COLLAR: Cascaded Object-Level Latent Refinement for High-Fidelity Conditional Generation

Xinlong Zhang, Jia Wei, Xiaoyu Zhang, Teng Zhou, Chengyu Lin, Yongchuan Tang*
College of Computer Science, Zhejiang University, Hangzhou, Zhejiang, China
{xinlzhang, weijia_77, xiaoyzhang, tengzhou, linchengyu, yctang}@zju.edu.cn

Abstract

Achieving high-fidelity object-level control in Diffusion Transformers remains a significant challenge despite the introduction of structural priors like depth and Canny maps. Current object-level conditional generation methods frequently suffer from visual artifacts and struggle to maintain precise control over objects within small localized regions. To address these limitations, we propose Cascaded Object-Level Latent Refinement (COLLAR), a training-free framework that progressively optimizes object-level features via the Field-of-View (FoV) expansion. First, we propose the Cross-Scale Semantic Alignment (CSSA) module to address spatial-semantic gaps by injecting object-level features into extended-FoV branches via attention mechanisms. To further optimize these features, the Cyclic Feature Injection (CFI) module introduces a reciprocal background feedback mechanism. It leverages a frequency-based adaptive strategy to selectively update the global backbone with context-aligned local information. Finally, the extended-FoV branch serves as a hub for integrating context-aligned object-level features into the global generation process. Experiments on COCO-MIG and COCO-POS demonstrate competitive performance across spatial fidelity, semantic alignment, and image quality, with COLLAR achieving the strongest spatial alignment under Canny conditioning and the lowest FID on COCO-MIG under depth conditioning.

1 Introduction

In recent years, text-to-image diffusion models [35, 30, 12] have achieved remarkable success in image generation. Building upon their generative capability, the integration of spatial cues further empowers models to grasp underlying scene structures [46, 27, 48], thereby enabling precise pixel-level control. Such controllability is indispensable for a wide range of practical applications, including professional design[8, 43], advertising[9], and animation[39, 13]. However, as the demand for fine-grained control grows, relying solely on text prompts and a single conditional image often falls short in complex, multi-object scenarios. As illustrated in Figure 1, existing methods still struggle with unavoidable semantic leakage. Specifically, when the global structure contains similar objects, the model fails to accurately render specified attributes (e.g., color) onto the target regions.

To alleviate the above problems, existing studies [51, 14, 45, 5] typically introduce layout information to achieve object-level control in complex scenarios. These methods are mainly divided into two categories: training-free [5, 10, 53] and training-based methods [14, 45, 41]. Training-free methods include pixel-level feature aggregation and attention-based feature modulation. Pixel-level feature aggregation [37, 19] offers strong control over object locations. Due to distribution discrepancies across features of distinct objects, the fusion process is prone to distribution bias, which inevitably introduces artifacts into the final images [5, 10]. Attention-based manipulation [51, 53, 6] enables control by reinforcing semantic activations in targeted regions, but it exhibits instability when handling

*Corresponding author.

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

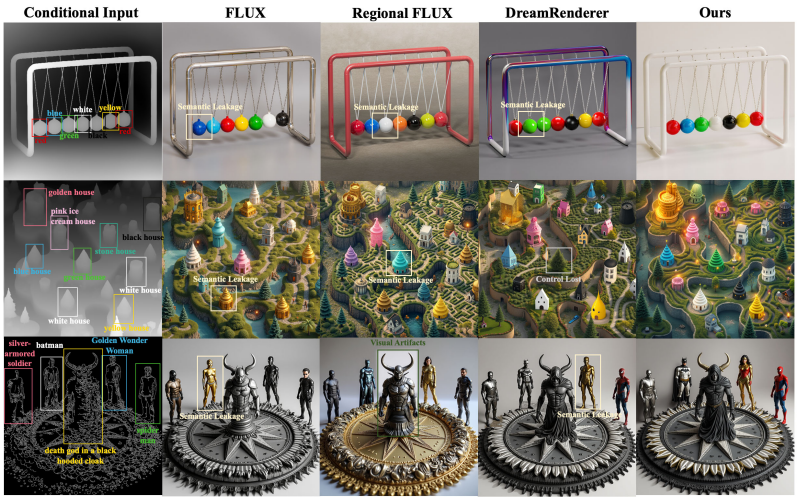


Figure 1: Our proposed framework, COLLAR, is a plug-and-play module specifically designed for object-level conditional generation. It facilitates precise, fine-grained semantic control across diverse spatial modalities. Notably, our approach excels at mitigating attribute leakage and enhancing the generation of small-scale objects.

small-scale objects. Tuning-based methods [52, 11] use bounding-box-guided adapters to control object position and appearance. However, they introduce additional parameters, increasing training complexity and resource overhead. Moreover, such adapters struggle to generalize to other models. The limitations of prior work, as discussed above, motivate the need for further research to address these challenges in object-level conditional generation.

In this paper, we propose a novel framework named COLLAR (**C**ascaded **O**bject-**L**evel **LA**tent **R**efinement). It is a training-free method to achieve high-fidelity object-level conditional image generation. Inspired by [20], COLLAR introduces a paradigm of progressive feature refinement and adaptive fusion. Specifically, it consists of two core modules: (a) Cross-Scale Semantic Alignment (CSSA) module and (b) Cyclic Feature Injection (CFI) module. The CSSA module uses attention-based manipulation to dynamically integrate bounding-box-restricted local features into corresponding extended Field-of-View (FoV) branches. This design strictly preserves local semantic alignment while enhancing the spatial contextual awareness of features. The CFI module establishes a reciprocal background feedback mechanism. It extracts global background features to extended-FoV branches and feeds object-level features into the global branch. Concurrently, we design a frequency-based adaptive injection strategy to selectively inject context-aligned object-level features into the global backbone. Serving as a central hub for feature optimization, extended-FoV branches allow object-level features to smoothly blend into the global generation process while strictly retaining core semantics. We extensively evaluate COLLAR on the COCO-POS and COCO-MIG benchmarks. Extensive quantitative and qualitative experiments demonstrate that COLLAR outperforms existing baselines across multiple metrics. It helps reduce visual artifacts while maintaining competitive image quality and spatial fidelity.

Our contributions are summarized as follows:

- We propose COLLAR, a training-free framework that achieves high-fidelity object-level conditional generation under global structural constraints.
- We introduce the Cyclic Feature Injection technique that facilitates reciprocal feature feedback, enabling local object semantics to seamlessly blend into the global generation process.
- We design an object-level adaptive injection strategy. By dynamically terminating feature injection, it helps reduce structural artifacts and improves inference speed.
- Extensive quantitative and qualitative evaluations demonstrate that our approach outperforms existing methods in object-level conditional generation.

2 Related Work

2.1 Text-to-Image Generation

Text-to-image generation aims to learn a visual-semantic mapping from semantic space to visual representations [33, 34]. Recently, the probabilistic diffusion paradigm [15] has made significant progress in both generation fidelity and text-image alignment [36]. Building on this paradigm, Stable Diffusion [35] demonstrates significant improvements in generation efficiency and fidelity by denoising in a compressed latent space [35, 30]. SDXL [30] also follows this design philosophy. With the advent of the Diffusion Transformer (DiT) architecture [28, 7], models such as SD3 [12] adopt transformer-based backbones with joint attention mechanisms to implicitly model cross-modal semantic correspondences. This design facilitates bidirectional interaction across different modalities, alleviating the information bottleneck inherent [46]. To handle complex text prompts, these models commonly use larger text encoders, such as T5 [32, 1].

2.2 Conditional Text-to-Image Generation

To enhance T2I models' ability for spatial geometric structures, researchers have extended diffusion models to accommodate diverse spatial conditions [46, 27, 47]. Early studies were predominantly built upon the Stable Diffusion framework, with representative works including ControlNet [46] and T2I-Adapter [27]. These methods introduce trainable structural control modules that enable precise injection of geometric priors (e.g., depth maps [22], Canny edges [44], and segmentation maps [21]), while preserving the original generative capabilities of the pre-trained model. Subsequently, Uni-ControlNet [49] and UniControl [31] proposed a novel unified control paradigm by constructing universal conditional encoders capable of hybrid representations for multimodal control signals. To achieve multi-condition controllable generation, Cocktail [17] and Cross-ControlNet [25] have explored novel feature-aggregation strategies. As diffusion model architectures evolve toward DiT, approaches like FLUX-Depth [2] and EasyControl [48] leverage LoRA's [16] parameter efficiency to guide large-scale DiT networks in effectively learning complex spatial-structure priors while maintaining high image fidelity.

2.3 Layout-to-Image Generation

Layout-to-image (L2I) generation aims to achieve object-level compositional image synthesis conditioned on instance information [50]. Existing methods can be broadly categorized into training-based [45, 41, 23, 40, 52, 11] and training-free approaches[5, 10, 6, 37, 19]. Training-based methods typically introduce novel adapters to model layout information for object-level controllable generation, such as GLIGEN [23], InstanceDiffusion [40], and CreatiLayout [45]. Training-free methods often explicitly manipulate cross-attention or joint attention layers to associate object descriptions with local image features, thereby enabling multi-object control. For example, DreamRenderer introduced a Hard Text Attribute Binding mechanism to ensure that the text embedding of each instance correctly aligns with its visual information [51]. Despite their effectiveness in controlling fine-grained attributes such as object position, color, and style, some of these methods remain prone to semantic leakage and edge artifacts when combined with global spatial conditioning.

