination of global-text cross-attention and layout-guided cross-attention is applied to the features.

Global Text Cross Attention (GCA) provides global semantic and stylistic guidance by computing attention between the latent features and the global prompt embedding $E(P)$, and then projected to compute K and V. Similarly, no mask is applied:

$$Q = ZW_Q, \quad K = E(P)W_K, V = E(P)W_V, \quad M = \text{None}. \tag{8}$$

Layout Guided Cross Attention (LCA) is aimed for precise layout control. A guidance embedding G_i for each layout region by concatenating the text embedding $E(d_i)$ with a position embedding $E_{\text{pos}}(b_i)$ derived from the bounding box b_i is constructed:

$$G_i = [E(d_i), \text{MLP}(\text{Fourier}(b_i))] \tag{9}$$

The position embedding $E_{\text{pos}}(b_i)$ is computed by encoding the Fourier features of b_i transformed by a multilayer perceptron (MLP). The MLP is already part of the pretrained layout-to-image model [30]. This embedding G_i is used to compute K and V, while a hard attention mask M derived from b_i enforces spatial constraints.

$$\begin{aligned} &Q = ZW_Q, \quad K = G_iW_K, \quad V = G_iW_V, \\ &M = \text{Mask}(b_i), \quad \text{Mask}(b_i) = \begin{cases} 0 & \text{if } p \in b_i, \\ -\infty & \text{otherwise} \end{cases} \end{aligned} \tag{10}$$

We observe a fundamental dilemma when dealing with large bounding boxes (Figure 5). For boxes designed to depict a single object, uniformly distributed attention often leads to object duplication within the region. In contrast, boxes meant for multiple objects or scenes (such as trees) require uniform attention to ensure coverage and completeness.

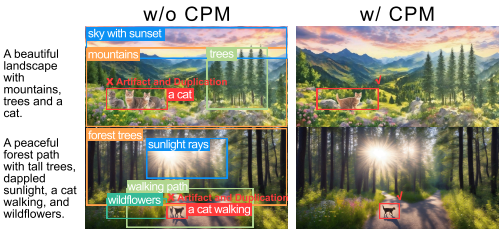

Fig. 5: Resolving object duplication with Conditional Position Mask (CPM).

To address this, we propose a Conditional Position Mask (CPM) strategy that modulates the contribution of layout-guided cross-attention based on the semantic content of each bounding box b_i. Rather than applying the mask within the attention computation, CPM acts as a spatial weighting factor when combining the outputs of global-text cross-attention and layout-guided cross-attention.

Formally, to compute CPM_i the mask of the i^{th} bounding box, the value $\text{CPM}_i(p)$ at pixel location p is defined as:

$$\text{CPM}_i(p) = \begin{cases} \exp\left(-\frac{u_p^2+v_p^2}{2\sigma^2}\right) & \text{if } p \in b_i \text{ and } i \in S_{\text{single}}, \\ 1 & \text{if } p \in b_i \text{ and } i \in S_{\text{multi}}, \\ 0 & \text{if } p \notin b_i. \end{cases} \tag{11}$$

Here (u_p, v_p) are normalized coordinates within the bounding box b_i. S_{single} and S_{multi} are sets of indices for single-object and multi-object prompts, respectively. This center-focused Gaussian decay is designed to counteract the model's tendency to tile features when applying uniform attention over large regions. By anchoring the core object semantics at the center and smoothly attenuating towards the edges, it effectively suppresses peripheral duplication. We develop an LLM-based agent to automatically distinguish multi-object prompts from single-object ones, assigning the former to S_{multi} and the latter to S_{single}.

