SPATIALCOMPOSER: 3D SPATIAL OBJECT INSERTION VIA IMAGE GAUSSIAN COMPOSITION

Anonymous authors
Paper under double-blind review

ABSTRACT

With the rapid advancement of open-world image generation models in recent years, a series of image editing tasks have achieved excellent performance. However, considering object insertion as a representative example, this task still presents three primary challenges. First, the inserted object should maintain identity consistency with the reference object while preserving the original scene in non-edited regions. Second, the spatial position and scale of the inserted object should be reasonable and align with user expectations. Third, the inserted object should harmonize with other image components, typically involving object style and surface illumination harmonization. To address these challenges, we propose SpatialComposer, which leverages depth-aware image Gaussians to construct a spatially-structured scene representation from a single scene image and models object insertion as Gaussian composition, thereby achieving effective preservation of scene and object identity while enabling precise control over the scale and 3D spatial position of the inserted object. Subsequently, based on pre-trained diffusion generative models, we introduce a simple yet effective refinement method for the object harmonization process. By designating only the Gaussian components corresponding to the inserted object as trainable parameters, SpatialComposer avoids unintended modifications to other regions while simultaneously addressing both object-scene integration and scene detail preservation. Furthermore, recognizing that current object insertion benchmarks lack consideration for depth-aware position control, we construct a specialized benchmark featuring high-resolution scene images with substantial depth complexity. Comprehensive evaluations demonstrate that SpatialComposer achieves comparable or superior performance over state-of-the-art object insertion approaches across all three aforementioned challenges.

1 INTRODUCTION

Object insertion is a fundamental task in image editing, defined as the process of seamlessly integrating a specified object into a target location within a scene. With the open-sourcing of recent diffusion-based image generation models (e.g., Stable Diffusion (Rombach et al., 2022b), and FLUX (Labs, 2024)), contemporary approaches to object insertion have evolved significantly beyond previous methods (Tripathi et al., 2019; Zhang et al., 2020; Liu et al., 2021; Niu et al., 2022; Cong et al., 2020; Ling et al., 2021; Sofiiuk et al., 2021; Cong et al., 2022).

Current object insertion methods typically fall into two categories: *training-based methods* (Song et al., 2025; Wang et al., 2025; Chen et al., 2024b) that fine-tune variants of pre-trained inpainting models conditioned on reference object images to control the generated content, and *training-free methods* (Chen et al., 2024c; Wang et al., 2024) that manipulate intermediate feature representations and attention mechanisms during the model's forward inference process. These methods enable users to perform intuitive personalized customization of existing real or synthesized images using arbitrary object references, thereby providing accessible and efficient automated tools for creating and modifying artistic works and visual content.

Despite the promising results achieved by existing methods, object insertion is still facing three main challenges: *1). Object and scene consistency*. Due to the randomness introduced by the forward diffusion process and the information loss introduced by the encoder-decoder downsampling

of latent diffusion models, it is challenging for current methods to maintain detail consistency in terms of both object and scene. *2). Depth and scale controllability*. Existing approaches often control insertion through object-level masks or rectangular bounding box combined with optional text prompts. However, these merely specify two-dimensional positioning and cannot accurately control the depth-related spatial position and scale of the inserted object. Our experimental results demonstrate that even when depth information is provided through text prompts, previous object insertion methods still fail to meet expectations in many cases. *3). Style and illumination harmony*. Current methods often struggle to produce style consistency, natural lighting and color harmonization that are required to achieve seamless visual integration between the inserted object and the scene.

In this paper, we propose a novel depth-aware object insertion method based on Gaussian Kerbl et al. (2023) representation, which we term SpatialComposer. In order to tackle the above three challenges, our approach contains three main steps: Gaussian fitting, Gaussian composition, and Object refinement. In the first step, by leveraging a pre-trained monocular depth estimation network, combined with a back-projection Gaussian initialization strategy, we efficiently construct consistent object and scene Gaussian representations with meaningful spatial structure. In the second step, object insertion is implemented through the composition of object Gaussians with scene Gaussians. Through user-specified scaling and translation operations applied to the object Gaussians, our approach achieves precise control over both the depth and scale of the inserted object. In the third step, following the composition of object and scene representations, we propose a simple yet effective refinement method for inserted objects based on pre-trained inpainting diffusion models and illumination harmonization models. With separable scene Gaussians and object Gaussians, we designate only the object Gaussian components as trainable parameters, thereby preventing unintended modifications to other scene regions. This design circumvents the limitations of pre-trained diffusion generative models regarding resolution constraints and reconstruction fidelity. It also preserves the integrity of unmodified scene areas while harnessing the supervisory signals provided by pre-trained diffusion models for harmonizing inserted objects with the scene.

The existing open-source TF-ICON benchmark (Lu et al., 2023b) for image-guided object insertion, which encompasses both realistic and stylized scenes, consists of generated scene images at 512×512 resolution with relatively low quality and simple spatial structures, where insertion positions do not involve spatial relationships with other components in the scene. Therefore, this benchmark is insufficient to meet our evaluation requirements. To fill this research gap, we also collect a novel benchmark dataset comprising over 200 high-resolution scene images with complex spatial structures and over 200 object images spanning more than 9 categories, including animals, vegetables, food, people, vehicles, and others. The scene images encompass over 100 indoor and outdoor real photographs and over 100 artistic images across multiple artistic styles, including oil painting, cartoon, anime, watercolor, and pixel art. Based on these scene and object collections, we constructed over 200 semantically coherent scene-object insertion cases. We name the dataset Depth-Aware Object Insertion (DAOI) Dataset, and will release it in the near future.

Ablation experiments demonstrate the effectiveness of our proposed image Gaussian representation and initialization method. Using only a single image and its corresponding relative depth estimates, we can rapidly construct scene Gaussians with meaningful depth information while maintaining high-quality image reconstruction. Leveraging the intuitive compositionality of Gaussian representations, scaling and translating the object Gaussians enables precise control of object scale and placement at any spatial location within the scene Gaussians. The proposed refinement module exhibits strong generalization capabilities and excellent performance without requiring additional fine-tuning of pre-trained models. Furthermore, unlike existing methods, our approach is not constrained by scene image resolution and can be applied to ultra-high-resolution images while preserving fine-grained scene details. The method also demonstrates robust performance across diverse artistic styles, confirming its practical applicability for handling various application scenarios.

In summary, our main contributions include: (1) We propose a depth-aware image Gaussian representation and a back-projection initialize strategy to enhance the performance of Gaussian fitting while providing meaningful depth information. This achieves preservation of non-edited scene regions and object identity consistency, while enabling precise control over depth-related spatial positioning and scale during object insertion. To our knowledge, our work is the first to consider depth-related spatial positioning in object insertion. (2) We introduce a simple yet effective object refinement method that achieves robust object refinement without the need for fine-tuning on task-specific data. (3) We collect a high-resolution object insertion benchmark featuring complex

scene spatial structures, providing a high-quality evaluation platform for advancing research in this domain. (4) Experimental results demonstrate the superior performance and robustness of SpatialComposer across diverse stylistic and real scenes.

2 Related Work

2.1 Training-based object insertion

Built upon pre-trained text-to-image or inpainting models, training-based object insertion methods employ techniques such as LoRA (Hu et al., 2022) and Adapters (Houlsby et al., 2019) to fine-tune the models using paired data, i.e., scene images with and without specific objects. These approaches typically introduce new network structures to accept the scene image, reference object, and spatial mask as conditions. Several works (Yang et al., 2023; Song et al., 2022; Chen et al., 2024a; Canet Tarrés et al., 2024; Yuan et al., 2024; Kulal et al., 2023; Zhang et al., 2023a; Song et al., 2024; He et al., 2024) extract features from reference objects through pre-trained visual encoders and trainable modules, then influence UNet features via cross-attention, summation, or custom fusion mechanisms to control generation. Some methods (Zhang et al., 2023b; Chen et al., 2024b) use composite images formed by directly pasting reference objects into scene regions as image conditions. Insert Anything and UniCombine (Song et al., 2025; Wang et al., 2025) adopt Diffusion Transformer (DiT) architectures, enabling operations between latent variables and condition tokens for generation control. To achieve better environmental harmony, Zerocomp (Zhang et al., 2025) trains models conditioned on depth, surface normal, albedo, and shading of both objects and scenes. Objectmate (Winter et al., 2024b) concatenates reference objects and scene into a 2×2 grid as diffusion input to produce coherent insertion results. Anydoor (Chen et al., 2024b) combines a trainable ID extractor for identity features with high-pass filters for spatial information to generate insertion results under dual feature control. While achieving satisfactory results, these methods require large-scale paired training data and suffer from insufficient generalization ability. To avoid the over-reliance on large datasets, DreamCom and DreamEdit (Lu et al., 2023a; Li et al., 2023) leverage DreamBooth's (Ruiz et al., 2023) approach, fine-tuning embeddings with only several reference images. DreamEdit (Li et al., 2023) then employs DDIM Inversion (Mokady et al., 2023) for noising-denoising processes, while DreamCom (Lu et al., 2023a) uses masked attention control for generation. OmniPaint (Yu et al., 2025) reduces data requirements by training separate insertion and removal models, then leveraging their inverse relationship through cycle consistency loss for optimization. To reduce paired data dependency, ObjectDrop (Winter et al., 2024a) first trains an object removal model on a small dataset, and then uses it to collect a large synthetic dataset for insertion model training. Overall, when there exists a significant gap between training and inference data, the preservation of non-edited scene regions and object identity, as well as the harmony between objects and scenes, exhibit substantial degradation. These methods are also unable to control the depth-related spatial positioning of object insertion. In contrast, SpatialComposer is training-free, enables precise control over 3D spatial positioning, and has been validated to deliver stable performance across real-world scenes and a variety of stylistic scenes.