3 Methods

3.1 Preliminaries

FLUX is a state-of-the-art diffusion-based text-to-image generation model. Its core architecture leverages the T5 text encoder to map complex textual prompts into high-dimensional sequence embeddings. These embeddings are subsequently fused with image features via multimodal joint-attention layers, enabling bidirectional interaction across modalities. FLUX is built upon the multimodal DiT, enabling the semantic alignment of image and text embeddings within a unified embedding space. During the generation process, image features are iteratively denoised via Flow Matching and then decoded into high-fidelity images by a pre-trained autoencoder. Furthermore, FLUX supports conditional image generation tasks—such as FLUX-Depth and FLUX-Canny—by concatenating control images (encoded by the VAE) with the latent features along the channel dimension, thereby achieving precise spatial guidance.

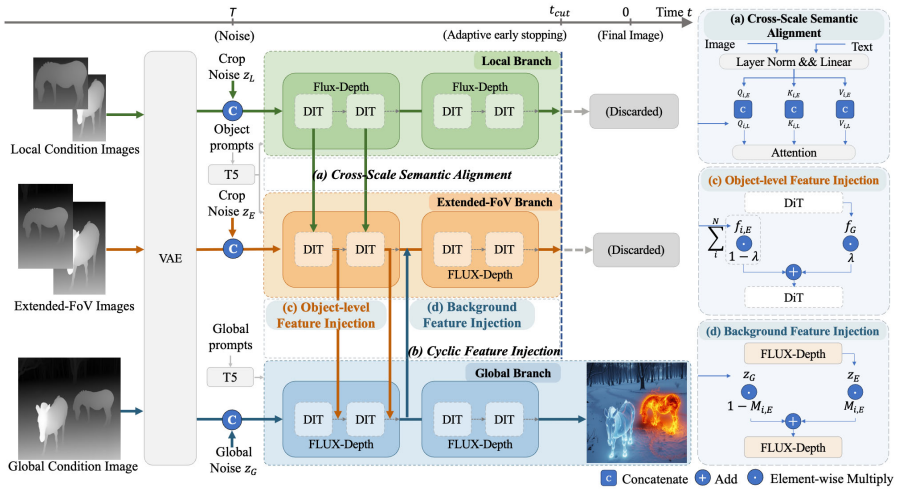


Figure 2: Overview of COLLAR. We freeze the FLUX-Depth and use COLLAR to achieve object-level conditional T2I generation. COLLAR contains two processing modules. By progressively optimizing object-level features, our method adaptively integrates them into the global generation process while helping maintain image quality.

In each MM-DiT layer, image and text features are first normalized via adaptive Layer Normalization (AdaLN), where the normalization parameters are derived from fused timestep and CLIP text embeddings. The normalized image and text features are projected into queries, keys, and values. The corresponding features are then concatenated along the sequence dimension. Before being fed into the self-attention module, the query and key features are augmented with rotary position embeddings (RoPE). This process can be formalized as follows:

$$Q = \text{RoPE}([Q^T, Q^I]), \qquad K = \text{RoPE}([K^T, K^I]), \qquad V = [V^T, V^I], \tag{1}$$

$$\text{Attention}(Q, K, V) = \text{Softmax}(\frac{QK^T}{\sqrt{d}})V. \tag{2}$$

where Q^I, K^I, V^I denote the query, key, and value features for the image modality, and Q^T, K^T, V^T for the text modality. d is the dimension of the key vectors. This mechanism facilitates bidirectional interaction between modalities, enabling better alignment between text and image features.

3.2 Pipeline

As illustrated in Figure 2, we propose a progressive feature optimization and fusion strategy to improve local fidelity across different instances. Our pipeline consists of two core modules: (i) a Cross-Scale Semantic Alignment (CSSA) module, and (ii) a Cyclic Feature Injection (CFI) module. In the CSSA module (Figure 2(a)), we use semantic features from local branches to guide the generation of extended-field-of-view branches, making it take into account both details and context. Specifically, the conditional image crops within each bounding box serve as inputs for local branches to generate semantically aligned details. The extended-FoV branch leverages the fine-grained features from local branches as a semantic reference to dynamically reinforce its own object-level representations during denoising. In the CFI module (Figure 2(b)), we explicitly align local features with the global background, enabling the seamless integration of instances without compromising fidelity. Specifically, we extract background features from the global branch that spatially align with the extended region of the extended-FoV branch. Subsequently, these extracted features are injected into the extended-FoV branch to smooth object semantics. Finally, we design an object-level early-stopping strategy, and the global branch adaptively integrates these object features to synthesize the final high-fidelity image. Serving as a central harmonization hub, the extended branch effectively bridges the spatial scales, and the resulting harmonized object representations are finally fed back into the global backbone.

3.3 Cross-Scale Semantic Alignment

Given the impressive performance of pretrained models, existing object-level generation methods typically extract instance-conditioned features and then fuse them into the global stream using bounding boxes. However, these methods exhibit inherent limitations when global spatial conditions are introduced to guide the generation process. For small-scale objects, the inherently limited resolution of local conditional inputs is further exacerbated by the VAE encoder's downsampling, leading to a severe loss of fine-grained spatial details. As a result, generated instances often lack fine details and fail to achieve precise semantic alignment with their textual descriptions, thereby compromising subsequent multi-instance feature fusion. Expanding the field of view provides more spatial context for object generation. However, background content and nearby objects may weaken alignment with the instance prompt.

Therefore, we propose the CSSA module, which explicitly injects object features from the local branch into the extended-FoV branch to enhance semantics. To achieve this, we employ an attention-based feature fusion mechanism, as illustrated in Figure 2(a). We extract the $Q_{i,L}$, $K_{i,L}$, and $V_{i,L}$ from the local branch corresponding to the i-th instance, and feed them into the DiT layer of the corresponding extended-FoV branch. Within this layer, these features are concatenated with extended-branch features $Q_{i,E}$, $K_{i,E}$, and $V_{i,E}$ along the sequence dimension as follows:

$$\widetilde{Q}_{i,E} = [\text{RoPE}_E(Q_{i,E}), \text{RoPE}_L(Q_{i,L})], \tag{3}$$

$$\widetilde{K}_{i,E} = [\text{RoPE}_E(K_{i,E}), \text{RoPE}_L(K_{i,L})], \tag{4}$$

$$\widetilde{V}_{i,E} = [V_{i,E}, V_{i,L}], \tag{5}$$

$$\text{RoPE}_L = \text{Crop}(\text{RoPE}_E, b_i) \tag{6}$$

RoPE_E and RoPE_L respectively represent rotary position embeddings for the extended-FoV branch and the local branch, and b_i represent the bounding box of the instance i. Then, these concatenated features are fed into Equation (2) to execute feature fusion. It is worth noting that in the DiT layer, text and image features interact densely, and text features also encode visual information. To fully preserve the semantic information from local branches, we feed both their text and image features into the extended-FoV branch. Crucially, we explicitly align the position embeddings by treating the extended-FoV branch as the reference coordinate space. Specifically, position embeddings for features from the local branch are directly assigned the exact spatial subset of indices corresponding to their shared overlapping region. This strict relative alignment empowers the extended branch to precisely localize and attend to the local priors. Furthermore, we also support initializing local branches with distinct noise latents for object i. By aggregating features from multiple parallel local branches and injecting them into the extended-FoV branch, the semantic information is enriched.

3.4 Cyclic Feature Injection

Considering that the extension-FoV branch is still guided by the instance's prompt, forcefully integrating optimized local features (obtained by Section 3.3) directly into global features can easily lead to visual disharmony. Therefore, we propose the CFI module. To clarify this module, we formalize it step by step.

Background Feature Injection. First, we extract the background feature from the global branch that spatially aligns with the extended region of the extended-FoV branch (Figure 2(d)). Let b_i denote the initial bounding box of the object i, and we can obtain the extended-FoV bounding box, $b_{i,E}$, by applying a spatial expansion operation with a scale factor γ:

$$b_{i,E} = \text{Expand}(b_i, \gamma) \tag{7}$$

Then, we crop the corresponding regional features from the global branch z_G:

$$z_E^{global} = \text{Crop}(z_G, b_{i,E}) \tag{8}$$

Finally, we replace the background features in the extended-FoV branch with the cropped global features while retaining the features within the object region:

$$z_E \leftarrow z_E^{global} \odot (1 - M_{i,E}) + z_E \odot M_{i,E} \tag{9}$$

where $M_{i,E}$ represents the object region in the extended-FoV branch.