The final composite cross-attention computation is then expressed as:

$$\text{Attn}_{\text{cross}} = \text{GCA}_P(Z) + \sum_{i=1}^{N} \text{CPM}_i \odot \text{LCA}_{d_i}(Z) \tag{12}$$

Global Context Dispatch (Dispatch) After the global attention fusion, the updated global latent $Z_{t,k}$ contains a unified plan for the entire scene. This stage is responsible for dispatching that global plan back to each independent local view path. We define a splitting operation, denoted as Split, which acts as the inverse of the Sync operation:

$$z_{t,k}^{(i)} = \text{Split}(Z_{t,k}, v_i), \quad \forall i \in [1, I] \tag{13}$$

Each updated local latent $z_{t,k}^{(i)}$ then proceeds through the standard operations of the U-Net's subsequent layers (e.g., convolution, normalization):

$$z_{t,k+1}^{(i)} = \text{UBlock}(z_{t,k}^{(i)}) \tag{14}$$

In summary, our proposed SFD workflow is integrated into MultiDiffusion, where both Global Self-Attention (SA) and Cross-Attention (CA) components are configured to operate during the sampling process.

6 Experiment

6.1 Experimental Setting

Benchmark. To evaluate instance layout control in panoramic settings, we introduce *Pano-Layout-Bench*, comprising 1,341 unique bounding-box layout-prompt pairs. Inspired by the proven spatial reasoning capabilities of LLMs in visual planning [9,18], the dataset was constructed via a semi-automated pipeline: initial scene descriptions and bounding boxes were generated by a multimodal LLM (GPT-4o [22]) using designed templates, followed by manual refinement to ensure logical coherence and spatial realism. The benchmark covers three aspect ratios: 1:2 (412 samples), 1:3 (456 samples), and 1:4 (473 samples). The dataset features an average of 4.86 bounding boxes per image, with box counts ranging from 2 to 8 to cover diverse scenarios. The construction and more statistics are provided in the Supplementary Material.

Baselines. We compare InstaPano with several layout-controlled methods, including MultiDiffusion [2], SyncDiffusion [15], and MAD [24]. Furthermore, we evaluate against direct generation approaches, namely InstanceDiffusion [33], IFAdapter [34], and CreatiLayout [39] to show the efficacy in ILCPG task.

Evaluation Metrics. We evaluate all methods across four critical dimensions:

- **Layout Fidelity:** We compute mIoU, AP, AP50, and AR by comparing generated objects against specified bounding boxes using GroundingDINO [20].
- **Text-Image Consistency:** We use CLIP Score [11] for global text-image alignment and Local CLIP Score for region-specific consistency with local descriptions.
- **Stylistic Coherence:** We employ Intra-LPIPS [42] to measure perceptual similarity across overlapping regions, with lower scores indicating smoother visual transitions.
- **Visual Quality:** We adopt an aesthetic score predictor [26] to evaluate the overall visual appeal and perceptual quality of the generated panoramas.

Implementation Details. Joint diffusion based methods are built on the Stable Diffusion XL [23] backbone. Our InstaPano integrates IFAdapter [34], a layout-to-image model that employs layout-guided masked cross-attention for spatial control within the UNet framework. The σ in CPM is set to be 0.15. For the self attention fusion duration, it operates for the first 10 steps, and cross attention operates throughout the entire process. Self-attention fusion and text cross attention fusion are applied at every layer of the UNet, while layout-guided cross attention follows the design of the underlying layout-to-image model and is applied only at the layers specified therein. For all experiments, the denoising process consists of 30 sampling steps. For the baseline methods in line with MultiDiffusion, the bootstrapping stage is 10, a parameter intended to allow the generated content to more closely fit the exact bounding box.

6.2 Main Results

Quantitative Results. Table 1 shows that InstaPano achieves the best overall performance under the background-only prompt setting. It improves layout fidelity over the strongest baselines by +0.06 mIoU and +0.09 AP50, while also obtaining higher text-image consistency, with +0.52 CLIP and +0.78 Local CLIP gains. In addition, InstaPano reduces Intra-LPIPS compared with MAD by -0.0342, indicating better stylistic coherence. Although the aesthetic score slightly decreases, our method maintains competitive visual quality while significantly improving layout precision and content consistency.

We further evaluate InstaPano with holistic prompts, denoted as InstaPano*. Unlike region-based methods that use a background-only global prompt and specify objects only through box-level prompts, the holistic prompt summarizes both the background and all object descriptions. This setting further improves layout accuracy, text-image alignment, and visual quality, with gains of +0.08 mIoU, +1.78 CLIP, and +0.31 Aesthetic Score over the background-only setting.