2.2 Training-Free Object Insertion

Instead of parameter updating, training-free object insertion methods manipulate the intermediate features or attention maps during the inference stage. TF-ICON (Lu et al., 2023b) employs a three-branch inference process using initial noise from DDIM inversion of scene and reference object images, along with composite noise created by placing resized object noise into the target scene region. During forward inference, features from scene and object reconstruction branches guide cross-attention computation, with outputs injected into the composite generation branch. Similarly, the method in (Li et al., 2024) uses three-branch inference with self-attention feature fusion to preserve object features while enabling text-guided attribute modification. FreeCompose (Chen et al., 2024c) iteratively optimizes the editing branch through key-value replacement and DDS-loss (Hertz et al., 2023) between reconstruction and editing branches. PrimeComposer (Wang et al., 2024) introduces a correlation diffuser that computes cross-attention between target region features and reference object features within UNet self-attention layers. During denoising, noisy latents are processed by both the correlation diffuser and pre-trained Stable Diffusion UNet, with correlation diffuser attention maps injected into UNet self-attention layers for appearance preservation. Region-constrained

cross-attention ensures object-related prompt tokens only influence the target insertion region. Constrained by the reconstruction and generation capabilities of the underlying foundation models, these methods exhibit limitations in preserving non-edited regions and maintaining object identity. Likewise, they do not provide effective control over the spatial placement of inserted objects.

3 METHOD

The pipeline of SpatialComposer is illustrated in Fig. 1. It comprises three main components: Gaussian fitting, Gaussian composition and object refinement. We begin by introducing the foundational concepts of 3D Gaussians in Sec. 3.1, and we also introduce the latent diffusion model in Appendix A. Subsequently, we provide detailed descriptions of our proposed Depth-Aware Image Gaussian (DA-ImgGS) representation in Sec. 3.2 and the Diffusion-Based Object Refinement method in Sec. 3.3.

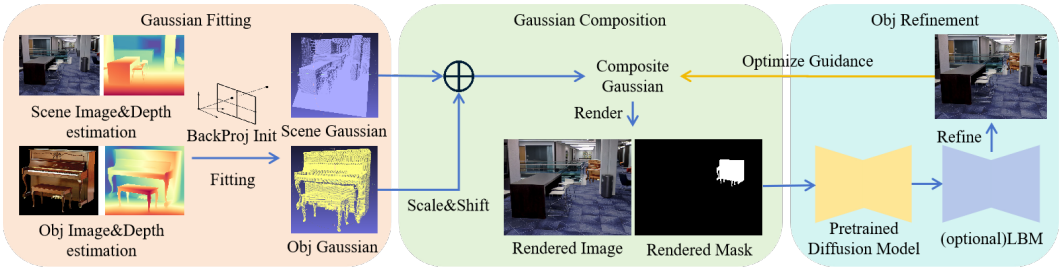

Figure 1: We first initialize and fit both scene and object images using our proposed depth-aware image Gaussian representation. The object is then scaled and positioned at the desired location through scaling and translation operations to generate the composed Gaussians. Subsequently, we employ a refinement method based on pre-trained diffusion models to optimize the object Gaussians, thereby achieving harmonization between the object and the scene.

3.1 3D GAUSSIAN SPLATTING (3DGS)

3DGS model is an explicit representation method that models scenes through a set of 3D Gaussians Kerbl et al. (2023). The spatial distribution of each 3D Gaussian $\mathcal{G}$ can be determined by its mean $\boldsymbol{\mu} \in \mathbb{R}^3$ and covariance matrix $\Sigma \in \mathbb{R}^{3\times3}$:

$$\mathcal{G}(\mathbf{x}) = e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}. \tag{1}$$

Additionally, to support subsequent image rendering, the parameters of each Gaussian also include spherical harmonics (SH) $\mathbf{c} \in \mathbb{R}^k$ representing color and opacity $\alpha \in \mathbb{R}$. Since the covariance matrix defines an ellipsoid in space, it can be further decomposed and represented as $\Sigma = RSS^TR^T$, where $R \in \mathbb{R}^{3\times3}$ is the rotation matrix and $S \in \mathbb{R}^{3\times3}$ is the scale matrix. Since the scale matrix S is a diagonal matrix, it can be represented as $S = \text{diag}([s_x, s_y, s_z])$. The rotation matrix R, processed orthogonality, can be constructed from a vector $\mathbf{v} = [r_w, r_x, r_y, r_z]$. To render images, given camera pose W, the 3D Gaussians need to be approximately projected onto the pixel coordinates along the depth dimension. The covariance matrix in the pixel coordinates can then be defined as:

$$\Sigma_{pix} = JW\Sigma W^TJ^T, \tag{2}$$

where J is the Jacobian matrix of the affine approximation of the projection transformation. The color at each pixel can then be obtained through alpha-blending of N overlapping Gaussians at that pixel in depth order:

$$\mathbf{c}_{pix} = \sum_i^N \mathbf{c}_i\alpha_i\prod_j^{i-1}(1-\alpha_j), \tag{3}$$

where $\mathbf{c}_i$ and α_i represent the color and density of each Gaussian at that pixel, respectively, which can be obtained through weighting the corresponding Gaussian's spherical harmonics (SH) and opacity by the covariance matrix Σ. Since the entire rendering process is differentiable, all parameters of the Gaussians can be optimized in an end-to-end manner based on the loss computed from the rendered images.

3.2 DEPTH-AWARE IMAGE GAUSSIAN (DA-IMGGS)

We modify the parameter configuration, initialization, and optimization of standard 3D Gaussian to develop depth-aware image Gaussian. This adaptation accommodates our application requirements, enhances fitting performance on single images, while endowing the scene representation with usable spatial structure.

Standard 3DGS defines Gaussian means in world coordinates as 3D positions. During rendering, these world coordinates are transformed to camera coordinates through the camera's extrinsic and intrinsic matrices before being projected onto the image plane, enabling image rendering from arbitrary viewpoints. However, for applications requiring rendering only from a fixed camera pose, we define our depth-aware image Gaussian directly in Normalized Device Coordinates (NDC). NDC represents a standardized coordinate system in computer graphics where coordinates are normalized to the range $[-1, 1]$. This coordinate system enables graphics rendering to adapt seamlessly across display devices with different resolutions and aspect ratios.

The covariance matrix in standard 3DGS represents an ellipsoid in 3D space, which projects to ellipses on different image planes during rendering. Since image Gaussians only require rendering onto a fixed image plane, we reduce the covariance matrix to 2D, where $\Sigma_{2d} \in \mathbb{R}^{2\times2}$ represents an elliptical Gaussian distribution parallel to the fixed image plane. In standard 3DGS, object colors observed from different viewpoints depend on multiple factors including viewing angle, lighting conditions, and material properties. Higher-order spherical harmonics decompose this lighting and material information into coefficient sets, providing efficient and smooth representations of complex lighting effects such as shadows, specular highlights, and viewpoint-dependent color variations. However, since image Gaussian rendering operates from a fixed viewpoint without 3D scene lighting variations, simple RGB values sufficiently capture the required surface colors. This reduction in Gaussian representation parameters improves scene fitting quality. Finally, the parameters of our depth-aware image Gaussian include: mean $\mathbf{u} \in \mathbb{R}^3$, where the 2D covariance matrix is decomposed via Cholesky decomposition as $\Sigma_{2d} = L \cdot L^T$, requiring storage of only three elements from the lower triangular matrix L, denoted as $L = (l_{11}, l_{21}, l_{22})^T \in \mathbb{R}^3$, opacity $\alpha \in \mathbb{R}$, and RGB color representation $\mathbf{c} \in \mathbb{R}^3$.

