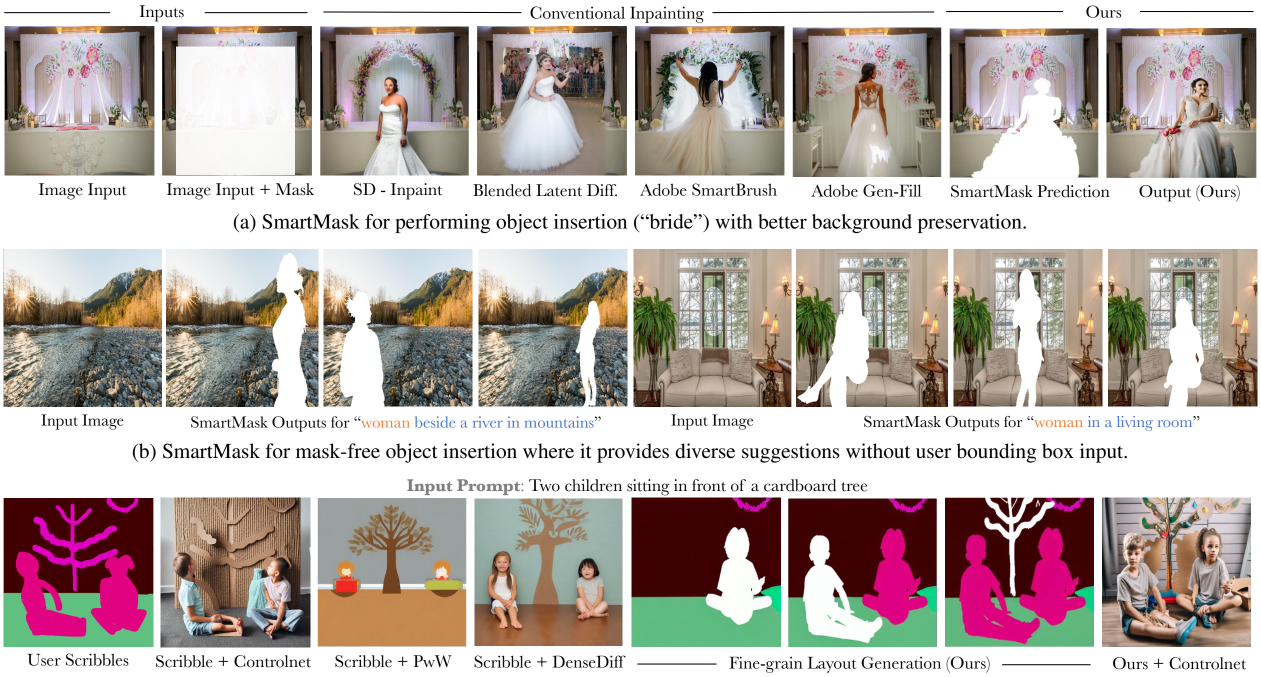
SmartMask: Context Aware High-Fidelity Mask Generation for Fine-grained Object Insertion and Layout Control

Jaskirat Singh[1,2] Jianming Zhang[1] Qing Liu[1] Cameron Smith[1] Zhe Lin[1] Liang Zheng[2]
[1]Adobe Research [2]Australian National University
https://smartmask-gen.github.io



(a) SmartMask for performing object insertion ("bride") with better background preservation.

(b) SmartMask for mask-free object insertion where it provides diverse suggestions without user bounding box input.

(c) SmartMask for fine-grain layout design for layout to image generation.

Figure 1. ***Overview***. We introduce *SmartMask* which allows a novice user to generate high-fidelity masks for fine-grained object insertion and layout control. The proposed approach can be used for object insertion (a-b) where it not only allows for image inpainting with better background preservation (a) but also provides diverse suggestions for mask-free object insertion at different positions and scales. We also find that when used iteratively *SmartMask* can be used for fine-grained layout design (c) for better quality semantic-to-image generation.