Object-level Feature Injection. Following refinement in the CSSA module and background-feature injection (Figure 2(d)), object-level features in the extended-FoV branch seamlessly fuse local semantics with spatial context. Consequently, we extract these features and precisely inject them into the corresponding regions of the global branch (Figure 2(c)). Specifically, at the output of each selected DiT layer, the global feature f_G is updated via a weighted summation with the feature set $\{f_{i,E}\}_1^N$ extracted from extended-FoV branches:

$$f_G \leftarrow f_G \odot (1 - M_G) + \left[\lambda f_G + \frac{1-\lambda}{\max\left(\sum_{i=1}^N M_{i,G}, 1\right)} \sum_{i=1}^{N} (f_{i,E} \odot M_{i,G}) \right] \odot M_G \tag{10}$$

where $M_{i,G}$ denotes the globally aligned mask of object i, and M_G is the union of all such masks. The weighted fusion is applied only within the object regions, while features outside these regions are entirely provided by the global branch. Due to the spatial resolution mismatch between the extended-FoV branch and the global branch, we explicitly zero-pad the extracted features $f_{i,E}$ and map their masks to the global feature grid before performing weighted summation.

Object-level Early-Stopping Strategy. The guidance step of object-level feature injection significantly impacts the final image quality. Insufficient steps lead to inadequate semantic activation, whereas excessive injection compromises global coherence. To achieve a trade-off, we propose an object-level early-stopping strategy. At each timestep t, we employ the masks $M_{i,G}$ and $M_{i,E}$ to extract the local latent features from the global and the extended-FoV branch i, respectively. These features are then processed by a high-pass filter $\mathcal{H}(\cdot)$:

$$H_{i,G}^{(t)} = \mathcal{H}(z_G^{(t)} \odot M_{i,G}), \quad H_{i,E}^{(t)} = \mathcal{H}(z_{i,E}^{(t)} \odot M_{i,E}) \tag{11}$$

We compute token-wise cosine similarities between the resulting high-frequency components:

$$s_i^{(t)} = \text{CosSim}(H_{i,G}^{(t)}, H_{i,E}^{(t)}) \tag{12}$$

The global and extended-FoV branches share initial latents in corresponding object regions. Their similarities can therefore be high early in the denoising process. We apply min–max normalization to the token-wise similarities within each object region. Lower normalized values indicate positions with relatively lower similarity between the branches. We average these values to obtain the final stopping score $s_i^{(t)}$. The predefined threshold σ is set to 0.85 by default. If $s_i^{(t)} \geq \sigma$, the feature injection for instance i is terminated at the timestep t in the global branch (Figure 2). The intuition behind this design is that high-pass residuals are sensitive to local edges and contours, as supported by the analyses in Section B.1. As denoising progresses, continued object-level feature injection may disrupt global coherence and introduce visual artifacts. Overall, this strategy significantly reduces inference costs while helping the final images remain natural and high-quality. In addition, we set the maximum guidance step S.

4 Experiments

4.1 Implementation Details

To evaluate the effectiveness of COLLAR, we conduct experiments on FLUX.1-Depth and FLUX.1-Canny [2]. During inference, the FlowMatchEulerDiscreteScheduler is employed with 20 sampling steps for all benchmarks. For structural guidance, the Classifier-Free Guidance scale is set to 10.0 for depth-conditioned generation and 30.0 for Canny-conditioned generation. All experiments are conditioned on boxes, object-level descriptions, text prompts, and structural images.

4.2 BaseLines

We compare COLLAR with state-of-the-art (SOTA) training-free, object-level conditional generation methods, including DreamRenderer [51], 3DIS [53], and Regional-FLUX [5]. Given that COLLAR employs a latent-level patch transplantation strategy, we also include several representative methods, such as NoiseCollage [37] and GrounDit [19]. To ensure a fair comparison, we omit the plug-and-play strategy R&B [42] from GrounDit, as it is used for the cross-attention module and is not directly applicable to FLUX's joint-attention architecture. Furthermore, we also integrate COLLAR into the latest conditional generation models, such as EasyControl [48] and SD3 [12], to demonstrate its compatibility and effectiveness.

4.3 Evaluation Benchmarks

We evaluate the effectiveness of COLLAR using the COCO-POS and COCO-MIG [51]. We randomly select 300 samples from the COCO-2014 dataset [24]. For each sample, 8 random seeds are used for generation, yielding a total of 2,400 images per method for evaluation. In COCO-MIG and COCO-POS, we assess local fidelity with respect to attribute binding, spatial accuracy, semantic consistency, and image quality under depth and Canny maps. In Section 4.8 and Section C, we examine COLLAR's integration versatility. Specifically, we conduct a comparative analysis between the original backbone models and their COLLAR-enhanced versions.

4.4 Evaluation Metrics

We use the following metrics for measuring the performance of different models: (1) **Spatial Fidelity**: mean Intersection over Union (mIoU) and mean Average Precision (mAP) are employed to assess the grounding accuracy between synthesized objects and the prescribed layouts. To extract object-level bounding boxes for evaluation, we utilize SAM3 [4] as an off-the-shelf detector. Furthermore, the Object Success Ratio (OSR) and the Image Success Ratio (ISR) are computed to quantitatively evaluate multi-instance generation. OSR denotes the ratio of correctly rendered instances, while ISR indicates the percentage of images where all objects are correctly rendered. For COCO-POS (Table 1), ISR and OSR evaluate category correctness only, without considering color attributes. For COCO-MIG (Table 2), a successful instance must have both the correct category and color attribute; (2) **Semantic Consistency**: Beyond spatial accuracy, we employ Global CLIP (G-CLIP), Local CLIP (L-CLIP), HPSv3 [26], and PickScore [18] to evaluate text-image alignment; (3) **Image Quality**: We utilize the Fréchet Inception Distance (FID) to evaluate the visual quality and photorealism of the synthesized images.

4.5 Quantitative Comparison

Table 1 and Table 2 collectively present a comprehensive evaluation of our proposed COLLAR, demonstrating its effectiveness in balancing precise spatial condition adherence with high generative fidelity. As detailed in Table 1, under the Depth modality, our method achieves the best FID (16.3) and highly competitive aesthetic performance, ranking second in HPSv3 (6.15). On COCO-MIG under depth conditioning, NoiseCollage achieves slightly higher OSR and mIoU than COLLAR, but has a higher FID of 20.4 compared with 16.3 for COLLAR. Its advantage in spatial alignment may come from the dense spatial cues in depth maps. The robustness of COLLAR is more evidently observed under the challenging Canny modality. Without dense spatial cues, several baseline models struggle to maintain structural integrity. In contrast, COLLAR achieves the highest OSR (68.9) and ISR (40.8) on COCO-POS. On COCO-MIG, it obtains the highest PickScore (0.153) and a competitive FID of 18.2. To further validate this capability, Table 2 provides a fine-grained evaluation on multi-instance generation. As the scene complexity increases from n_2 to n_6 objects, methods such as Regional-FLUX and GrounDit experience noticeable performance degradation. In contrast, COLLAR generally maintains the highest or second-highest success rates across both modalities. Particularly under the Canny constraint, our approach achieves the highest OSR across all instance tiers while Table 1 reports the highest overall COCO-MIG OSR (51.8) and ISR (22.4).

Table 1: Quantitative comparisons on the COCO-POS and COCO-MIG benchmarks. COCO-POS spatial metrics evaluate category correctness only, whereas COCO-MIG spatial metrics require both the correct category and attribute. **Bold** and underline represent the best and second-best methods, respectively. The two Canny FIDs reported as 18.2 are not assigned a second-place marker.

		COCO-POS				COCO-MIG								
Input	Method	ISR↑	OSR↑	mIoU↑	mAP↑	G-CLIP↑	L-CLIP↑	ISR↑	OSR↑	mIoU↑	mAP↑	FID↓	HPSv3↑	Pick↑
Depth	FLUX	62.6	81.7	74.9	83.2	31.40	26.97	24.1	46.0	43.7	47.5	/	6.14	**0.156**
	3DIS	63.4	82.9	75.5	83.8	31.46	27.85	42.3	65.9	61.0	68.2	18.3	5.88	0.143
	Regional-FLUX	64.9	83.4	76.0	84.9	30.79	27.65	37.6	63.8	59.2	66.0	19.8	6.00	0.135
	NoiseCollage	65.7	84.0	76.4	85.9	31.58	**28.23**	**47.8**	**71.7**	**66.0**	**74.9**	20.4	5.83	0.140
	GrounDit	65.3	83.9	75.8	85.3	31.21	27.98	43.0	69.3	63.3	71.1	19.6	5.64	0.126
	DreamRenderer	66.5	**85.8**	**78.2**	86.6	**31.73**	28.15	42.0	68.1	63.1	70.0	17.6	**6.18**	0.151
	Ours	**66.7**	84.8	76.9	**86.7**	31.63	28.16	46.8	71.0	65.0	73.8	**16.3**	6.15	0.146
Canny	FLUX	32.6	60.0	56.3	62.1	30.23	25.59	9.0	29.6	31.8	32.6	/	6.04	0.149
	3DIS	34.3	62.3	57.4	65.4	30.21	26.41	16.5	42.5	41.8	45.5	18.2	6.02	0.144
	Regional-FLUX	36.4	64.5	59.3	66.5	29.63	26.41	15.5	42.8	41.9	45.2	19.3	5.91	0.132
	NoiseCollage	34.6	63.0	57.8	65.1	30.36	26.43	16.2	43.7	42.3	46.8	20.7	5.68	0.139
	GrounDit	36.8	64.9	59.2	67.3	30.27	26.81	19.3	47.1	45.0	49.6	19.2	5.86	0.131
	DreamRenderer	39.9	67.3	61.7	69.4	30.66	26.93	17.5	45.4	44.3	47.8	**17.3**	**6.25**	0.151
	Ours	**40.8**	**68.9**	**62.2**	**71.9**	**30.80**	**27.21**	**22.4**	**51.8**	**48.6**	**55.5**	18.2	6.24	**0.153**



Figure 3: Qualitative comparison between our method and state-of-the-art baselines. Visualization results demonstrate that COLLAR effectively preserves local structural integrity while maintaining precise attribute alignment.