When supervised with only a single image and its corresponding monocular depth estimation, random initialization of Gaussians followed by optimization produces floating Gaussian components in mid-air. These components disrupt the correct spatial structure of the scene and lead to poor fitting performance. Therefore, we adopt a back-projection initialization method based on the input image and depth estimation. The mean of each initial Gaussian component in NDC is determined using the normalized pixel coordinates and depth estimation of the corresponding pixel. Color values are initialized directly from the pixel RGB values. The initial axis lengths of the ellipse corresponding to the covariance matrix are computed based on the scene image resolution, focal length, and pixel depth values by analyzing the distribution range of initial Gaussian components in NDC space. This back-projection initialization approach enables rapid fitting of image Gaussians while preventing the emergence of floating Gaussian components. The supervision during optimization combines multiple loss terms: L1 and SSIM losses between the rendered and source scene images, regularization terms for occupancy and covariance, and L1 loss with respect to the depth estimation values:

$$\mathcal{L}_{gs} = (1 - \lambda_{ssim})\mathcal{L}_1 + \lambda_{ssim}\mathcal{L}_{ssim} + \mathcal{L}_{1\_depth} + \lambda_{reg\_\alpha}\mathcal{L}_{reg\_\alpha} + \lambda_{reg\_\Sigma}\mathcal{L}_{reg\_\Sigma}, \tag{4}$$

we set λ_{ssim} to 0.2, λ_{reg_α} and λ_{reg_Σ} to 0.01. After fitting Gaussians to the scene and object images, the object can be placed at any spatial location within the scene Gaussians at arbitrary scale by scaling and translating the object Gaussians.

3.3 DIFFUSION-BASED OBJECT REFINEMENT

After achieving depth-aware object insertion through Gaussian combination of scene and object, the inserted object requires further refinement to harmonize its style and surface illumination with the surrounding environment. We propose a simple yet effective object refinement approach that leverages a pretrained diffusion model to harmonize the rendered image. Diffusion models pretrained on large-scale datasets inherently capture diverse output distributions encompassing various visual styles and lighting conditions. For cross-domain object insertion tasks (real object insert to stylized scene), we directly exploit the knowledge embedded in the pretrained FLUX.1-Fill-dev model to refine the inserted object, thereby achieving style harmonization between the object and scene.

We denote the combined Gaussians of scene and object as $\mathcal{G}com$, where R represents the differentiable rendering process. Based on the combined Gaussians, we can simultaneously render the composite image and the opacity of object Gaussians to obtain the object mask: $\mathbf{I}_{com}, \mathbf{M}_{obj} = R(\mathcal{G}_{com})$. We introduce an coefficient $s \in [0, 1]$ to control the refinement strength. When $s = 0$, the method outputs the original image without modification, while $s = 1$ performs complete inpainting of the masked region. Let T denote the total number of inference steps in the diffusion model, corresponding to T different noise levels. Given refinement strength s, the corresponding noise level is computed as $T' = \lceil sT \rceil$. Let $\mathcal{E}$ denote the encoder of the diffusion model. The latent variable corresponding to the composite image is $\mathbf{z}_{com} = \mathcal{E}(\mathbf{I}_{com})$. The initial latent variable for the inference process is then defined as:

$$\mathbf{z}_{T'} = \sigma(T^{'})\boldsymbol{\epsilon} + (1 - \sigma(T^{'}))\mathbf{z}_{com}, \boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I}). \tag{5}$$

DiT models such as FLUX process inputs by dividing them into patches and converting the values within each patch into tokens. Let the size of each patch be $k = h \times w$, with a total of N patches. We apply similar processing to $\mathbf{M}_{obj}$, transforming it into a patch-wise binary mask $\mathbf{M}_{patch} \in \{0, 1\}^N$:

$$m_n = \left[\sum_{i=1}^{k} \mathbf{M}_{obj\_n,i} > 0 \right] \in \{0, 1\}. \tag{6}$$

We denote the diffusion model used for refinement as G_{refine}, which follows the flow matching framework. At each time step t during the refinement process, we first perform a single denoising step on the noisy latent variables.

$$\mathbf{z}_{t-1} = \mathbf{z}_t - \Delta t \cdot G_{refine}(\mathbf{z}_t, t|c), \tag{7}$$

where c represents the textual condition for the model. In our implementation, we employ a simple template format `"XXX style of a XXX"` that provides basic descriptions of object categories and styles. Subsequently, we perform an overwrite operation on the latent variables based on the patch-wise mask.

$$\mathbf{z}_{t-1} = \mathbf{M}_{patch} \odot \mathbf{z}_{t-1} + (1 - \mathbf{M}_{patch}) \odot \mathbf{z}_{ref}(t - 1), \tag{8}$$

with

$$\mathbf{z}_{\text{ref}}(t) = \begin{cases} \sigma(t)\boldsymbol{\epsilon} + (1 - \sigma(t))\mathbf{z}_{com} & t > 0, \\ \mathbf{z}_{com} & t = 0. \end{cases} \tag{9}$$

This ensures precise reconstruction of the original scene image in regions outside the object mask.

In real-scene object insertion, object refinement focuses primarily on achieving consistency between object surface lighting and the surrounding environment. Since widely-used open-source text-to-image diffusion models and inpainting models such as Stable Diffusion or FLUX possess limited knowledge in lighting harmonization, we additionally leverage the pre-trained illumination harmonization model LBM (Chadebec et al., 2025), denoted as G_{light}. For object refinement in real-scene, we first apply low-strength refinement to harmonize the inserted object, then perform additional lighting harmonization on the refined image using G_{light}. The refinement process for real-scene domain object insertion can be expressed as:

$$\mathbf{I}_{harmony} = G_{light}(\text{Refinement}(\mathbf{I}_{com}, s)). \tag{10}$$

After completing the refinement process on the rendered image following object insertion, the harmonized image from the diffusion model can directly provide supervision for Gaussian parameter updates due to the differentiable nature of the rendering process. However, during the refinement process, diffusion models often introduce unwanted modifications or detail loss in non-target regions due to constraints in their generative capabilities and supported resolutions. Therefore, we set only the object Gaussians as learnable to avoid affecting the scene Gaussians.

4 Experiments

4.1 Experiment Setup

Data Preparation Due to the current lack of high-quality object insertion datasets that feature complex spatial relationships between inserted objects and scenes while encompassing both same-domain and cross-domain scenarios, we construct a high-resolution dataset containing diverse real

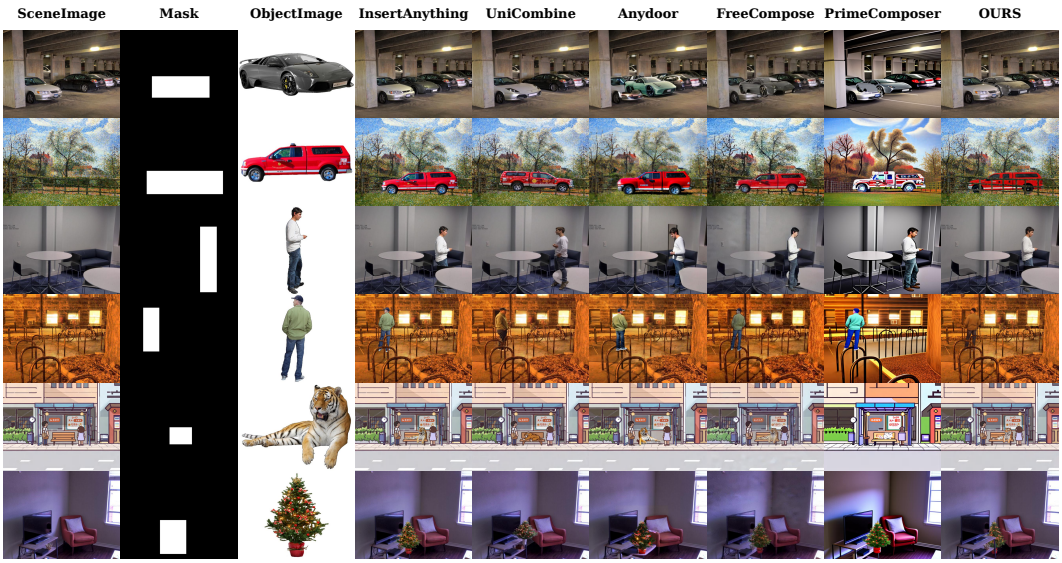

Figure 2: Visual comparison results between our SpatialComposer and baselines. Zoom in to observe details.

photographs and artistic style scenes with complex spatial structures to validate our proposed method. For scene images, we select high-quality indoor and outdoor photographs from the DIODE dataset (Vasiljevic et al., 2019) that meet our spatial complexity requirements. We also incorporate high-resolution artworks featuring various oil painting styles from the WikiArt dataset (Saleh & Elgammal, 2024), extract qualifying frames from anime and cartoon videos, and generate additional scene images in other styles using GPT-Image-1 (OpenAI, 2024) and FLUX-pro-1.1-ultra. This process yields 225 high-quality scene images across multiple domains. For object images, we collect web resources and utilize SAM (Kirillov et al., 2023) to segment objects from the COCO (Lin et al., 2014) dataset, obtaining 255 object images spanning more than 9 categories including animals, vegetables, food, people, and vehicles. Based on these scenes and objects, we provide over 200 semantically coherent scene-object combination cases for evaluation. All subsequent experiments are conducted on this carefully curated dataset. As a supplement, we also report results on the existing TF-ICON benchmark in the Appendix F.