Abstract

The field of generative image inpainting and object insertion has made significant progress with the recent advent of latent diffusion models. Utilizing a precise object mask can greatly enhance these applications. However, due to the challenges users encounter in creating high-fidelity masks, there is a tendency for these methods to rely on more coarse masks (e.g., bounding box) for these applications. This results in limited control and compromised background content preservation. To overcome these limitations, we introduce SmartMask, which allows any novice user to create detailed masks for precise object insertion. Combined with a ControlNet-Inpaint model, our experiments demonstrate that SmartMask achieves superior object insertion quality, preserving the background content more effectively than previous methods. Notably, unlike prior works the proposed approach can also be used even without user-mask guidance, which allows it to perform mask-free object insertion at diverse positions and scales. Furthermore, we find that when used iteratively with a novel instruction-tuning based planning model, SmartMask can be used to design detailed layouts from scratch. As compared with user-scribble based layout design, we observe that SmartMask allows for better quality outputs with layout-to-image generation methods.

1. Introduction

Multi-modal object inpainting and insertion has gained widespread public attention with the recent advent of large-scale language-image (LLI) models [1–3, 19, 24, 27, 30, 37, 38]. A novice user can gain significant control over the inserted object details by combining text-based conditioning with additional guidance from a coarse bounding box or user-scribble mask. The text prompt can be used to describe the object semantics, while the coarse mask provides control over the position and scale of the generated object.

While convenient, the use of a coarse mask for object insertion suffers from two main limitations. **1) First,** the use of a coarse mask can often be undesirable as it tends to also modify the background regions surrounding the inserted object [30, 37] (refer Fig. 1a). In order to minimize the background artifacts, recent works [37, 40] also explore the use of user-scribble based free-form mask instead of a bounding-box input. However, while feasible for describing coarse objects (*e.g*., mountains, teddy bear *etc*.), the generation of accurate free-form masks for objects with a number of fine-grain features (*e.g*., humans) can be quite challenging especially when limited to coarse user-scribbles. **2)** Furthermore, generating variations in position and scale of the inserted object can also be troublesome as it requires the user to provide a new scene-aware free-form mask to satisfy the geometric constraints at the new scene location.

To address these drawbacks, we introduce *SmartMask*, a context-aware diffusion model which allows a novice user to directly generate fine-grained mask suggestions for precise object insertion. In particular, given a semantic object description (*e.g. kid*) and the overall scene context, *SmartMask* generates scene-aware precise object masks which can then be used as input to a ControlNet-inpaint model [36, 41] to perform object insertion while preserving the contents of background image. As compared with coarse-mask based inpainting methods, we find that *SmartMask* provides a highly convenient and controllable method for object insertion and can be used 1) *with user-inputs (bounding box, scribbles etc.)*: where the user can specify location and shape for the target object , or 2) *in a mask-free manner:* where the model automatically generates diverse suggestions for object insertion at diverse positions and scales.

In addition to object insertion, we also find that SmartMask can be used for fine-grain layout design. Existing segmentation-to-image (S2I) generation methods (*e.g.* ControlNet [41]) enable the generation of controllable image outputs from user-scribble based semantic segmentation maps. However, generating a good quality semantic layout can be quite challenging if the user wants to generate a scene with objects that require fine-grain details for best description (*e.g*., humans, chairs *etc*.). To address this challenge, we show that *SmartMask* when used with a novel instruction-tuning [34] based planning model allows the user to iteratively generate the desired scene layout from scratch. As compared with scribble based layout generation, we find that the proposed approach allows the users to better leverage existing S2I generation methods (*e.g.* ControlNet [41]) for higher quality output generation.

The main contributions of the paper are: 1) We propose *SmartMask* which allows any novice user to generate precise object masks for finegrained object insertion with better background preservation. 2) We show that unlike prior works, the proposed approach can also be used for mask-free object insertion. 3) Finally, we demonstrate that *SmartMask* can be used iteratively to generate detailed semantic layouts, which allows users to better leverage existing S2I generation methods for higher quality output generation.

2. Related Work

Diffusion based multi-modal image inpainting [1–3, 19, 24, 27, 30, 37, 38] has gained widespread attention with advent of text-conditioned diffusion models [24, 29, 30, 32, 39]. Despite their efficacy, these methods use coarse bounding box or user-scribble based masks for object inpainting which leads to poor background preservation around the inserted object. In contrast, *SmartMask* directly allows user to generate precise masks for the target object, which can then be combined with ControlNet-Inpaint [30, 41] to insert the target object while better preserving background contents.