Table 1: Quantitative comparison of different methods on Pano-Layout-Bench. ↑ indicates higher is better, and ↓ indicates lower is better. The best results are highlighted in **bold**. For fairness, all methods including ours are evaluated with background-only prompts. Holistic prompt settings indicated by gray * are shown for reference.

Method	Layout Fidelity				Text-Image Consistency		Stylistic Coherence	Visual Quality
	mIoU ↑	AP ↑	AP50 ↑	AR ↑	CLIP ↑	Local CLIP ↑	Intra-LPIPS ↓	Aesthetic Score ↑
MultiDiffusion	0.57	0.17	0.29	0.37	30.07	25.91	0.6865	5.89
SyncDiffusion	0.52	0.13	0.25	0.31	29.10	25.06	0.6625	**6.07**
MAD	0.57	0.21	0.35	0.38	29.44	25.96	0.6007	5.79
InstaPano	**0.63**	**0.25**	**0.44**	**0.44**	**30.59**	**26.74**	**0.5665**	5.81
InstaPano*	0.71	0.25	0.45	0.49	32.37	27.45	0.6038	6.12

Fig. 6: Qualitative comparison under background-only prompts. InstaPano better aligns generated objects with specified boxes and preserves global consistency, while baselines often suffer from misalignment and fragmented contexts. Best viewed magnified on screen.

Qualitative Results. Figure 6 shows that InstaPano produces more accurate object placement and more coherent panoramic structures. In contrast, baseline methods frequently generate misplaced objects, inconsistent details, or broken cross-view semantics.

6.3 Efficacy of Joint Diffusion

Directly applying L2I models to wide-aspect-ratio panoramas introduces a out-of-distribution challenge, often causing *layout collapse* over the extended canvas. As shown in Table 2, CreatiLayout and InstanceDiffusion suffer severe drops in layout fidelity under direct panoramic inference. Although IFAdapter is relatively robust, InstaPano further improves semantic consistency (+0.84 CLIP), spatial accuracy (+0.02 mIoU), and perceptual visual quality (+0.15 Aesthetic Score).

InstaPano shows slightly lower AP/AP50 than IFAdapter, mainly due to a precision–recall trade-off. While our method better covers target regions, as re-

Table 2: Quantitative comparison between direct generation and InstaPano. InstaPano demonstrates superior performance in layout fidelity and visual quality.

Method	Layout Fidelity				Text-Image Consistency		Stylistic Coherence	Visual Quality
	mIoU ↑	AP ↑	AP50 ↑	AR ↑	CLIP ↑	Local CLIP ↑	Intra-LPIPS ↓	Aesthetic Score ↑
InstanceDiffusion	0.46	0.11	0.16	0.25	30.56	24.49	0.6031	5.46
IFAdapter	0.69	**0.26**	**0.47**	0.48	31.53	27.20	**0.5749**	5.97
CreatiLayout	0.34	0.03	0.06	0.12	32.32	23.67	0.5951	5.73
InstaPano*	**0.71**	0.25	0.45	**0.49**	**32.37**	**27.45**	0.6038	**6.12**

Table 3: Post-hoc sample fusion vs. holistic layout-aware attention fusion. Under the same IFAdapter backbone and holistic prompt setting, InstaPano consistently outperforms the MAD post-hoc sample fusion variant, validating holistic layout-aware panorama generation.

Method	mIoU ↑	AP50 ↑	AR ↑	CLIP ↑	Local CLIP ↑	Aesthetic Score ↑
MAD+IFAdapter*	0.56	0.28	0.36	30.02	26.96	5.88
InstaPano*	**0.71**	**0.45**	**0.49**	**32.37**	**27.45**	**6.12**

flected by higher mIoU and AR, the holistic generation process may introduce additional semantically consistent instances outside annotated boxes, which are counted as false positives by AP-based metrics. As further discussed in the Supplementary Material, InstaPano's advantage becomes clearer when foreground repetition is reduced by background prompts.

Fig. 7 further shows that direct panorama generation often leads to missing objects and visible artifacts, whereas InstaPano preserves more complete object structures and more stable spatial reasoning across the panoramic canvas. These results demonstrate that joint diffusion framework mitigates the OOD issue of direct panoramic inference while improving global coherence and visual quality.