Experiment Details For Gaussian fitting, we employ the Adam optimizer with learning rates of 0.01 for means, $2.5e-3$ for opacities, 0.01 for RGB values, and 0.1 for covariances. We fix the global random seed to 42. Since our method involves only forward inference of the refinement model and Gaussian fitting, a single A6000 GPU is sufficient to support object refinement using FLUX.1-Fill-dev.

Baselines To validate our method's effectiveness, we compare against both training-based and training-free object insertion approaches. The training-based baselines include InsertAnything, UniCombine, and AnyDoor, while the training-free baselines comprise Freecompose, Primecomposer.

4.2 Quantitative Evaluation

Existing automatic evaluation metrics primarily measure the preservation of non-edited regions by calculating the distance between low-level features of non-edited regions in the generated results and corresponding regions in the scene image. To evaluate the preservation of non-edited regions in the scene, we compute the PSNR between these regions and their counterparts in the original scene image. For object identity preservation, previous works (Lu et al., 2023b; Wang et al., 2024; Chen et al., 2024c) typically measure the distance between high-level semantic features or low-level perceptual features extracted by pre-trained visual encoders from the edited region in the result image and the reference object image. Our method considers depth-related spatial positioning during object insertion. After the model reasonably handles occlusion relationships of inserted objects,

Table 1: Quantitative comparison of different methods. **Bold** indicates the best result, underline denotes the second best, and † marks training-free methods.

Method	Auto-Metric			User-Study			
	$\textbf{PSNR}_{bg}\uparrow$	$\textbf{Harm}_{Style}\downarrow$	$\textbf{Harm}_{Real}\uparrow$	Qual	Harm	Reas	Cons
AnyDoor	31.52	3.21	0.78	2.94	2.11	2.30	2.21
UniCombine	32.03	3.16	**0.87**	4.96	6.71	4.32	5.61
InsertAnything	36.82	3.18	0.79	8.92	7.17	7.08	9.38
Freecompose†	23.56	3.27	0.81	1.75	1.84	1.38	1.75
Primecomposer†	16.82	3.12	0.81	0.55	0.74	0.74	0.46
SpatialComposer†	**46.99**	**3.08**	0.84	**80.88**	**81.43**	**84.19**	**80.61**

the feature distance between edited region and the reference object image actually increases. Concurrently, when a substantial domain gap exists between the target scene and the reference object, modifications to the object's color and texture are required to ensure seamless integration and visual harmony. However, these refinement adjustments inherently result in degraded performance on the corresponding quantitative metrics. Consequently, we exclude these quantitative measures from our experimental evaluation. Furthermore, previous work lacks effective metrics for measuring whether inserted objects harmonize with the surrounding environment. Therefore, we design automatic harmony assessment metrics specifically for object insertion in stylized scenes and real scenes, respectively. For object insertion in stylized scenes, we compute the style loss (Karras et al., 2019) based on the Gram Matrix between the edited region in the result image after object insertion and the original scene image to measure style harmony. For object insertion in real scenes, harmony focuses more on the consistency between object surface lighting and the scene, which we measure by computing the consistency of brightness and color distributions between the object and surrounding scene. We provide detailed implementation of the harmony evaluation metrics for both stylized and real-world scenes in Appendix B. For the reasonableness of inserted objects' spatial positioning and scale within scenes, this proves difficult to assess through automatic evaluation metrics.

Given the current lack of automatic evaluation metrics with high alignment to human perception for such interactive generation and editing tasks, this field primarily relies on user studies for result assessment. In our user study, we recruited 40 participants to evaluate results generated by SpatialComposer and baseline approaches across four dimensions: overall image quality, harmony between object and environment, the reasonableness of inserted objects' spatial positioning and scale, and consistency between the object and reference object image, as well as consistency between non-edited regions in the scene and the original scene image. For each case, participants identified the best-performing method across these four evaluation dimensions. We subsequently computed the average vote percentage for each method across all four dimensions. The results presented in Table 1 demonstrate that SpatialComposer achieves comparable or superior performance to existing approaches in automatic evaluation metrics. Our dataset comprises scenes primarily derived from real photographs and authentic artworks, which high resolution, complex spatial structures at object insertion locations, and diverse styles. These characteristics pose significant challenges for existing object insertion methods. SpatialComposer overcomes these difficulties through the introduction of depth-aware Gaussian representations and the effective utilization of diffusion models pre-trained on large-scale data, thereby achieving significantly superior performance in the user study.

4.3 Qualitative Evaluation

As shown in Fig. 2, InsertAnything demonstrates the best performance among the baselines, showing some understanding of object depth relationships and displaying reasonable depth relations in certain insertion results. However, its capabilities for style harmonization in cross-domain insertion and illumination processing in real scenes remain limited, as shown in the second, fourth and fifth rows in Fig. 2. UniCombine exhibits stronger style harmonization capabilities for cross-domain object insertion but frequently generates results with significant errors or unreasonable outcomes, as demonstrated by the spatial positions of objects in the third and sixth rows of Fig. 2. AnyDoor shows limitations in object-scene harmony and the visual quality in edited regions. FreeCompose

lacks understanding of depth information and demonstrates limited harmonization performance. Beyond the issues encountered by other methods, PrimeComposer also exhibits significant problems in preserving the original scene image. Overall, SpatialComposer achieves comparable or superior performance in terms of image quality, reasonableness of spatial positioning and scale, consistency of object identity with non-edited regions of the scene, and harmonization between the scene and objects. More results are provided in Appendix D. Due to file size constraints, we provide vector-format versions of the visualization comparison figures in the supplementary material.

4.4 Ablation Study

We conduct ablation studies to validate the effectiveness of our proposed Gaussian representation, initialization method, and refinement foundation model, as well as to examine the impact of strength coefficient in our refinement method.

Gaussian representation and initialization method We simplify standard 3D Gaussians to better adapt them to our task and employ a back-projection-based initialization method. This approach not only significantly enhances scene fitting quality but also achieves superior spatial structure representation. As shown in Fig. 3, our proposed Gaussian representation and initialization method substantially improve the fitting quality of scene reconstruction. In Fig. 6, we visualize the Gaussian means under different configurations, demonstrating that our proposed Gaussian settings and initialization method yield the most reasonable spatial structure, which serves as the foundation for subsequent object insertion operations.

Refinement Foundation Model In the refinement processing of inserted objects, our pipeline is compatible with different foundation models. Fig. 4 (a) compares the refinement results across different foundation models. Our proposed refinement method achieves the best overall performance in maintaining object identity while achieving style consistency between objects and scenes.

Illumination Harmonization Model in Real Scenes In Fig. 4 (b), we validate the limitations of pretrained text-to-image and inpainting models when addressing object surface illumination harmonization in real scenes, as well as the necessity of incorporating pretrained illumination harmonization models. The illumination harmonization model better captures environmental lighting information, including intensity, color, and direction.

Refinement Strength Finally, we validate the impact of the strength coefficient s on the results during the refinement process. As shown in Fig. 5, higher refinement strength produces results that are more harmonious with the scene style but simultaneously weakens the preservation of object identity. An appropriate value should be determined based on the degree of domain gap between the inserted object and the scene. In our experiments, we select $s = 0.6$ as the default setting for stylized scene object insertion and $s = 0.2$ for real scene object insertion.

5 Conclusion and Limitation

In this paper, we introduce SpatialComposer, which effectively reconstructs high-fidelity scene representations with meaningful depth structure, enabling precise control over object scale and 3D spatial positioning during insertion. To address insertion disharmony, we employ a refinement method based on pretrained diffusion models, achieving reasonable harmonization that aligns object and lighting with the surrounding scene. Furthermore, we constructed the DAOI dataset, in which we collected over 200 high-resolution scene images with complex spatial structures and diverse styles, along with over 200 multi-category object images. Experiments demonstrate our method's superiority, achieving comparable or superior performance across multiple evaluation dimensions.