Mask-free object placement has been studied in the context of image compositing methods [9, 20, 22, 25, 35, 42, 44, 46], where given a cropped RGB object instance and a target image, the goal is to suggest different positions for the target object. In contrast, we study the problem of mask-free object insertion using text-only guidance. Lee *et al*. [18] propose a GAN-based approach to directly model a distribution of potential object locations. However, the learned distribution is *w.r.t* to a specific object class (*e.g*., cars) which limits its generalizability for diverse use-cases.

Semantic-layout to image generation methods have been explored to enable controllable image synthesis from user-scribble based semantic segmentation maps [8, 12, 17, 23, 26, 33, 45]. Recently, [4, 6, 14, 31] propose a cross-attention based training-free approach for controlling the overall scene layout from coarse user-scribbles using text-conditioned diffusion models. Zhang *et al*. [41] propose a versatile ControlNet model which allows the users to control the output layout on a more fine-grained level through an input semantic map. While effective, generating desired semantic layout with scribbles can itself be quite challenging for scenes with objects that require fine-grain details for best description (*e.g*., humans). *SmartMask* helps address this problem by allowing users to generate more fine-grained layouts to facilitate better quality S2I generation.

3. Our Method

Given an input image $\mathcal{I}$, object semantic label $\mathcal{T}_{obj}$ and a textual description $\mathcal{T}_{context}$ describing the final scene con-