4.6 Qualitative Comparison

Figure 3 illustrates the qualitative comparison between our COLLAR and several state-of-the-art baselines under various spatial conditions. While the vanilla FLUX model generates high-quality images, it often struggles to follow spatial layouts. Methods such as 3DIS and DreamRenderer occasionally lead to semantic leakage, especially in controlling small-scale instances. Furthermore, Regional-FLUX and NoiseCollage lack explicit bidirectional interaction between local and global features. This may contribute to visual artifacts in complex scenes. In contrast, our method demonstrates superior spatial fidelity and object-level precision. As shown in the results, generated instances more closely match the input descriptions, with clearer boundaries in overlapping regions. COLLAR also helps preserve instance attributes and overall image realism. The higher visual quality of small objects further validates our model's robust control capability. In summary, these qualitative results

Table 2: Quantitative comparisons on the COCO-MIG benchmark.

Input	Method	Image Success Ratio↑					Object Success Ratio↑				
		n_2	n_3	n_4	n_5	n_6	n_2	n_3	n_4	n_5	n_6
Depth	FLUX	44.6	18.7	6.1	1.0	0.0	62.1	50.6	42.3	39.0	29.0
	3DIS	63.1	38.9	25.6	13.0	16.1	77.7	71.9	62.6	55.2	54.7
	Regional-FLUX	59.2	35.8	18.3	9.4	7.1	74.4	68.8	62.1	57.5	50.9
	NoiseCollage	**66.8**	**44.5**	**37.8**	21.9	16.4	**79.6**	**74.8**	**73.3**	66.9	60.5
	GrounDit	61.4	41.1	28.5	16.1	17.9	76.5	71.7	68.3	63.6	**62.3**
	DreamRenderer	63.7	40.2	24.7	8.9	12.5	77.8	71.7	64.9	62.2	58.7
	Ours	65.9	42.6	32.8	**22.9**	**20.7**	79.0	73.3	69.5	**67.7**	62.1
Canny	FLUX	18.4	5.0	0.6	0.0	0.0	37.9	34.7	25.6	25.3	20.1
	3DIS	26.7	14.6	6.7	2.1	5.7	49.6	48.3	38.0	34.5	36.5
	Regional-FLUX	28.8	11.0	3.5	**3.6**	0.7	49.8	46.1	40.0	39.5	35.5
	NoiseCollage	25.1	17.3	9.0	1.0	1.1	47.5	50.3	42.6	40.4	35.4
	GrounDit	31.0	16.1	9.9	2.1	**7.9**	49.9	49.6	47.0	41.4	44.4
	DreamRenderer	29.5	14.2	8.7	2.6	2.1	51.5	49.8	43.5	39.0	39.0
	Ours	**34.6**	**22.9**	**11.3**	2.6	5.0	**56.1**	**55.7**	**50.7**	**43.9**	**48.1**

Table 3: Component configurations at $K = 50$. OFI and OFI+BFI use $S = 5$; full CFI uses $S = 10$.

Input	Method	L-CLIP↑	ISR↑	OSR↑ n_3	n_4	n_5	mIoU↑	mAP↑	FID↓	HPSv3↑
Depth	FLUX	26.97	24.1	50.6	42.3	39.0	43.7	47.5	/	6.14
	FLUX + CSSA + OFI	28.1	44.6	73.6	67.7	64.5	64.1	72.2	16.6	6.14
	FLUX + CSSA + OFI + BFI	28.1	44.2	73.4	67.7	64.2	64.2	72.2	16.2	6.22
	FLUX + CSSA + CFI	**28.2**	**45.3**	**73.8**	**68.3**	**65.6**	**64.5**	**73.1**	16.8	6.17

confirm that our approach effectively balances high-fidelity image synthesis with rigorous object-level controllability, outperforming existing methods in complex multi-instance generation.

4.7 Ablation Study

Table 3 compares component configurations with $K = 50$. CSSA with Object-level Feature Injection (OFI) improves spatial control over FLUX. Adding Background Feature Injection (BFI) improves FID and HPSv3, with a small decrease in ISR. The full CFI configuration attains the highest spatial scores in this comparison, but uses a larger maximum injection window ($S = 10$) than OFI and OFI+BFI ($S = 5$). Its improvement therefore reflects both the component configuration and the guidance schedule. At the same $S = 5$ limit, Table 10 reports lower ISR/mAP for CFI (42.9/71.4) than OFI (44.6/72.2), alongside better FID (15.6 vs. 16.6) and HPSv3 (6.27 vs. 6.14). These results demonstrate a trade-off between spatial control and image quality. Additional hyperparameter ablations appear in Section B.

Inference Efficiency Analysis. We evaluate 35 depth-conditioned COCO-MIG samples with six objects each on an NVIDIA A800 GPU. Table 4 reports latency and peak VRAM. COLLAR takes 19.01 s, compared with 25.15 s for Regional-FLUX, reducing latency by 24.4%, while using more memory (35.57 vs. 33.89 GiB). It is also faster than NoiseCollage and DreamRenderer, but slower than FLUX and 3DIS. Early stopping reduces COLLAR's latency from 30.23 to 19.01 s (37.1%) without reducing peak VRAM. These measurements characterize the computation cost alongside the quality and control results in Table 1.

Table 4: Computational efficiency comparison (1024×1024). ES: early-stopping.

Methods	FLUX	3DIS	Regional-FLUX	GrounDit	NoiseCollage	DreamRenderer	Ours(w/o ES)	Ours
Inference Time(s)	10.12	16.83	25.15	19.61	39.81	36.71	30.23	19.01
Peak VRAM(GiB)	33.82	33.83	33.89	33.85	33.86	33.84	35.57	35.57

4.8 Integration with other FLUX-based Conditional Methods

Table 5 demonstrates the compatibility and effectiveness of our proposed method when integrated with the state-of-the-art conditional generation frameworks OminiControl [38] and EasyControl [48]. The quantitative results reveal that our cascaded refinement strategy improves the object-level

controllability of both baselines, particularly in instance rendering and spatial alignment. Notably, when applied to OminiControl, our method yields a substantial increase in ISR (from 18.62 to 27.00) and mAP (from 44.70 to 57.77). The performance gains are even more pronounced for EasyControl, where integrating our strategy increases the OSR from 47.61 to 64.93 and increases the ISR from 25.46 to 36.83. These results show that COLLAR improves object-level control on both backbones, with comparable G-CLIP and higher L-CLIP scores. Such quantitative leaps are further corroborated by the qualitative visualizations (Figure 4), which clearly illustrate our method's superior ability to harmonize complex multi-object layouts and preserve geometric fidelity compared to the original baselines.

Table 5: Integration with FLUX-based conditional generation methods on COCO-MIG. Within each backbone pair, **bold** denotes the better result.

Input	Method	G-CLIP↑	L-CLIP↑	ISR↑	OSR↑	mIoU↑	mAP↑
Depth	OminiControl	31.22	27.25	18.62	43.50	41.78	44.70
	OminiControl + Ours	**31.30**	**27.88**	**27.00**	**56.02**	**53.16**	**57.77**
	EasyControl	**30.97**	26.24	25.46	47.61	45.25	48.54
	EasyControl + Ours	30.94	**27.34**	**36.83**	**64.93**	**59.78**	**66.37**

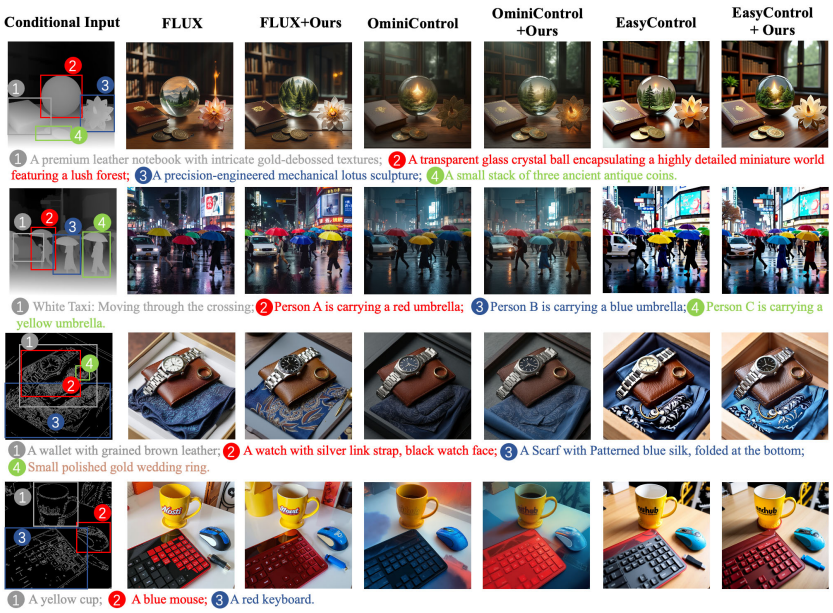


Figure 4: Qualitative results demonstrating the generalizability of our method as a plug-and-play module.