In our task setting, given only a single image of the insertion object, we cannot construct a complete Gaussian representation of the object and can only model the visible surface shown in the single image. Consequently, in the subsequent Gaussian composition operations, our method only supports scaling and translation of the object Gaussians, but does not support rotation of the object Gaussians. Reconstructing individual object Gaussians from single object images represents an actively pursued research direction in the 3D Gaussian field. With the rapid advancement of techniques in this domain, incorporating object 3D Gaussian reconstruction models based on a single object image may provide an effective approach to address the limitations of our method.

Ethics Statement

This work adheres to the ICLR Code of Ethics. Our research focuses on image editing tasks, where all data are sourced from open-source datasets, generated through model API calls, and publicly available internet data, with no involvement of personal privacy information. All data used in this study are publicly available and were used in accordance with their respective licenses and terms of use. We have properly cited all data sources and respected the original creators' rights. While our method demonstrates improvements in object insertion, we recognize that like other AI technologies, it could potentially be misused to generate unsafe visual content. We encourage responsible deployment and further research into safety mechanisms. We believe this work contributes positively to the research community and poses minimal ethical concerns when used responsibly.

Reproducibility Statement

We have made every effort to ensure the reproducibility of our work. Our constructed dataset DAOI will be publicly available alongside our framework and experimental source code upon publication, while TF-ICON serves as a public benchmark. Our proposed method is detailed in Sec. 3, which includes the hyperparameters involved in the loss terms. Additional experimental details such as optimizers, learning rates, and random seeds are provided in Sec. 4. We commit to making the complete framework and experimental source code, as well as the constructed benchmark, publicly available as soon as possible after publication.

References

Gemma Canet Tarrés, Zhe Lin, Zhifei Zhang, Jianming Zhang, Yizhi Song, Dan Ruta, Andrew Gilbert, John Collomosse, and Soo Ye Kim. Thinking outside the bbox: Unconstrained generative object compositing. In *European Conference on Computer Vision*, pp. 476–495. Springer, 2024.

Clément Chadebec, Onur Tasar, Sanjeev Sreetharan, and Benjamin Aubin. Lbm: Latent bridge matching for fast image-to-image translation. *arXiv preprint arXiv:2503.07535*, 2025.

Jiaxuan Chen, Bo Zhang, Qingdong He, Jinlong Peng, and Li Niu. Mureobjectstitch: Multi-reference image composition. *arXiv preprint arXiv:2411.07462*, 2024a.

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

Zhekai Chen, Wen Wang, Zhen Yang, Zeqing Yuan, Hao Chen, and Chunhua Shen. Freecompose: Generic zero-shot image composition with diffusion prior. In *European Conference on Computer Vision*, pp. 70–87. Springer, 2024c.

Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. Dovenet: Deep image harmonization via domain verification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8394–8403, 2020.

Wenyan Cong, Xinhao Tao, Li Niu, Jing Liang, Xuesong Gao, Qihao Sun, and Liqing Zhang. High-resolution image harmonization via collaborative dual transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18470–18479, 2022.

Jixuan He, Wanhua Li, Ye Liu, Junsik Kim, Donglai Wei, and Hanspeter Pfister. Affordance-aware object insertion via mask-aware dual diffusion. *arXiv preprint arXiv:2412.14462*, 2024.

Amir Hertz, Kfir Aberman, and Daniel Cohen-Or. Delta denoising score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2328–2337, 2023.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pp. 2790–2799. PMLR, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (TOG)*, 42(4):1–14, 2023.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Sumith Kulal, Tim Brooks, Alex Aiken, Jiajun Wu, Jimei Yang, Jingwan Lu, Alexei A Efros, and Krishna Kumar Singh. Putting people in their place: Affordance-aware human insertion into scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17089–17099, 2023.

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

Pengzhi Li, Qiang Nie, Ying Chen, Xi Jiang, Kai Wu, Yuhuan Lin, Yong Liu, Jinlong Peng, Chengjie Wang, and Feng Zheng. Tuning-free image customization with image and text guidance. In *European Conference on Computer Vision*, pp. 233–250. Springer, 2024.

Tianle Li, Max Ku, Cong Wei, and Wenhu Chen. Dreamedit: Subject-driven image editing. *arXiv preprint arXiv:2306.12624*, 2023.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Jun Ling, Han Xue, Li Song, Rong Xie, and Xiao Gu. Region-aware adaptive instance normalization for image harmonization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9361–9370, 2021.

Liu Liu, Zhenchen Liu, Bo Zhang, Jiangtong Li, Li Niu, Qingyang Liu, and Liqing Zhang. Opa: object placement assessment dataset. *arXiv preprint arXiv:2107.01889*, 2021.

Lingxiao Lu, Jiangtong Li, Bo Zhang, and Li Niu. Dreamcom: Finetuning text-guided inpainting model for image composition. *arXiv preprint arXiv:2309.15508*, 2023a.

Shilin Lu, Yanzhu Liu, and Adams Wai-Kin Kong. Tf-icon: Diffusion-based training-free cross-domain image composition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2294–2305, 2023b.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6038–6047, 2023.

Li Niu, Qingyang Liu, Zhenchen Liu, and Jiangtong Li. Fast object placement assessment. *arXiv preprint arXiv:2205.14280*, 2022.

OpenAI. gpt-image-1. `https://platform.openai.com/docs/models#gpt-image`, 2024. Accessed: 2025-09-24.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022a.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *arXiv preprint arXiv:2112.10752*, 2022b.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22500–22510, 2023.

Babak Saleh and Ahmed Elgammal. Wikiart: A large-scale dataset of artworks, dec 2024. URL https://service.tib.eu/ldmservice/dataset/wikiart--a-large-scale-dataset-of-artworks.

Konstantin Sofiiuk, Polina Popenova, and Anton Konushin. Foreground-aware semantic representations for image harmonization. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 1620–1629, 2021.

Wensong Song, Hong Jiang, Zongxing Yang, Ruijie Quan, and Yi Yang. Insert anything: Image insertion via in-context editing in dit. *arXiv preprint arXiv:2504.15009*, 2025.

Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, and Daniel Aliaga. Objectstitch: Generative object compositing. *arXiv preprint arXiv:2212.00932*, 2022.

Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, He Zhang, Wei Xiong, and Daniel Aliaga. Imprint: Generative object compositing by learning identity-preserving representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8048–8058, 2024.

Shashank Tripathi, Siddhartha Chandra, Amit Agrawal, Ambrish Tyagi, James M Rehg, and Visesh Chari. Learning to generate synthetic data via compositing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 461–470, 2019.

Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z. Dai, Andrea F. Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R. Walter, and Gregory Shakhnarovich. DIODE: A Dense Indoor and Outdoor DEpth Dataset. *CoRR*, abs/1908.00463, 2019. URL http://arxiv.org/abs/1908.00463.

Haoxuan Wang, Jinlong Peng, Qingdong He, Hao Yang, Ying Jin, Jiafu Wu, Xiaobin Hu, Yanjie Pan, Zhenye Gan, Mingmin Chi, et al. Unicombine: Unified multi-conditional combination with diffusion transformer. *arXiv preprint arXiv:2503.09277*, 2025.

Yibin Wang, Weizhong Zhang, Jianwei Zheng, and Cheng Jin. Primecomposer: Faster progressively combined diffusion for image composition with attention steering. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 10824–10832, 2024.

Daniel Winter, Matan Cohen, Shlomi Fruchter, Yael Pritch, Alex Rav-Acha, and Yedid Hoshen. Objectdrop: Bootstrapping counterfactuals for photorealistic object removal and insertion. In *European Conference on Computer Vision*, pp. 112–129. Springer, 2024a.

Daniel Winter, Asaf Shul, Matan Cohen, Dana Berman, Yael Pritch, Alex Rav-Acha, and Yedid Hoshen. Objectmate: A recurrence prior for object insertion and subject-driven generation. *arXiv preprint arXiv:2412.08645*, 2024b.

Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18381–18391, 2023.

Yongsheng Yu, Ziyun Zeng, Haitian Zheng, and Jiebo Luo. Omnipaint: Mastering object-oriented editing via disentangled insertion-removal inpainting. *arXiv preprint arXiv:2503.08677*, 2025.

Ziyang Yuan, Mingdeng Cao, Xintao Wang, Zhongang Qi, Chun Yuan, and Ying Shan. Customnet: Object customization with variable-viewpoints in text-to-image diffusion models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 10976–10984, 2024.

Bo Zhang, Yuxuan Duan, Jun Lan, Yan Hong, Huijia Zhu, Weiqiang Wang, and Li Niu. Controlcom: Controllable image composition using diffusion model. *arXiv preprint arXiv:2308.10040*, 2023a.