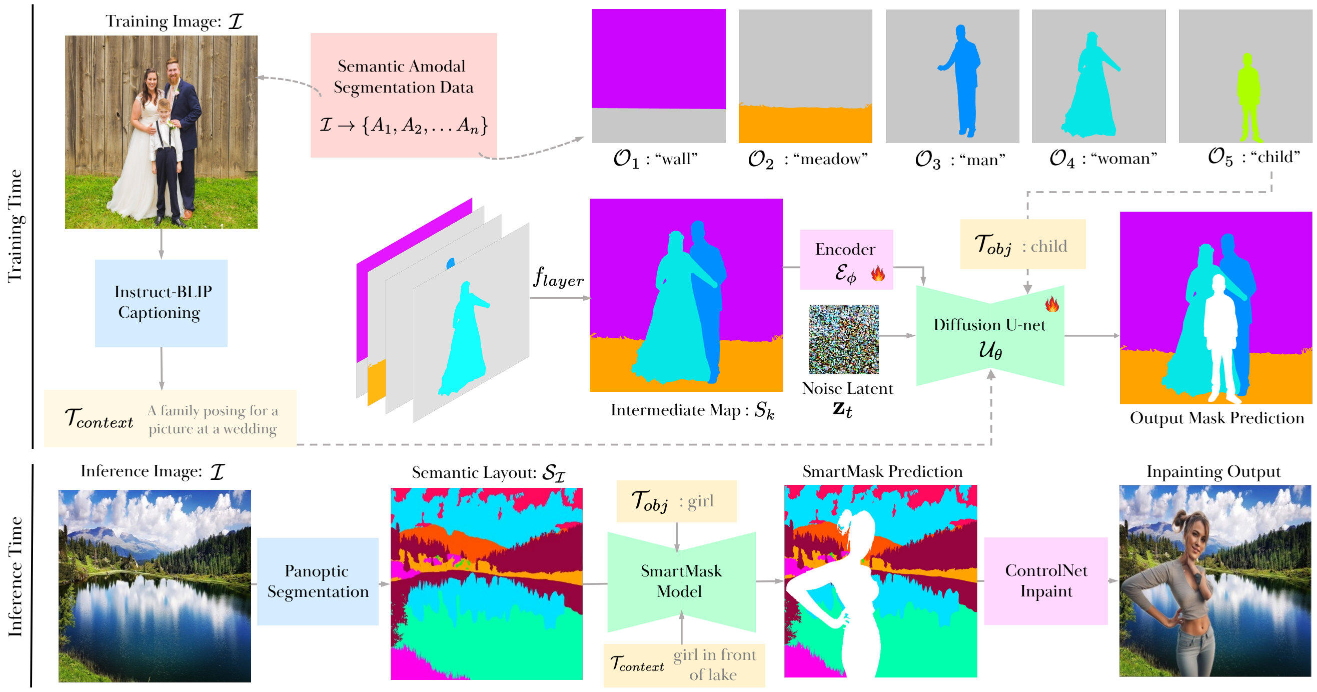

Training Time
Training Image: $\mathcal{I}$
Semantic Amodal Segmentation Data
$\mathcal{I} \rightarrow \{A_1, A_2, \dots A_n\}$
$\mathcal{O}_1$: "wall"
$\mathcal{O}_2$: "meadow"
$\mathcal{O}_3$: "man"
$\mathcal{O}_4$: "woman"
$\mathcal{O}_5$: "child"
Instruct-BLIP Captioning
f_{layer}
Intermediate Map : S_k
Encoder $\mathcal{E}_\phi$
Noise Latent $\mathbf{z}_t$
$\mathcal{T}_{obj}$: child
Diffusion U-net $\mathcal{U}_\theta$
Output Mask Prediction
$\mathcal{T}_{context}$ A family posing for a picture at a wedding
Inference Time
Inference Image: $\mathcal{I}$
Panoptic Segmentation
Semantic Layout: $\mathcal{S}_\mathcal{I}$
$\mathcal{T}_{obj}$: girl
SmartMask Model
$\mathcal{T}_{context}$ girl in front of lake
SmartMask Prediction
ControlNet Inpaint
Inpainting Output

Figure 2. ***Method Overview.*** A key idea behind *SmartMask* is to leverage semantic amodal segmentation data [28, 47] in order to obtain high-quality paired training annotations for mask-free single or multi-step object insertion. During training *(top)*, given a training image $\mathcal{I}$ with caption C, we stack k ordered instance maps $\{A_1, A_2, \dots A_k\}$ to obtain an intermediate semantic map S_k. The diffusion model is then trained to predict the instance map A_{k+1}, conditional on the semantic map S_k, $\mathcal{T}_{obj} \leftarrow \mathcal{O}_{k+1}$ and scene context $\mathcal{T}_{context} \leftarrow C$. During inference *(bottom)*, given a real image $\mathcal{I}$, we first use a panoptic segmentation model to compute semantic map $\mathcal{S}_I$. The generated semantic layout is then directly used as input to the trained diffusion model in order to predict the fine-grained mask for the inserted object.

text, our goal is predict a fine-grained mask $\mathcal{M}_{obj}$ for the target object. The object mask $\mathcal{M}_{obj}$ could then be used as input to a ControlNet-Inpaint model for fine-grained object insertion (Sec. 4.1) or used to design detailed semantic layouts from scratch (Sec. 4.3). For instance, *SmartMask* could be used for single object insertion, where given an image $\mathcal{I}$ depicting a man on a bench, $\{\mathcal{T}_{obj}$: *'woman'*$\}$ and $\{\mathcal{T}_{context}$: *'a couple sitting on a bench'*$\}$, our goal is to predict fine-grain binary mask $\mathcal{M}_{woman}$ which places the *'woman'* in a manner such that the resulting scene aligns with overall scene context of a *'a couple sitting on a bench'*.

Similarly, *SmartMask* could also be used for multiple object insertion. For instance, given an image $\mathcal{I}$ depicting a wedding, the user may wish to add multiple objects {*'man'*, *'woman'* and *'kid'*} such that the final scene aligns with the $\{\mathcal{T}_{context}$: *'a family posing for a picture at a wedding'*$\}$. Unlike prior image inpainting methods which are limited independently adding each object to the scene, the goal of the *smartmask* model is to add each object in a manner such the generated *'man'*, *'woman'* and *'kid'* appear to be *'a family posing for a picture at a wedding'*. (refer Fig. 2a).

In the next sections, we describe the key *SmartMask* components in detail. In particular, in Sec. 3.1 we discuss how *SmartMask* can leverage semantic amodal segmentation data in order to obtain high-quality paired annotations for mask-free object insertion. We then discuss a simple data-adaptation strategy which allows the user to also control the position, shape of the inserted object using coarse inputs (bounding-box, scribbles) in Sec. 3.2. Finally, in Sec. 3.3 we propose a visual-instruction tuning [21] based planning model which when used with *SmartMask* , allows for generation of detailed semantic layouts from scratch.