6.4 Ablation Study

Post-hoc sample fusion vs. holistic layout-aware attention fusion. To verify that the improvement does not simply stem from adopting a stronger L2I backbone, we construct a controlled variant, MAD+IFAdapter, which uses the same IFAdapter backbone and holistic prompt setting as InstaPano*. As shown in Table 3, MAD+IFAdapter still underperforms InstaPano in layout fidelity, text-image consistency, and visual quality. This suggests that post-hoc sample fusion is less suitable for L2I-based panoramic generation. In contrast, InstaPano enables all boxes and object prompts to interact on a shared panoramic canvas through holistic layout-aware attention fusion.

Self-attention fusion duration. We further vary the duration for which Global Self-Attention (SA) fusion is enabled during denoising, while keeping cross-attention fusion active throughout. Table 4 shows that disabling SA weakens stylistic coherence, whereas longer SA fusion improves AP/AP50 and Intra-LPIPS but slightly reduces CLIP. We adopt $t = 10$ as the default setting, as it

Fig. 7: Comparison with direct generation methods. Direct inference often produces missing objects and artifacts, while InstaPano maintains more complete object structures and better visual quality.

establishes global structure early while maintaining a balanced trade-off across metrics. More ablations are reported in the Supplementary Material.

6.5 Computational Efficiency and Memory Overhead

We profile the layer-wise *incremental* memory overhead of the Sync-Fuse-Dispatch (SFD) workflow during inference on an NVIDIA vGPU (48GB). As shown in Figure 8, the Fusion stage is lightweight, adding only 40–80 MB, while the main overhead comes from Synchronization. At high-resolution layers, Sync introduces about 660 MB due to retaining multi-view features and allocating the aggregated global canvas. Overall, InstaPano achieves precise layout control with manageable inference-time memory overhead.

7 Conclusion

In this paper, we propose InstaPano, a training-free, zero-shot framework for precise instance layout control and global semantic coherence in wide-aspect-ratio panorama generation. Instead of following the conventional joint diffusion paradigm that processes layouts independently and fuses them post hoc, InstaPano shifts toward holistic layout-aware panorama generation. Through the

Table 4: Ablation results of applying Global Self-Attention (SA) fusion for different durations t during denoising.

t (steps)	Layout Fidelity				Text-Image Consistency		Stylistic Coherence	Visual Quality
	mIoU ↑	AP ↑	AP50 ↑	AR ↑	CLIP ↑	Local CLIP ↑	Intra-LPIPS ↓	Aesthetic Score ↑
$t = 0$	**0.68**	0.21	0.37	0.45	**32.34**	**27.71**	0.6140	**6.18**
$t = 10$	**0.68**	0.24	0.42	0.45	32.19	27.62	0.5752	6.12
$t = 20$	**0.68**	0.24	0.43	**0.46**	31.83	27.55	0.5613	6.07
$t = 30$	0.67	**0.26**	**0.46**	0.43	31.53	27.36	**0.5478**	6.15

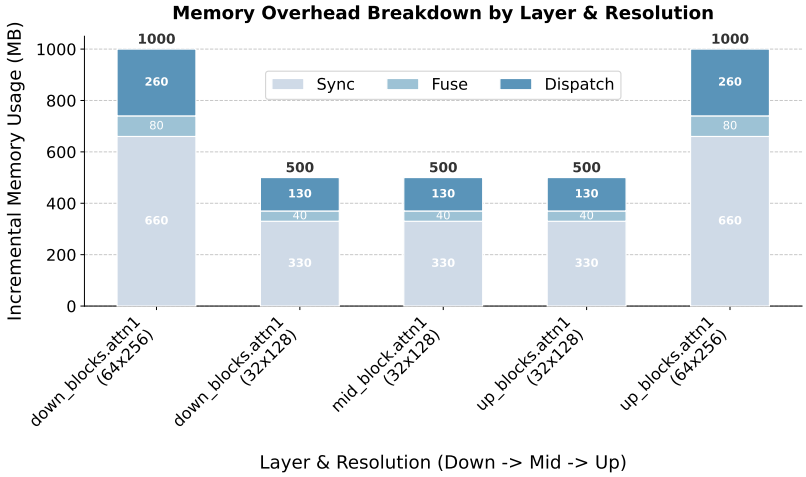

Fig. 8: Incremental memory overhead of the SFD workflow. Sync is the main bottleneck due to tensor aggregation, while Fuse remains lightweight.