Lingzhi Zhang, Tarmily Wen, Jie Min, Jiancong Wang, David Han, and Jianbo Shi. Learning object placement by inpainting for compositional data augmentation. In *European Conference on Computer Vision*, pp. 566–581. Springer, 2020.

Xin Zhang, Jiaxian Guo, Paul Yoo, Yutaka Matsuo, and Yusuke Iwasawa. Paste, inpaint and harmonize via denoising: Subject-driven image editing with pre-trained diffusion model. *arXiv preprint arXiv:2306.07596*, 2023b.

Zitian Zhang, Frédéric Fortier-Chouinard, Mathieu Garon, Anand Bhattad, and Jean-François Lalonde. Zerocomp: Zero-shot object compositing from image intrinsics via diffusion. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 483–494. IEEE, 2025.

A Latent Diffusion Model

In this work, we leverage pre-trained text-conditioned latent diffusion models (Rombach et al., 2022a) to guide the object refinement process. These models operate in a learned latent space through an encoder-decoder architecture, where $\mathcal{E}(\cdot)$ and $\mathcal{D}(\cdot)$ represent the encoder and decoder components, respectively. The diffusion process involves forward noise addition followed by reverse denoising operations within this latent representation. Given an input image $\mathbf{x}_0$, we first encode it into the latent space as $\mathbf{z}_0 = \mathcal{E}(\mathbf{x}_0)$. During training, this latent representation is progressively corrupted through the forward diffusion process, transforming $\mathbf{z}_0$ into $\mathbf{z}_t$:

$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}), \tag{11}$$

for $t \in [1, T]$, where $\bar{\alpha}_t = \prod_{s=1}^{t} 1 - \beta_s$, and β_s represents the variance schedule at timestep s. Subsequently, a denoising U-Net is trained to predict the added noise conditioned on c using the following objective function:

$$\mathcal{L} = \mathbb{E}_{\mathcal{E}(\mathbf{x}_0),\boldsymbol{\epsilon}\sim\mathcal{N}(0,\mathbf{I}),t}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, c)\|_2^2\right], \tag{12}$$

where $\boldsymbol{\epsilon}_\theta$ represents the denoising U-Net.

B Implementation of Harmony Evaluation Metrics

The automatic evaluation metrics employed by existing works do not specifically measure the harmony between inserted objects and scenes. Based on the different aspects of harmony that humans focus on in real scenes versus stylized scenes, we design and adopt different harmony measurement approaches accordingly.

B.1 Stylized Scenes Harmony Evaluation Metric

For object insertion in stylized scenes, we follow the StyleLoss setting in StyleGAN (Karras et al., 2019). We compute the Gram Matrix between the edited region in the result image after object insertion and the original scene image, and measure the harmony between the inserted object and stylized scene based on the Gram Matrix. Given a composite image I_c, scene image I_s, and binary mask M indicating the inserted object region, we extract deep convolutional features using a pre-trained VGG-19 network. Following established practices in neural style transfer, we select features from five representative layers (conv1_1, conv2_1, conv3_1, conv4_1, conv5_1) to capture style information at different levels of visual abstraction. The object region is extracted by applying the mask to the composite image: $I_o = I_c \odot M$, where $\odot$ denotes element-wise multiplication. For each selected layer l, we compute the Gram matrix G^l to encode style information through feature channel correlations. Given feature map $F^l \in \mathbb{R}^{N_l \times M_l}$ where N_l is the number of channels and $M_l = H_l \times W_l$ represents spatial dimensions, the Gram matrix is computed as:

$$G^l_{i,j} = \frac{1}{N_l M_l} \sum_{k=1}^{M_l} F^l_{i,k} F^l_{j,k}. \tag{13}$$

This formulation captures the correlations between different feature channels, which has been shown to effectively represent texture and style characteristics.

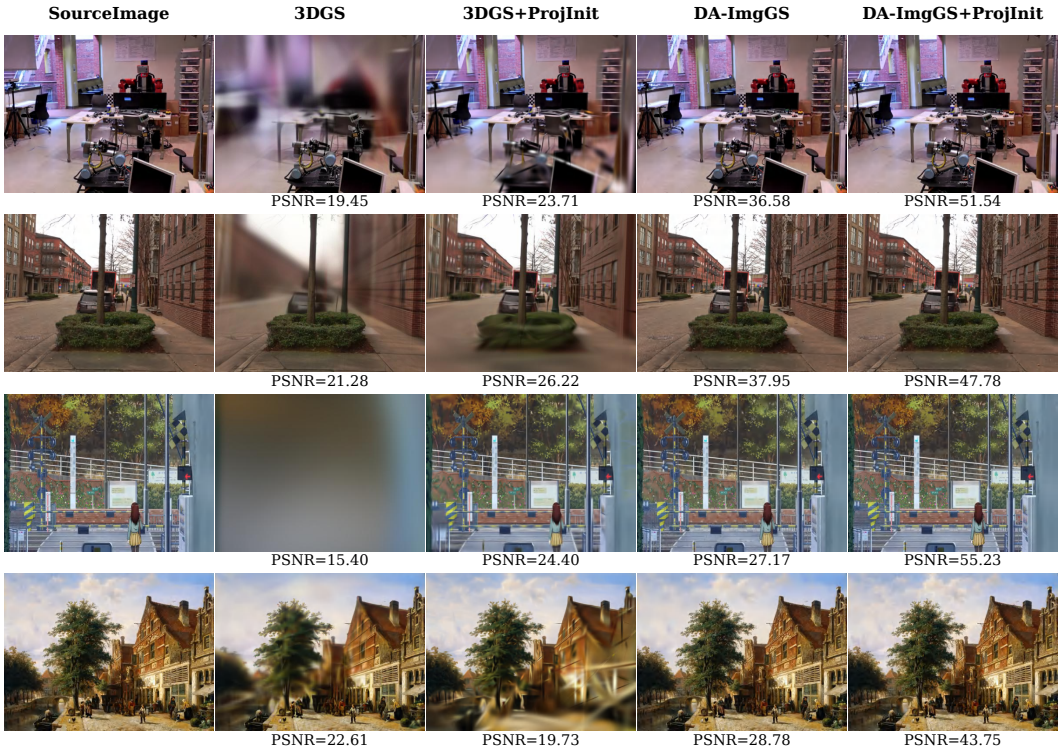

Figure 3: Comparison of fitting performance under different Gaussian representations and initialization strategies. Both the proposed depth-aware image Gaussian and initialization strategy demonstrate improved fitting quality.

The style discrepancy between the inserted object and scene is measured by comparing their respective Gram matrices across multiple layers. For each layer l, we compute the style loss using the Frobenius norm:

$$\mathcal{L}_s^l = \|G_b^l - G_o^l\|_F^2, \tag{14}$$

where G_b^l and G_o^l are the Gram matrices for scene and object regions respectively. The final style consistency score is obtained by averaging losses across all selected layers:

$$\mathcal{L}_s = \frac{1}{|\mathcal{L}|} \sum_{l \in \mathcal{L}} \mathcal{L}_s^l. \tag{15}$$

This multi-scale approach ensures comprehensive style evaluation from low-level textures to high-level semantic features. Lower values indicate better style harmony between the inserted object and scene, while higher values suggest greater stylistic discrepancy. The metric can be computed efficiently for batch evaluation and provides a quantitative foundation for comparing different object insertion methods.

B.2 Real Scenes Harmony Evaluation Metric

For object insertion in real scenes, we primarily consider the consistency between object surface lighting and environmental lighting, measuring the harmony between objects and real scenes by computing the consistency of brightness distribution and color distribution between the object and surrounding scene.