5 Conclusion

We present COLLAR, a training-free, plug-and-play framework for precise instance-level control in spatially-conditioned image generation. It introduces two core modules: (1) Cross-Scale Semantic Alignment, which leverages an extended Field-of-View branch to seamlessly bridge local semantics with global spatial contexts; and (2) Cyclic Feature Injection, which employs a frequency-based adaptive strategy for the seamless blending of local details into the global backbone. Experiments show that COLLAR achieves competitive performance and can be integrated into different generative models.

Limitation. COLLAR's local and extended-FoV branches increase computation relative to FLUX and 3DIS. Although it runs faster than NoiseCollage and DreamRenderer in our comparison, its peak memory usage is higher. Lightweight latent fusion or inference distillation may reduce this cost. Resolving conflicts between local textual semantics and spatial geometry also remains challenging; a semantic-structural-aware feature-scaling strategy is a direction for future work.

Acknowledgments

We would like to thank the reviewers for their thoughtful comments. The work is supported in part by the Zhejiang "Leading Goose" R&D Program (No. 2025C02211) and the Ningbo "Yongjiang Science Innovation 2035" Key Technology Breakthrough Plan (Nos. 2024Z254 and 2025Z056).

References

[1] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. 2022. eDiff-I: Text-to-image diffusion models with an ensemble of expert denoisers. *Preprint*, arXiv:2211.01324.

[2] Black Forest Labs. 2024. Introducing FLUX.1 tools: Depth and Canny structural conditioning. `https://bfl.ai/blog/24-11-21-tools`. Official release page for FLUX.1 Depth and FLUX.1 Canny; accessed 2026-08-23.

[3] Kendrick Boyd, Kevin H. Eng, and C. David Page. 2013. Area under the precision-recall curve: Point estimates and confidence intervals. In *Machine Learning and Knowledge Discovery in Databases*, volume 8190 of *Lecture Notes in Computer Science*, pages 451–466. Springer.

[4] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, and 19 others. 2025. Sam 3: Segment anything with concepts. *Preprint*, arXiv:2511.16719.

[5] Anthony Chen, Jianjin Xu, Wenzhao Zheng, Gaole Dai, Yida Wang, Renrui Zhang, Haofan Wang, and Shanghang Zhang. 2024. Training-free regional prompting for diffusion transformers. *Preprint*, arXiv:2411.02395.

[6] Hongyu Chen, Yiqi Gao, Min Zhou, Peng Wang, Xubin Li, Tiezheng Ge, and Bo Zheng. 2024. Enhancing prompt following with visual control through training-free mask-guided diffusion. *Preprint*, arXiv:2404.14768.

[7] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. 2024. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *Proceedings of the International Conference on Learning Representations*.

[8] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. 2024. Anydoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6593–6602.

[9] Yiyun Chen and Weikai Yang. 2025. Refadgen: High-fidelity advertising image generation. *Preprint*, arXiv:2508.11695.

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

[11] Bo Cheng, Yuhang Ma, Liebucha Wu, Shanyuan Liu, Ao Ma, Xiaoyu Wu, Dawei Leng, and Yuhui Yin. 2024. Hico: Hierarchical controllable diffusion model for layout-to-image generation. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

[12] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 12606–12633. PMLR.

[13] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. 2024. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *International Conference on Learning Representations*.

[14] Runze He, Bo Cheng, Yuhang Ma, Qingxiang Jia, Shanyuan Liu, Ao Ma, Xiaoyu Wu, Liebucha Wu, Dawei Leng, and Yuhui Yin. 2025. Plangen: Towards unified layout planning and image generation in auto-regressive vision language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18143–18154.

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851.

[16] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

[17] Minghui Hu, Jianbin Zheng, Daqing Liu, Chuanxia Zheng, Chaoyue Wang, Dacheng Tao, and Tat-Jen Cham. 2023. Cocktail: Mixing multi-modality controls for text-conditional image generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

[18] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. 2023. Pick-a-pic: an open dataset of user preferences for text-to-image generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

[19] Phillip Y. Lee, Taehoon Yoon, and Minhyuk Sung. 2024. Groundit: grounding diffusion transformers via noisy patch transplantation. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

[20] Seung Hyun Lee, Jijun Jiang, Yiran Xu, Zhuofang Li, Junjie Ke, Yinxiao Li, Junfeng He, Steven Hickson, Katie Datsenko, Sangpil Kim, Ming-Hsuan Yang, Irfan Essa, and Feng Yang. 2025. Cropper: Vision-language model for image cropping through in-context learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 30010–30019.

[21] Danfeng Li, Hui Zhang, Sheng Wang, Jiacheng Li, and Zuxuan Wu. 2025. Seg2any: Open-set segmentation-mask-to-image generation with precise shape and semantic control. In *Proceedings of the 39th Conference on Neural Information Processing Systems*.

[22] Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. 2024. Controlnet++: Improving conditional controls with efficient consistency feedback. In *Proceedings of the European Conference on Computer Vision*, pages 129–147. Springer.

[23] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. 2023. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521.

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pages 740–755. Springer.

[25] Xiang Liu, Junjun Jiang, Wei Han, Kui Jiang, and Xianming Liu. 2026. Cross-controlnet: Training-free fusion of multiple conditions for text-to-image generation. In *The Fourteenth International Conference on Learning Representations*.

[26] Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. 2025. Hpsv3: Towards wide-spectrum human preference score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15086–15095.

[27] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. 2024. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(5):4296–4304.

[28] William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205.

[29] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. 2016. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 724–732.

[30] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2024. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*.

[31] Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, Stefano Ermon, Yun Fu, and Ran Xu. 2023. Unicontrol: A unified diffusion model for controllable visual generation in the wild. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

[32] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

[33] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr.

[34] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. Pmlr.

[35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695.

[36] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. 2022. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, volume 35, pages 36479–36494.

[37] Takahiro Shirakawa and Seiichi Uchida. 2024. Noisecollage: A layout-aware text-to-image diffusion model based on noise cropping and merging. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8921–8930.

[38] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. 2025. OminiControl: Minimal and universal control for diffusion transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14940–14950.

[39] Shuyuan Tu, Qi Dai, Zhi-Qi Cheng, Han Hu, Xintong Han, Zuxuan Wu, and Yu-Gang Jiang. 2024. Motioneditor: Editing video motion via content-aware diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7882–7891.

[40] Xudong Wang, Trevor Darrell, Sai Saketh Rambhatla, Rohit Girdhar, and Ishan Misra. 2024. Instancediffusion: Instance-level control for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6232–6242.

[41] Qiang Xiang, Shuang Sun, Binglei Li, Dejia Song, Huaxia Li, Yibo Chen, Xu Tang, Yao Hu, and Junping Zhang. 2025. Instanceassemble: Layout-aware image generation via instance assembling attention. In *Advances in Neural Information Processing Systems 38*.

[42] Jiayu Xiao, Henglei Lv, Liang Li, Shuhui Wang, and Qingming Huang. 2024. R&b: Region and boundary aware zero-shot grounded text-to-image generation. In *The Twelfth International Conference on Learning Representations*.

[43] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. 2025. Omnigen: Unified image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13294–13304.

[44] Denis Zavadski, Johann-Friedrich Feiden, and Carsten Rother. 2024. Controlnet-xs: Rethinking the control of text-to-image diffusion models as feedback-control systems. In *Proceedings of the European Conference on Computer Vision*, pages 343–362, Berlin, Heidelberg. Springer-Verlag.

[45] Hui Zhang, Dexiang Hong, Yitong Wang, Jie Shao, Xinglong Wu, Zuxuan Wu, and Yu-Gang Jiang. 2025. Creatilayout: Siamese multimodal diffusion transformer for creative layout-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18487–18497.

[46] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847.

[47] Xinlong Zhang, Zejian Li, Wei Li, Xiaoyu Zhang, Jia Wei, Chengyu Lin, and Yongchuan Tang. 2025. Objctrl: Object-based control relaxation for conditional text-to-image generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pages 10064–10073, New York, NY, USA. Association for Computing Machinery.

[48] Yuxuan Zhang, Yirui Yuan, Yiren Song, Haofan Wang, and Jiaming Liu. 2025. Easycontrol: Adding efficient and flexible control for diffusion transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19513–19524.

[49] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K. Wong. 2023. Uni-controlnet: All-in-one control to text-to-image diffusion models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23.

[50] Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. 2023. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22490–22499.