Sync-Fuse-Dispatch workflow, it integrates Global Attention Fusion into a pre-trained layout-to-image model, allowing all instance boxes and prompts to interact within a shared global context for coherent planning and fine-grained control. We further construct Pano-Layout-Bench for systematic evaluation. Extensive experiments show that InstaPano achieves superior layout fidelity, text-image consistency, and visual coherence over existing methods. Overall, our framework provides a scalable way to extend pre-trained layout-to-image models to instance-controlled panoramic generation without additional training.

Acknowledgements

This work was supported by National Natural Science Foundation of China (Grant No. 62576306) and Ministry of Education Planning Fund for Humanities and Social Sciences (Grant No. 25YJC760106). The authors would like to thank the anonymous reviewers for their constructive comments.

References

1. Avrahami, O., Fried, O., Lischinski, D.: Blended latent diffusion. ACM transactions on graphics (TOG) **42**(4), 1–11 (2023)
2. Bar-Tal, O., Yariv, L., Lipman, Y., Dekel, T.: Multidiffusion: fusing diffusion paths for controlled image generation. In: Proceedings of the 40th International Conference on Machine Learning. ICML'23, JMLR.org (2023)
3. Chen, M., Laina, I., Vedaldi, A.: Training-free layout control with cross-attention guidance. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 5343–5353 (2024)
4. Cheng, B., Ma, Y., Wu, L., Liu, S., Ma, A., Wu, X., Leng, D., Yin, Y.: Hico: Hierarchical controllable diffusion model for layout-to-image generation. Advances in neural information processing systems **37**, 128886–128910 (2024)
5. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. NIPS '21, Curran Associates Inc., Red Hook, NY, USA (2021)
6. Fang, C., Dong, Y., Luo, K., Hu, X., Shrestha, R., Tan, P.: Ctrl-room: Controllable text-to-3d room meshes generation with layout constraints. In: 2025 International Conference on 3D Vision (3DV). pp. 692–701. IEEE (2025)
7. Fang, C., Li, H., Liang, Y., Zheng, J., Mao, Y., Liu, Y., Tang, R., Zhou, Z., Tan, P.: Spatialgen: Layout-guided 3d indoor scene generation. In: 2026 International Conference on 3D Vision (3DV). pp. 1392–1402. IEEE (2026)
8. Feng, H., Zhang, D., Li, X., Du, B., Qi, L.: Dit360: High-fidelity panoramic image generation via hybrid training. arXiv preprint arXiv:2510.11712 (2025)
9. Feng, W., Zhu, W., Fu, T.J., Jampani, V., Akula, A., He, X., Basu, S., Wang, X.E., Wang, W.Y.: LayoutGPT: Compositional visual planning and generation with large language models. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 18225–18250. Curran Associates, Inc. (2023)
10. Frolov, S., Moser, B.B., Dengel, A.: Spotdiffusion: A fast approach for seamless panorama generation over time. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 2073–2081. IEEE (2025)
11. Hessel, J., Holtzman, A., Forbes, M., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 7514–7528 (2021)
12. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS '20, Curran Associates Inc., Red Hook, NY, USA (2020)
13. Kim, J., Koo, J., Yeo, K., Sung, M.: Synctweedies: A general generative framework based on synchronized diffusions. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024)
14. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
15. Lee, Y., Kim, K., Kim, H., Sung, M.: Syncdiffusion: Coherent montage via synchronized joint diffusions. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 50648–50660. Curran Associates, Inc. (2023)
16. Li, B., Hu, Y., Liu, S., Wang, X.: Control and realism: Best of both worlds in layout-to-image without training. In: Forty-second International Conference on Machine Learning (2025)

17. Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: Gligen: Open-set grounded text-to-image generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22511–22521 (2023)
18. Lian, L., Li, B., Yala, A., Darrell, T.: LLM-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. Transactions on Machine Learning Research (2024), featured Certification
19. Liu, A., Li, Z., Chen, Z., Li, N., Xu, Y., Plummer, B.A.: Panofree: Tuning-free holistic multi-view image generation with cross-view self-guidance. In: Computer Vision – ECCV 2024: 18th European Conference. pp. 146–164. Springer-Verlag, Berlin, Heidelberg (2024). https://doi.org/10.1007/978-3-031-73383-3_9
20. 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)
21. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International conference on machine learning. pp. 8162–8171. PMLR (2021)
22. OpenAI: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
23. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis (2023)
24. Quattrini, F., Pippi, V., Cascianelli, S., Cucchiara, R.: Merging and splitting diffusion paths for semantically coherent panoramas. In: European Conference on Computer Vision. pp. 234–251. Springer (2024)
25. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
26. Schuhmann, C.: Improved Aesthetic Predictor. https://github.com/christophschuhmann/improved-aesthetic-predictor (2022), accessed: 2025-07-30
27. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: Proceedings of the International Conference on Learning Representations (ICLR) (2021)
28. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: International Conference on Learning Representations (2021)
29. Sun, S., Xian, M., Yao, T., Xu, F., Capriotti, L.: Guided and variance-corrected fusion with one-shot style alignment for large-content image generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 7114–7121 (2025)
30. Tancik, M., Srinivasan, P.P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J.T., Ng, R.: Fourier features let networks learn high frequency functions in low dimensional domains. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS '20, Curran Associates Inc., Red Hook, NY, USA (2020)
31. Tang, S., Zhang, F., Chen, J., Wang, P., Furukawa, Y.: MVDiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. In: Thirty-seventh Conference on Neural Information Processing Systems (2023)
32. Wang, H., Xiang, X., Fan, Y., Xue, J.H.: Customizing 360-degree panoramas through text-to-image diffusion models. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 4933–4943 (2024)

33. Wang, X., Darrell, T., Rambhatla, S.S., Girdhar, R., Misra, I.: Instancediffusion: Instance-level control for image generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6232–6242 (2024)
34. Wu, Y., Zhou, X., Ma, B., Su, X., Ma, K., Wang, X.: Ifadapter: Instance feature control for grounded text-to-image generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15949–15959 (2025)
35. Xie, J., Li, Y., Huang, Y., Liu, H., Zhang, W., Zheng, Y., Shou, M.Z.: Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 7452–7461 (2023)
36. Xiong, Z., Xiong, W., Shi, J., Zhang, H., Song, Y., Jacobs, N.: Groundingbooth: Grounding text-to-image customization (2024)
37. Yang, X., Man, Y., Chen, J.K., Wang, Y.X.: Scenecraft: Layout-guided 3d scene generation. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024)
38. Zhang, C., Wu, Q., Cruz Gambardella, C., Huang, X., Phung, D., Ouyang, W., Cai, J.: Taming stable diffusion for text to 360∘ panorama image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
39. Zhang, H., Hong, D., Wang, Y., Shao, J., Wu, X., Wu, Z., Jiang, Y.G.: Creatilayout: Siamese multimodal diffusion transformer for creative layout-to-image generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18487–18497 (2025)
40. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023)
41. Zhang, Q., Song, J., Huang, X., Chen, Y., Liu, M.Y.: Diffcollage: Parallel generation of large content with diffusion models. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10188–10198. IEEE (2023)
42. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
43. Zheng, D., Zhang, C., Wu, X.M., Li, C., Lv, C., Hu, J.F., Zheng, W.S.: Panorama generation from nfov image done right. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21610–21619 (2025)
44. Zheng, G., Zhou, X., Li, X., Qi, Z., Shan, Y., Li, X.: Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22490–22499 (2023)
45. Zhou, D., Li, Y., Ma, F., Zhang, X., Yang, Y.: Migc: Multi-instance generation controller for text-to-image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6818–6828 (2024)
46. Zhou, D., Xie, J., Yang, Z., Yang, Y.: 3dis: Depth-driven decoupled instance synthesis for text-to-image generation. arXiv preprint arXiv:2410.12669 (2024)