3.1. SmartMask for Mask-Free Object Insertion

Semantic-Space Task Formulation. Directly learning a model for our task in *pixel* space can be quite challenging, as it would require large-scale collection of training data for single or multi-step object insertion while maintaining background preservation. To address this key challenge, a core idea of our approach is to propose an equivalent task formulation which allows us to leverage large-scale semantic amodal segmentation data [28, 47] for generating high-quality paired training data in the *semantic* space (Fig. 2).

SmartMask Training. In particular during training, given an image $\mathcal{I}$, with a sequence of ordered amodal semantic instance maps $\{A_1, A_2 \dots A_n\}$ and corresponding semantic object labels $\{\mathcal{O}_1, \mathcal{O}_2 \dots \mathcal{O}_n\}$, we first compute an intermediate layer semantic map as,

$$S_k = f_{layer}(\{A_1, A_2 \dots A_k\}) \quad where\ k \in [1, n]. \qquad (1)$$

where k is randomly chosen from $[1, n]$ and f_{layer} is a layering operation which stacks the amodal semantic segmentation maps from $i \in [1, k]$ in an ordered manner (Fig. 2).

We next train a diffusion-based mask prediction model $\mathcal{D}_\theta$ which takes as input the above computed intermediate semantic layer map S_k, textual description for next object $\mathcal{T}_{obj} \leftarrow \mathcal{O}_{k+1}$, overall caption $\mathcal{T}_{context} \leftarrow C_\mathcal{I}$ (for image $\mathcal{I}$) and learns to predict the binary mask $\{A_{k+1}\}$ for the next object. To this end, we first pass the intermediate semantic map S_k through a learnable encoder $\mathcal{E}_\phi$ to obtain the encoded features $\mathcal{E}_\phi(S_k)$. At any timestep t of the reverse diffusion process, the denoising noise prediction ϵ_t is then computed conditional jointly on previous noise latent $\mathbf{z}_t$, and model inputs $\{\mathcal{E}_\theta(S_k)$,$\mathcal{T}_{obj}$, $\mathcal{T}_{context}\}$ as,

$$\tilde{\epsilon}_{pred}(t) = \mathcal{U}_\theta(\mathbf{z}_t, \mathcal{E}_\phi(S_k), \mathcal{T}_{obj}, \mathcal{T}_{context}, t), \qquad (2)$$

where $\mathcal{U}_\theta$ represents the U-Net of the diffusion model $\mathcal{D}$.

The overall diffusion model $\mathcal{D}$ is then trained to predict the next layer A_{k+1} using the following diffusion loss,

$$\mathcal{L}_t(\theta, \phi) = \mathbf{E}_{t\sim[1,T],S_k,A_{k+1},\epsilon_t}[||\epsilon_t - \tilde{\epsilon}_{pred}(t)||^2], \qquad (3)$$

where T is total number of reverse diffusion steps, $\epsilon_t \sim \mathcal{N}(0, I)$ is sampled from a normal distribution and A_{k+1} represents ground truth binary mask for next object $\mathcal{O}_{k+1}$.

SmartMask Inference. During inference, given an input image $\mathcal{I}$, semantic object category $\mathcal{T}_{obj}$ and a textual description $\mathcal{T}_{context}$ describing the final scene context, we first use a panoptic semantic segmentation model [15] to obtain the corresponding semantic layout map $\mathcal{S}_\mathcal{I}$. The generated semantic layout $\mathcal{S}_\mathcal{I}$ is then directly used as input to the above trained diffusion model $\mathcal{D}_\theta$ in order to predict the fine-grained mask $\mathcal{M}_{obj}$ for the target object,

$$\mathcal{M}_{obj} = \mathcal{D}_\theta(\mathcal{E}_\phi(\mathcal{S}_I), \mathcal{T}_{obj}, \mathcal{T}_{context}). \qquad (4)$$