Given an inserted object with mask M_o and its surrounding scene region M_b, we extract the corresponding pixel sets P_o and P_b from the composite image I. The overall consistency score is defined as: $S_{overall} = w_b \cdot S_{brightness} + w_c \cdot S_{color}$, where $w_b + w_c = 1$. The brightness consistency $S_{brightness}$ evaluates luminance distribution alignment through three measures: (1) Jensen-Shannon divergence: $S_{JS} = 1 - JS(H_o||H_b)$ where H_o and H_b are normalized grayscale histograms; (2)

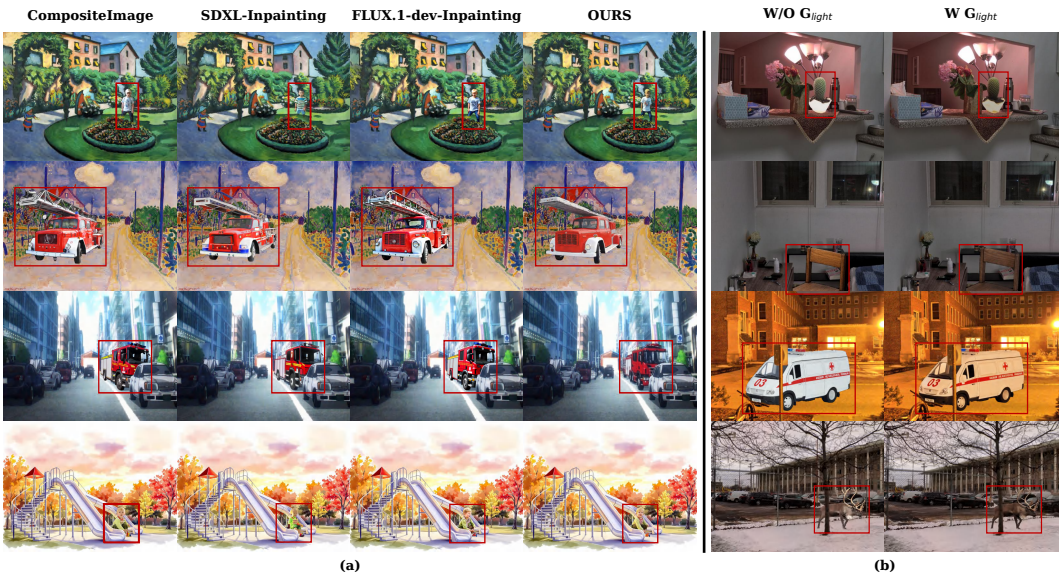

Figure 4: (a) demonstrates the compatibility of our pipeline with different pre-trained diffusion models and the effectiveness of our refinement method; (b) shows that incorporating pre-trained illumination harmonization models in real-scene object insertion significantly enhances the harmony between object surface lighting and the environment.

Figure 5: The impact of different object refinement intensities on the results. Higher refinement strength enables greater stylistic changes to the object but makes it more challenging to preserve object identity.

Wasserstein distance: $S_{EMD} = \exp(-W(P_o^{gray}, P_b^{gray})/\tau)$ for distribution matching; and (3) Statistical similarity: $S_{stat} = \exp(-(|\mu_o - \mu_b|/255 + |\sigma_o - \sigma_b|/255))$ for first and second-order statistics. These are combined as: $S_{brightness} = 0.4 \cdot S_{JS} + 0.4 \cdot S_{EMD} + 0.2 \cdot S_{stat}$.

The color consistency S_{color} is evaluated in the perceptually uniform CIE-LAB space to better reflect human visual perception. For each channel $c \in \{L^*, a^*, b^*\}$, we compute channel-specific scores S_c using the same JS divergence and Wasserstein distance formulations applied to the channel distributions. The final color score employs weighted integration: $S_{\text{color}} = 0.5 \cdot S_L + 0.25 \cdot S_a +$

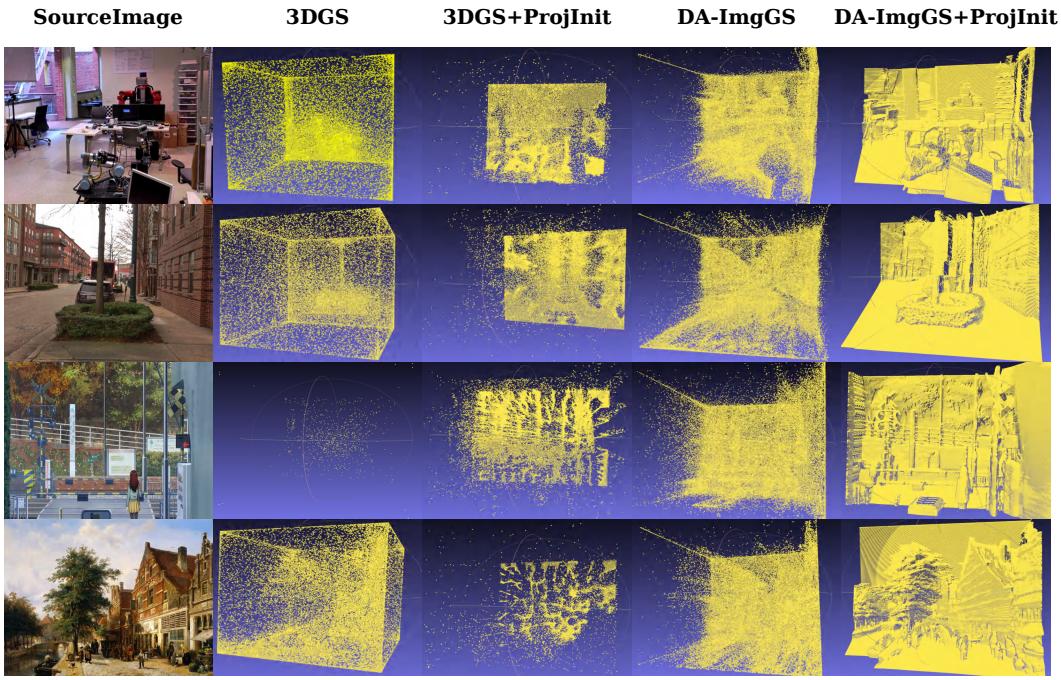

Figure 6: The adoption of depth-aware Gaussian splatting configuration combined with back-projection initialization strategy yields scene Gaussians with significantly more exploitable spatial structure.

$0.25 \cdot S_b$, with higher weight on luminance to reflect its dominance in visual perception. In our implementation, we set $w_b = 0.4$ and $w_c = 0.6$, and automatically generate surrounding scene regions through morphological dilation of the object mask. The metric ranges from 0 to 1, where higher scores indicate better photometric integration between the inserted object and the scene.

C Spatial Structure with Different Gaussian Settings

The depth-aware image Gaussian and back-projection initialization strategy adopted in our method not only effectively improves the Gaussian fitting quality, but also significantly optimizes the spatial structure of the Gaussian representation. We visualize the means of scene Gaussians fitted under different Gaussian and initialization settings respectively, and show them in Fig. 6. The joint employment of our proposed depth-aware image Gaussian with back-projection initialization strategy effectively captures the available spatial structure.

D Additional Visual Comparison Results

In Figs. 7 and 8, we present additional visualized comparisons of generation results.

E Multi-object Insertion Examples

SpatialComposer supports inserting multiple objects in a single scene. Since the Gaussians of these objects and the scene Gaussians are distinguishable, it avoids the accumulation of influence on other Gaussians when inserting multiple objects. In Fig. 9, we show some results of multiple object insertion in scenes.