[51] Dewei Zhou, Mingwei Li, Zongxin Yang, and Yi Yang. 2025. Dreamrenderer: Taming multi-instance attribute control in large-scale text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16712–16722.

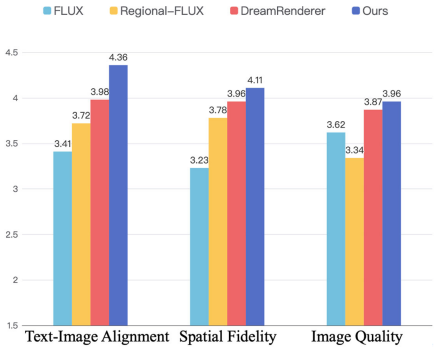
[52] Dewei Zhou, You Li, Fan Ma, Xiaoting Zhang, and Yi Yang. 2024. Migc: Multi-instance generation controller for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6818–6828.

[53] Dewei Zhou, Ji Xie, Zongxin Yang, and Yi Yang. 2024. 3dis: Depth-driven decoupled instance synthesis for text-to-image generation. *Preprint*, arXiv:2410.12669.

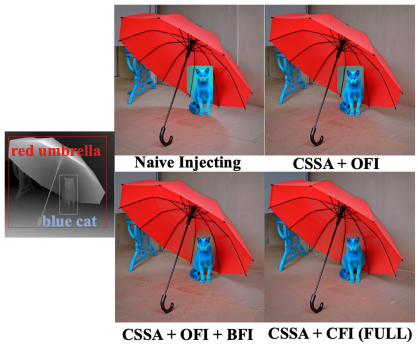
Appendix

A User Study

Figure 5 (a) shows a user study with 20 participants with a background in computer science, comprising 12 Master's students (60%) and 8 Ph.D. candidates (40%). Four methods are selected for evaluation: COLLAR, DreamRenderer, Regional-FLUX, and FLUX. We randomly selected 40 generated images from the COCO-MIG benchmark for each participant and asked them to rate each image on text-image alignment, image quality, and spatial fidelity. The rating scale ranged from 1 to 5 (highest). Fig. 5(a) shows that users preferred our method for spatial fidelity and image quality. A screenshot of the evaluation interface and the instructions provided to participants are shown in Figure 6.



(a) Users were asked to rate: text-image alignment, spatial fidelity, and image quality.



(b) Ablation study of model designs. Our method can achieve spatial fidelity and image quality.

Figure 5: Visualization results of User Study and Ablation Study.

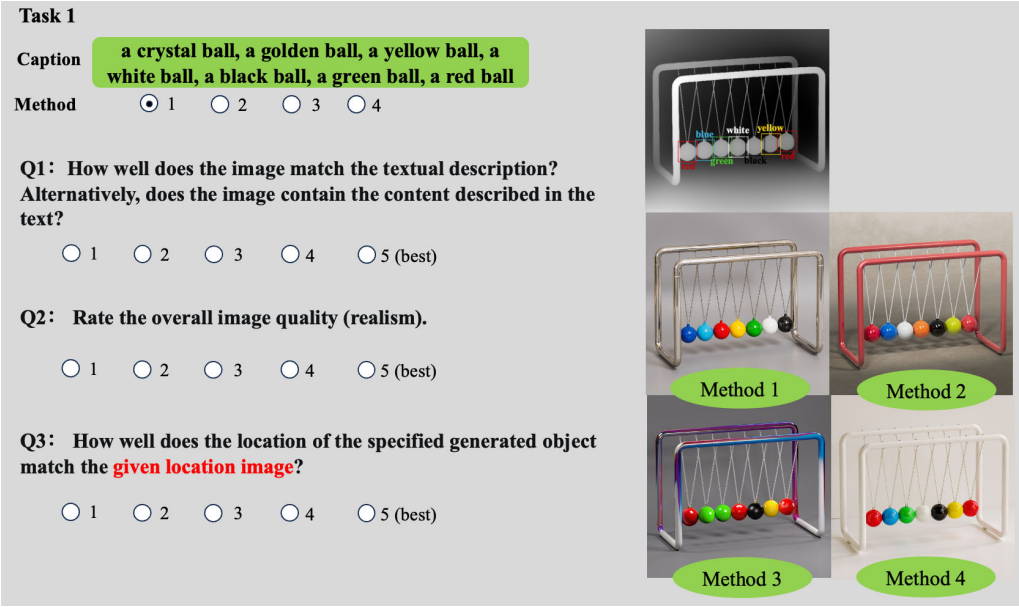


Figure 6: Screenshot of the instruction page and evaluation interface provided to participants during the user study.

B Ablation Study

B.1 Validation of the High-Pass Residual

We implement $\mathcal{H}(z) = z - \mathrm{AvgPool}_{3\times3}(z)$ using stride 1 and padding 1. Local averaging yields a low-pass-like component, while subtraction yields a high-pass-like residual. We examine their frequency content using channel-wise spatial 2-D FFTs of the final latents. Let r denote radial frequency normalized by the corner Nyquist radius. We define $r \leq 0.25$ as the low band and $r \geq 0.50$ as the high band. Each energy fraction is normalized by the corresponding component's total power. The middle band is omitted, so the two fractions need not sum to 100%.

Table 6: Frequency-band energy fractions (%). Avg.: average-pooled component; Res.: residual.

Condition	Kernel	Avg. low	Avg. high	Res. low	Res. high
Depth	3×3	96.39	0.79	2.17	74.20
Depth	5×5	98.64	0.39	9.94	48.00
Canny	3×3	95.59	0.93	2.49	72.60
Canny	5×5	98.28	0.49	11.12	45.32

Table 6 shows that local averaging concentrates energy in the low-frequency band. Compared with the 5×5 kernel, the 3×3 kernel produces a residual with a higher proportion of high-frequency energy and a lower proportion of low-frequency energy under both Depth and Canny conditioning. These results support our choice of the 3×3 kernel for extracting the high-pass residual.

We then examine whether the residual captures contour information after decoding. At the final denoising step, we decode the average-pooled component and the residual separately. The image decoded from the full latent serves as the reference. For Pixel Edge AP [3], we compare Scharr gradient magnitudes in each component image with Canny edges from the reference image. Boundary F1 [29] measures contour agreement with the reference image. As shown in Table 7, the residual achieves higher scores on both metrics under Depth and Canny conditioning.

Table 7: Contour responses of decoded latent components.

Condition	Component	Pixel Edge AP↑	Boundary F1↑
Depth	Average-pooled	0.06	0.23
	High-pass residual	**0.32**	**0.73**
Canny	Average-pooled	0.09	0.34
	High-pass residual	**0.48**	**0.83**

B.2 Effect of the Early-Stopping Signal

We compare full-latent and high-pass-residual signals for early stopping using the same independent set of 600 test samples as the object-size evaluation in Section D.1. Table 8 reports spatial control, image quality, and mean injection steps. The high-pass signal reduces mean injection steps from 8.25 to 7.29 and improves HPSv3 and PickScore, with decreases of only 0.13 and 0.31 percentage points in OSR and mIoU, respectively.

Table 8: Early-stopping signal comparison on the independent 600-sample test set under Depth conditioning.

Signal	OSR↑	mIoU↑	HPSv3↑	PickScore↑	Mean steps↓
Full latent	**69.77**	**64.46**	5.97	0.49	8.25
High-pass	69.64	64.15	**6.19**	**0.51**	**7.29**

We further group objects by their bounding-box area relative to image area, denoted by a. The stop rate is the percentage of objects that meet the similarity-based stopping criterion before the maximum injection step $S = 10$. Table 9 shows higher stop rates for the high-pass signal in all three size ranges, with larger objects stopping more frequently.

Table 9: Early-stop rates (%) by relative object area a under Depth conditioning.

Signal	$a < 5\%$	$5\% \leq a < 15\%$	$a \geq 15\%$
Full latent	14.20	34.01	51.14
High-pass	**15.34**	**41.57**	**70.83**

B.3 Hyperparameter Ablations

Effect of FoV Expansion Factor γ. The parameter γ determines the proportion of the extended field-of-view in the CSSA module. In Table 10, as γ increases from 0.2 to 1.0, we observe a consistent reduction in FID ($16.2 \rightarrow 15.3$). This improvement suggests that incorporating more surrounding spatial context through a larger FoV expansion may help preserve image quality. However, an excessively large FoV tends to relax the local spatial constraints, resulting in a slight decrease in ISR ($43.4 \rightarrow 42.3$) and mIoU ($63.4 \rightarrow 62.8$). We adopt $\gamma = 0.5$ as an empirical compromise between image quality and spatial fidelity.

Table 10: Ablation study on model designs and hyperparameters (under the injecting layers K =50).