3.2. Data Adaptation for Precise Mask Control

While the diffusion model trained in Sec. 3.1, allows the user to perform mask-free object insertion at diverse positions and scales, the user may also wish to obtain more direct control over the spatial location and details of the inserted object. To this end, we propose a simple train-time data adaptation strategy which allows *smartmask* to be easily adapted to diverse forms of user control (Fig. 4). In particular, given the intermediate layer map S_k computed using Eq. 1 and ground truth mask A_{k+1} for the next object, we replace the input S_k to the diffusion model as,

$$\tilde{S_k} = g(S_k, G_{obj}) = S_k \odot (1 - \alpha\, G_{obj}) + \alpha\, G_{obj}, \qquad (5)$$

where G_{obj} is the additional guidance input (*e.g.*, bounding box mask, coarse scribbles *etc.*) provided by the user and $\alpha = 0.7$ helps add additional guidance while still preserving the content of the original input S_k after data adaptation.

Training. In this paper, we mainly consider four main guidance inputs G_{obj} for additional mask control for adapting *SmartMask* . *1) Mask-free guidance:* in absence of any additional user inputs, we use $G_{obj} = \mathbf{0}^{H,W}$ which prompts the model to suggest fine-grained masks for object insertion at diverse positions and scales. *2) Bounding-box guidance:* we set G_{obj} as a binary mask corresponding to ground truth object mask A_{k+1}. *3) Coarse Spatial Guidance:* Expecting the user to provide precise bounding box for object insertion is not always convenient and can lead to errors if bounding box is not correct. We therefore introduce a coarse spatial control where user may provide a coarse spatial location and the model learns to infer the best placement of the object around the suggested region (Fig. 4c). During training, the same is achieved by setting G_{obj} as a coarse gaussian blob centered at the ground truth object mask A_{k+1}. *4) User scribbles:* Finally, we also allow the user to describe target object using free-form coarse scribbles, by setting G_{obj} as the dilated mask output of ground-truth object mask A_{k+1}.

Inference. At inference time, the additional guidance input G_{obj} (*e.g.* bounding box mask, coarse scribbles *etc.*) is directly provided by the user. Given an input image $\mathcal{I}$ with semantic layout $\mathcal{S}_\mathcal{I}$, we then use transformation from Eq. 5 as input to the adapted *SmartMask* model in order to generate object insertion suggestions with additional control.

3.3. Global Planning for Multi-Step Inference

While the original *SmartMask* model allows the user to generate fine-grained masks for single object insertion, we would also like to use *SmartMask* for iterative use cases such as multiple object insertion (Sec. A.1) or designing a fine-grained layout from scratch with large number (> 10) of scene elements. Such an iterative use of *SmartMask* would require the model to carefully plan the spatial location of each inserted object to allow for the final scene to be consistent with the final scene context description $\mathcal{T}_{context}$.

To achieve this, we train a visual-instruction tuning [21] based planning model which given the input semantic layout $\mathcal{S}_I$, learns to plan the positioning of different scene elements over long sequences. Given a semantic object description $\mathcal{T}_{obj}$ and final scene context $\mathcal{T}_{context}$, the global planning model provides several bounding box suggestions for object insertion. *SmartMask* model then uses the above predictions as coarse spatial guidance input G_{obj} to provide fine-grained mask suggestions for the next object. The above process can then be repeated in an iterative manner until all objects have been added to the scene (refer Fig. 7a).

The global planning model is trained in two stages. 1) **Feature Alignment.** Typical instruction tuning models are often trained on real images. In contrast, as discussed in Sec. 3.1 we would like to model our problem in semantic space as it allows us to leverage amodal segmentation data for training. To address this domain gap, we first finetune an existing LLaVA model [21] to understand the semantic inputs. To do this, given an intermediate semantic map S_k computed using Eq. 1, we finetune the projection matrix $\mathbf{W}$