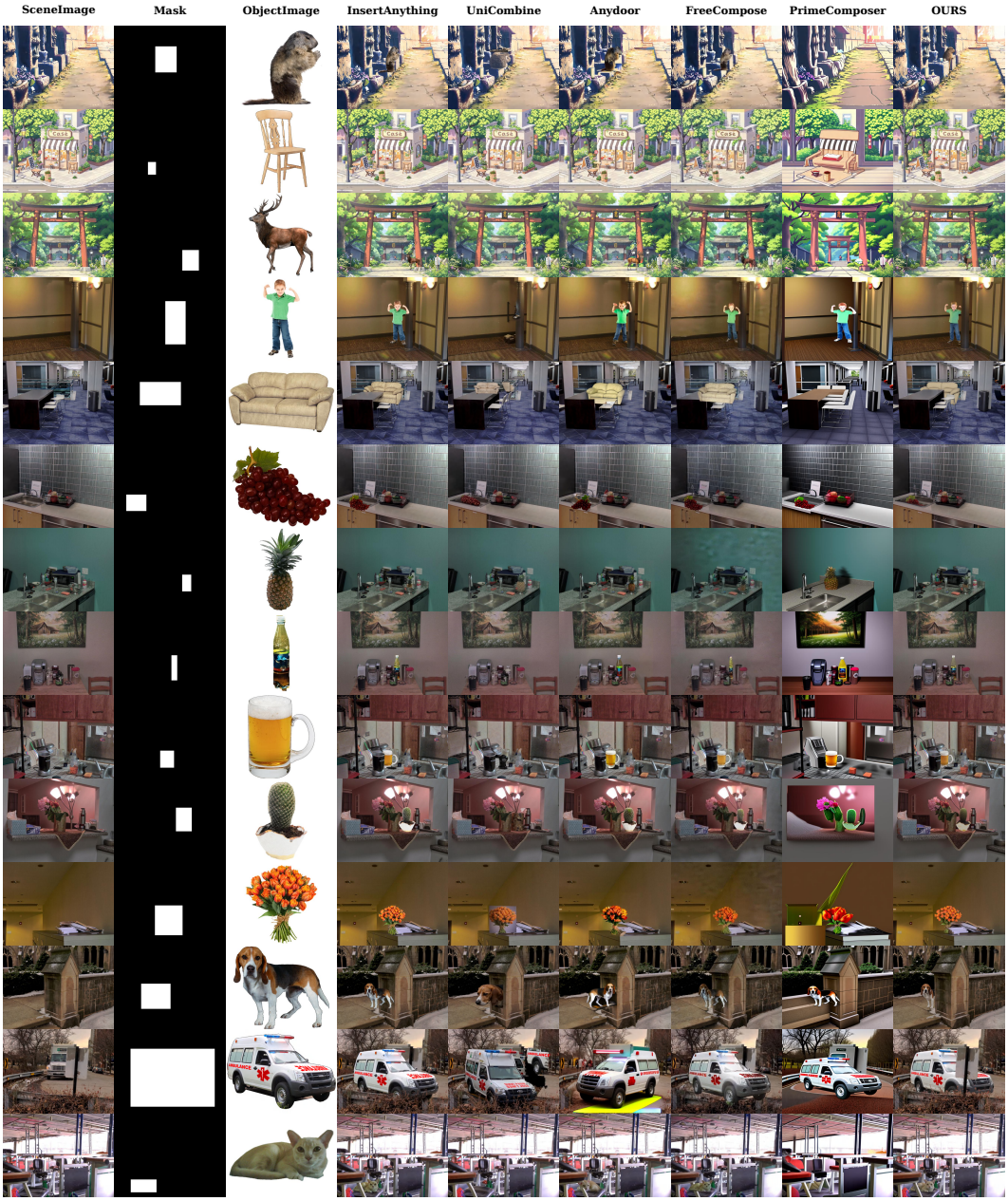

Figure 7: Visualization of comparative results, zoomed in for detailed observation.

F Results on TF-ICON Benchmark

SpatialComposer is a training-free object insertion approach that supports application to both stylized and real scenes. The existing works most similar to our method and task setting are TF-ICON, Primecomposer, and Freecompose. These works primarily use the benchmark proposed by TF-ICON, which contains 95 stylized scene object insertion cases and 237 real scene object insertion cases constructed from approximately 30 scene images and about 100 object images. The object images are constructed from real images through semantic segmentation, while the scene images are 256×256 pixel or 512×512 pixel Stable Diffusion generated images, with object insertion positions set in open areas of the images where complex spatial relationships do not exist. Due to the low scene resolution, generated nature of scenes, and simple object insertion position scenarios, this

Figure 8: Visualization of comparative results, zoomed in for detailed observation.

dataset cannot perfectly meet the requirements of our task. However, we also report results on this benchmark in this section.

We present a visual comparison of different methods on the TF-ICON benchmark in Fig. 10. Although the object insertion positions in this dataset do not involve complex spatial relationships, SpatialComposer consistently achieves performance that is comparable to or superior to existing approaches. For the quantitative metric setup on this dataset, we follow the configurations established in these prior works. We employ PSNR between the non-edited regions in the result image and the corresponding regions in the scene image to measure the preservation of non-edited areas. We use LPIPS and cosine similarity between CLIP image embeddings of the edited region in the result image and the reference object image to evaluate object identity preservation. As mentioned in our quantitative experiments in the main text, we believe that the evaluation results of these two

Table 2: Quantitative comparison of different methods on TF-ICON Dataset

Method	$\textbf{PSNR}_{bg}$	$\textbf{LPIPS}_{obj}$	$\textbf{CLIP}_{img\_img}$	$\textbf{CLIP}_{img\_text}$
AnyDoor	23.74	0.538	0.873	0.279
UniCombine	26.61	0.632	0.790	0.286
InsertAnything	<u>31.91</u>	0.432	0.881	0.285
Freecompose†	21.33	0.369	0.873	0.283
Primecomposer†	12.32	0.469	0.810	0.272
SpatialComposer†	33.14	0.468	0.831	0.307

quantitative metrics often do not align with human perceptual quality. We adopt the cosine similarity between the CLIP image embedding of the edited region in the result image and the CLIP text embedding of the corresponding text prompt for the edited region (formatted as `"XXX style / professional photograph of a XXX"`) to comprehensively measure both inserted object style and category. The results presented in Table 2 show that SpatialComposer achieves better performance in both the preservation of non-edited regions and the alignment of edited regions with the target style and object category.

G The Use of Large Language Models

During the writing process of this paper, we utilized large language models to enhance the manuscript quality, including employing large language models to correct grammatical errors and modify wording and expressions to achieve a more formal and academic style.

Figure 9: SpatialComposer supports multi-object insertion within a single scene while avoiding the adverse effects of multiple object insertion operations on other regions of the image, zoomed in for detailed observation.

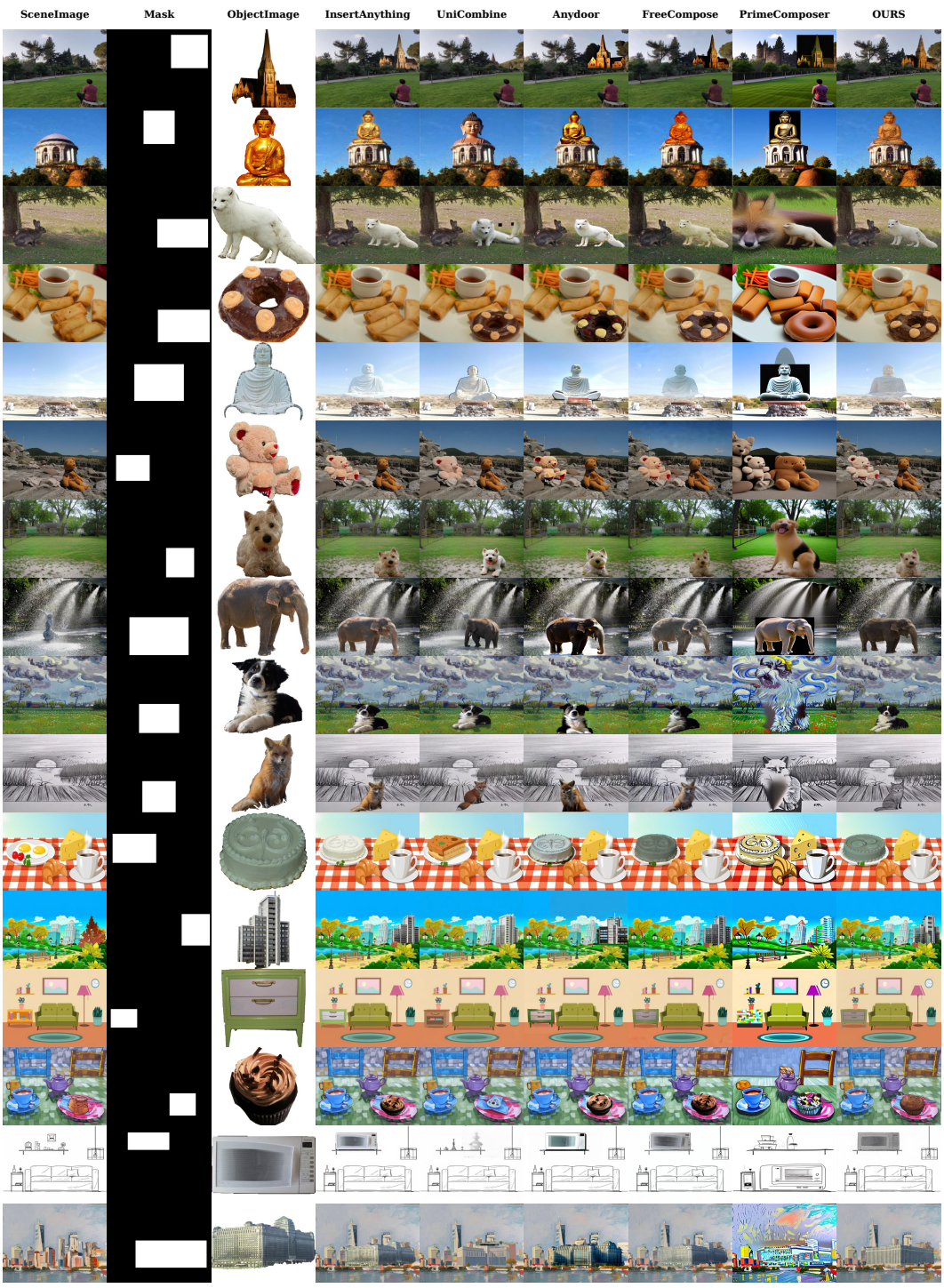

Figure 10: Visualization of comparative results on TF-ICON dataset, zoomed in for detailed observation.