Input	Method	L-CLIP↑	ISR↑	OSR↑ n_3	n_4	n_5	mIoU↑	mAP↑	FID↓	HPSv3↑
Depth	FLUX	26.97	24.1	50.6	42.3	39.0	43.7	47.5	/	6.14
	FLUX + CSSA + OFI									
	• $\gamma = 0.5, S = 5$	28.1	44.6	73.6	67.7	64.5	64.1	72.2	16.6	6.14
	FLUX + CSSA + OFI + BFI									
	• $\gamma = 0.5, S = 5$	28.1	44.2	73.4	67.7	64.2	64.2	72.2	16.2	6.22
	FLUX + CSSA + CFI									
	• $\gamma = 0.2, S = 5, \sigma = 0.85$	28.1	43.4	73.1	67.2	64.3	63.4	71.6	16.2	6.20
	• $\gamma = 0.5, S = 5, \sigma = 0.85$	28.1	42.9	72.8	66.5	63.6	63.3	71.4	15.6	6.27
	• $\gamma = 1.0, S = 5, \sigma = 0.85$	28.1	42.3	72.6	65.0	62.6	62.8	70.9	**15.3**	**6.31**
	• $\gamma = 0.5, S = 10, \sigma = 0.80$	28.1	41.7	72.5	65.0	62.5	62.7	70.7	15.8	6.25
	• $\gamma = 0.5, S = 10, \sigma = 0.85$	**28.2**	**45.3**	**73.8**	**68.3**	**65.6**	**64.5**	**73.1**	16.8	6.17

Effect of Guidance Step Control (σ and S). The parameters S and σ jointly determine when feature injection terminates. S sets the maximum number of guidance steps, while σ defines the similarity threshold for early stopping. With $\gamma = 0.5$ and $\sigma = 0.85$, increasing S from 5 to 10 improves mAP from 71.4 to 73.1 and ISR from 42.9 to 45.3. At $S = 10$, increasing σ from 0.80 to 0.85 improves mIoU from 62.7 to 64.5. These results suggest that permitting longer feature injection can improve spatial fidelity. However, the corresponding increase in FID and decrease in HPSv3 suggest a potential trade-off between spatial control and image quality.

Effect of the Number of Injecting Layers K and Weight λ. Table 11 provides a comprehensive quantitative analysis of the number of injecting layers K and the balance factor λ, revealing a clear performance scaling effect. Other hyperparameters follow the selected configuration presented in table 10. As K increases from 20 to 57, spatial accuracy metrics such as mIoU and mAP exhibit an overall upward trend, indicating that deeper feature injection is pivotal for reinforcing geometric constraints within the DiT architecture. Increasing K from 50 to 57 improves mIoU and mAP while reducing FID from 16.8 to 16.3. With $K = 57$, setting $\lambda = 0.5$ increases ISR to 47.2 and mAP to 74.9. This configuration also demonstrates enhanced robustness in complex multi-object scenarios, with the OSR for n_5 rising from 67.7 to 69.7, indicating improved instance-level control. Overall, we select $K = 57$ and $\lambda = 0$ for the main experiments because this configuration provides a better balance between image quality and spatial control. Compared with $\lambda = 0.5$, it reduces FID from 17.1 to 16.3 while retaining competitive instance-level control.

B.4 Effect of Object Count on Inference Cost

We evaluate computational cost using 35 samples for each object count and report mean inference time and mean peak GPU memory in Table 12. We use the same inference implementation as in Table 4. As the number of objects increases from two to six, mean inference time increases from 14.05 to 19.01 seconds, while mean peak GPU memory increases from 34.96 to 35.57 GiB. The measurements show approximately linear runtime growth over the tested range.

Table 11: Ablation study on the number of injecting layers K and weight λ.

Input	Method	L-CLIP↑	ISR↑	OSR↑ n_3	OSR↑ n_4	OSR↑ n_5	mIoU↑	mAP↑	FID↓
Depth	FLUX + CSSA + CFI								
	• $K = 20, \lambda = 0$	28.05	43.0	73.3	67.2	62.2	63.5	71.6	**15.9**
	• $K = 30, \lambda = 0$	28.08	43.3	72.4	66.4	64.0	63.7	71.7	16.6
	• $K = 40, \lambda = 0$	28.15	44.3	73.6	68.0	65.1	64.3	72.7	16.7
	• $K = 50, \lambda = 0$	28.2	45.3	73.8	68.3	65.6	64.5	73.1	16.8
	• $K = 57, \lambda = 0$ (Ours)	28.16	46.8	73.3	69.5	67.7	65.0	73.8	16.3
	• $K = 57, \lambda = 0.5$	**28.23**	**47.2**	**74.4**	**69.9**	**69.7**	**65.6**	**74.9**	17.1

Table 12: Inference cost for different numbers of objects. Each column averages 35 samples.

Number of objects	2	3	4	5	6
Mean inference time (s)	14.05	15.29	16.31	17.70	19.01
Mean peak GPU memory (GiB)	34.96	35.10	35.19	35.28	35.57

C Integration with SD3-based Conditional Methods

To demonstrate that our method can be generalized to other DiT-based models, Table 13 presents a quantitative comparison utilizing Stable Diffusion 3 (SD3) equipped with ControlNet as the backbone. The results clearly indicate that our method outperforms both the vanilla baseline and DreamRenderer in precise spatial control. Specifically, our approach achieves substantial improvements across all primary layout-alignment metrics, yielding the highest ISR (27.92), OSR (57.25), and mIoU (52.73). Furthermore, our method secures the best local semantic alignment. While the vanilla SD3+ControlNet retains slightly higher global metrics (e.g., G-CLIP and PickScore)—which is typical for native models generating relatively unconstrained scenes—it shows limited multi-instance spatial adherence. In contrast, our method successfully bridges this gap, achieving the best object-level controllability among the compared methods on the SD3 architecture while maintaining highly competitive overall aesthetic quality.

Table 13: Quantitative comparisons on COCO-MIG using SD3 with ControlNet. **Bold** and underline denote the best and second-best reported results, respectively.

Input	Method	G-CLIP↑	L-CLIP↑	ISR↑	OSR↑	mIoU↑	mAP↑	FID↓	HPSv3↑	Pick↑
Depth	SD3 + ControlNet	**33.22**	25.68	9.96	31.57	33.43	31.25	/	**5.502**	**0.4033**
	DreamRenderer	32.94	<u>26.50</u>	<u>17.50</u>	<u>46.16</u>	<u>44.29</u>	<u>47.44</u>	<u>21.53</u>	3.849	0.2429
	Ours	<u>33.02</u>	**27.15**	**27.92**	**57.25**	**52.73**	**58.54**	**18.66**	<u>5.435</u>	<u>0.3537</u>

D Additional Evaluation

D.1 Performance across Object Sizes

We evaluate OSR on COCO-MIG across three object-size groups, defined by the ratio a of target bounding-box area to image area: small ($a < 5\%$), medium ($5\% \leq a < 15\%$), and large ($a \geq 15\%$). We use a separate set of 600 randomly selected test samples, also used in the early-stopping analysis in Section B.2. An object is counted as successful only when both its category and color attribute are correct. Table 14 reports the results. With Canny conditioning, COLLAR achieves the highest OSR across all three size groups.

Table 14: COCO-MIG OSR (%) by object size.

Condition	Method	Small ($< 5\%$)	Medium (5–15%)	Large ($\geq 15\%$)
Depth	NoiseCollage	65.9	<u>73.6</u>	**73.2**
	GrounDiT	**67.4**	71.9	67.6
	DreamRenderer	62.9	68.5	<u>71.4</u>
	COLLAR	<u>66.7</u>	**73.7**	71.0
Canny	NoiseCollage	32.9	43.4	<u>53.0</u>
	GrounDiT	<u>42.7</u>	<u>48.0</u>	49.5
	DreamRenderer	36.4	45.2	52.8
	COLLAR	**47.4**	**52.8**	**54.0**

D.2 Additional Qualitative Results

Figures 7 and 8 show additional qualitative comparisons under depth and Canny conditioning, respectively.

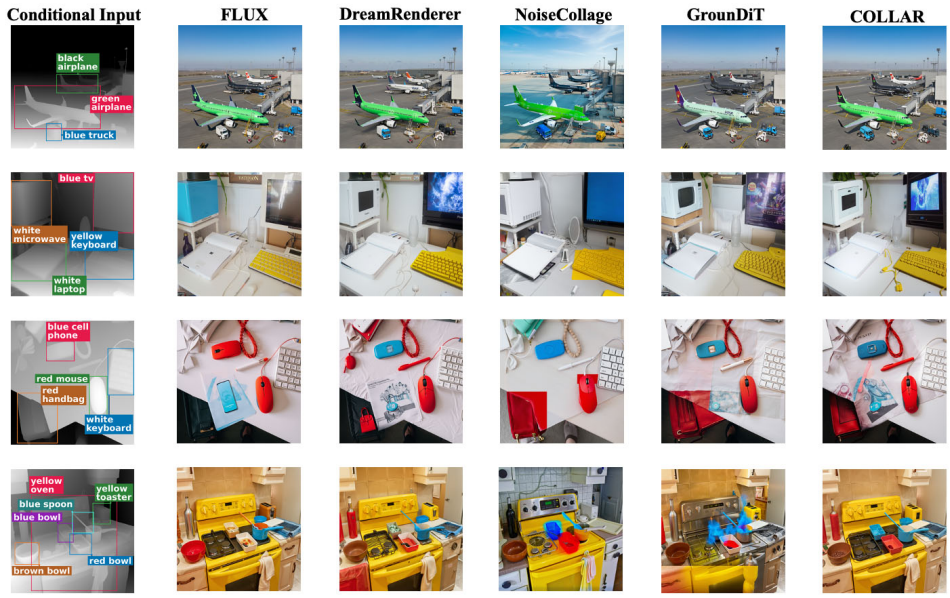


Figure 7: Additional qualitative comparisons under depth conditioning.

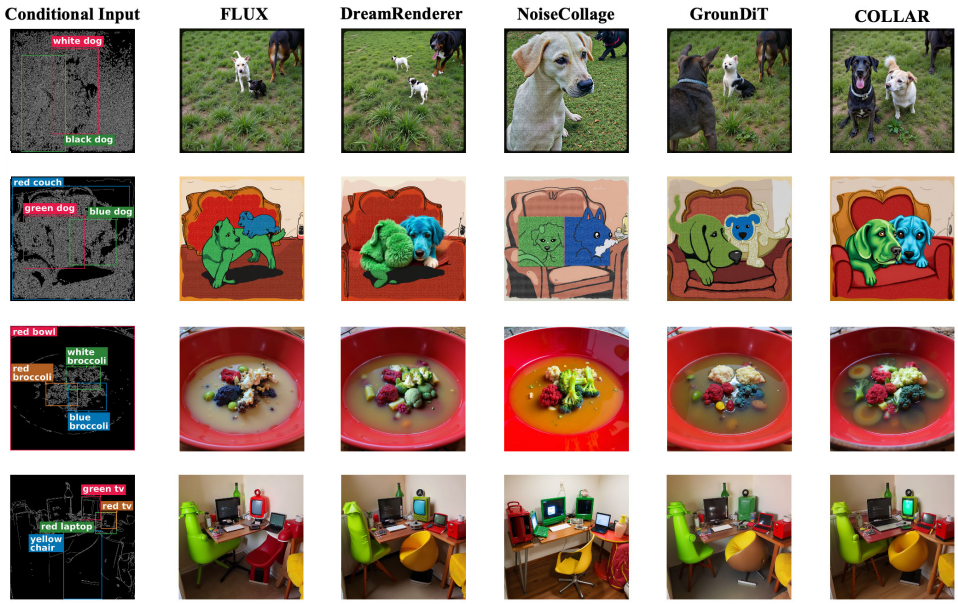


Figure 8: Additional qualitative comparisons under Canny conditioning.

E Pseudocode

Algorithm 1 illustrates the complete inference process of our method.

Algorithm 1 COLLAR: Cascaded Object-Level Latent Refinement

Input: Global prompt P, conditional image input C, bounding boxes $\{B_i\}$, object descriptions $\{O_i\}$.

Parameters: Threshold σ, maximum guidance steps S, FoV expansion factor γ, weight λ, high-pass filter $\mathcal{H}(\cdot)$, injection-layer set $\mathcal{K}$.
Initialize: $z_G^{(T)}, \{z_{i,L}^{(T)},\}, \{z_{i,E}^{(T)},\} \sim \mathcal{N}(0, \mathbf{I})$ {Global, Local, and Ex-FoV latents}
$\mathcal{A} \leftarrow \{1, \ldots, N\}$ {Set of active instances for feature injection}
$C_G \leftarrow C$
for $t = T, T-1, \ldots, 1$ **do**
  for each instance $i \in \mathcal{A}$ **do**
    $B_{i,E} \leftarrow \text{Expand}(B_i, \gamma)$ {Extended-FoV boxes}
    $C_{i,L} \leftarrow \text{Crop}(C, B_i), \quad C_{i,E} \leftarrow \text{Crop}(C, B_{i,E})$ {Branch conditions}
    $M_{i,E}, M_{i,G} \leftarrow \text{BuildMasks}(B_i, B_{i,E})$ {Ex-FoV and globally aligned masks}
    // Step 1: Local Semantic Extraction
    $\epsilon_{i,L}^{(t)}, \{(Q,K,V)_{i,L}^{(t,k)}\}_{k\in\mathcal{K}} \leftarrow \text{DiT}_L(z_{i,L}^{(t)}, O_i, C_{i,L})$ {Extract QKV from Local branch}
    $z_{i,L}^{(t-1)} \leftarrow \text{Scheduler}(z_{i,L}^{(t)}, \epsilon_{i,L}^{(t)}, t)$
    // Step 2: Global Background Latent Injecting
    $z_{i,E}^{(t)} \leftarrow \text{Fuse}(z_{i,E}^{(t)}, z_G^{(t)}, [B_{i,E}])$ {Inject global background into Ex-FoV latent}
    // Step 3: Extended-FoV Feature Fusion
    $\epsilon_{i,E}^{(t)}, \{f_{i,E}^{(t,k)}\}_{k\in\mathcal{K}} \leftarrow \text{DiT}_E(z_{i,E}^{(t)}, \{(Q,K,V)_{i,L}^{(t,k)}\}_{k\in\mathcal{K}}, O_i, C_{i,E})$
    $z_{i,E}^{(t-1)} \leftarrow \text{Scheduler}(z_{i,E}^{(t)}, \epsilon_{i,E}^{(t)}, t)$
    // Step 4: Structural Monitoring & Early-Stopping
    $\mathcal{S}_i^{(t)} \leftarrow \text{StopScore}\left(\mathcal{H}(z_G^{(t)} \odot M_{i,G}), \mathcal{H}(z_{i,E}^{(t)} \odot M_{i,E})\right)$
    if $\mathcal{S}_i^{(t)} \geq \sigma$ **or** $T - t \geq S$ **then**
      $\mathcal{A} \leftarrow \mathcal{A} \setminus \{i\}$ {Halt injection after alignment or S steps}
    end if
  end for
  // Step 5: Local Feature Injecting
  $\epsilon_G^{(t)} \leftarrow \text{DiT}_G(z_G^{(t)}, \{f_{i,E}^{(t,k)}\}_{k\in\mathcal{K}, i\in\mathcal{A}}, \{M_{i,G}\}_{i\in\mathcal{A}}, \lambda, P, C_G)$ {Masked fusion by Equation (10)}
  // Final Global Update
  $z_G^{(t-1)} \leftarrow \text{Scheduler}(z_G^{(t)}, \epsilon_G^{(t)}, t)$
end for
return Final image $x = \text{Decoder}(z_G^0)$

F Societal Impacts

Our proposed framework enables precise, spatially-conditioned image generation with multi-instance control, benefiting applications like creative design and synthetic dataset construction. However, such capabilities could be exploited to produce highly convincing, misleading content or sophisticated deepfakes. Additionally, as our approach builds upon existing generative foundation models, it inevitably inherits the societal biases and stereotypes embedded in their pre-training data. While we focus on addressing challenges in controllable generation, we emphasize that responsible deployment and the development of robust safeguards against privacy, bias, and fairness risks are essential future endeavors.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: See Section 1 and Abstract in the paper.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: See Section 5 in the paper.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [N/A]

Justification: I do not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: Yes. Detailed experimental setups, including hyperparameter configurations, and evaluation metrics, are comprehensively described in Section 4 and Section B.

   Guidelines:

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

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

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

Answer: [Yes]

Justification: Yes. For the double-blind review process, the source code and necessary data are included in the supplemental material. We commit to publicly releasing the code and data upon acceptance.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

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

Answer: [Yes]

Justification: Yes. Since our proposed framework is training-free, training-specific details are not applicable. However, all necessary inference details, including dataset configurations, hyperparameter choices, and evaluation metrics, are exhaustively documented in Section 4 and Section B.

Guidelines:

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

7. **Experiment statistical significance**

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

Answer:[No]

Justification: The paper does not include information about statistical significance of the experiments.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer:[Yes]

   Justification: Yes. The computer resources used for our experiments, including the specific hardware (NVIDIA A800 GPUs), memory constraints, and average execution times for image generation, are specified in Section 4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: Yes. We strictly follow the NeurIPS Code of Ethics, and have transparently discussed potential societal impacts in the paper.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: Yes. Both positive applications and potential negative risks, such as deepfakes, are transparently discussed in Section F.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The paper poses no such risks

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: Yes. All third-party assets, including models and code, are fully cited, and we strictly respect their associated licenses and terms of use.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [Yes]

    Justification: Yes. The source code introduced in this paper is well-documented, with a detailed README and reproduction instructions included in the supplementary material.

    Guidelines:

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

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

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

    Answer: [Yes]

    Justification: Yes. The full text of the instructions provided to participants, and screenshots of the evaluation interface are all included in Section A.

    Guidelines:

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

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

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

    Answer: [Yes]

    Justification: Yes. The user study posed no physical or psychological risks to participants, as it only involved viewing and rating AI-generated images. Participants were informed of the study's nature and their right to withdraw at any time. Given the non-invasive nature of the tasks, this study follows our institution's standard protocols and is exempt from formal IRB review.

    Guidelines:

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

16. **Declaration of LLM usage**

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

    Answer:[N/A]

    Justification: LLM is used only for writing, editing, or formatting purposes.

    Guidelines:

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