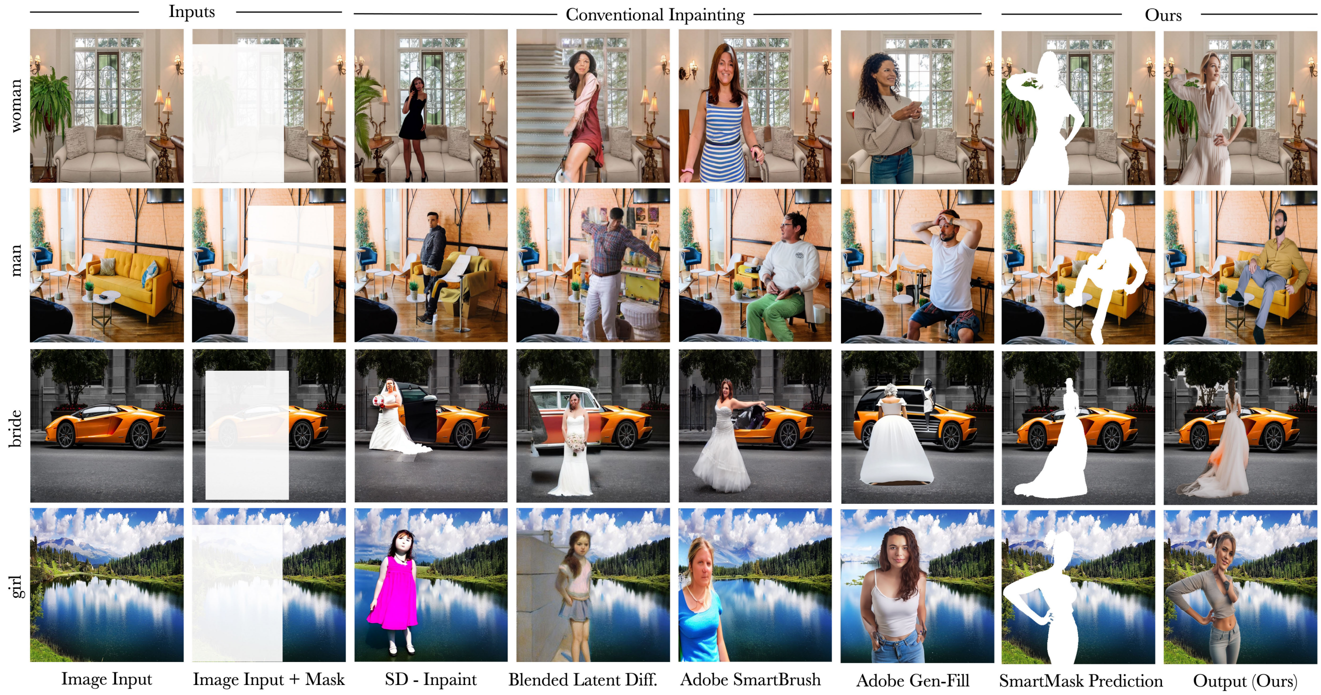


Figure 3. ***Qualitative Results for Image Inpainting.*** We observe that as compared to with state-of-the-art image inpainting [1, 2, 30, 37] methods, *SmartMask* allows the user to perform object insertion while better preserving the background around the inserted object.

of the LLaVA model [21] $\mathcal{H}$ to predict the semantic object labels $\{\mathcal{O}_1, \mathcal{O}_2 \ldots \mathcal{O}_k\}$ described in the current scene as,

$$\mathcal{L}_{align}(\mathbf{W}) = \mathcal{L}_{CE}(< \mathcal{O}_1, \mathcal{O}_2 \ldots \mathcal{O}_k >, \mathcal{H}(S_k)). \quad (6)$$

2) **Instruction-Tuning.** Finally, keeping the visual encoder weights for LLaVA model fixed, we next finetune both project matrix $\mathbf{W}$ and LLM weights Φ [43] for global object planning. In particular, given an intermediate semantic map $\mathcal{S}_k$, we first compute the bounding box coordinates $\mathcal{B}_{k+1} = \{x_{min}, y_{min}, x_{max}, y_{max}\}$ for the next object $\mathcal{T}_{obj} = \mathcal{O}_{k+1}$ using ground truth object mask $\mathcal{A}_{k+1}$. The LLaVA based planning model $\mathcal{H}$ is then trained as,

$$\mathcal{L}_{instruct}(\mathbf{W}, \Phi) = \mathcal{L}_{CE}(\mathcal{B}_{k+1}, \mathcal{H}(S_k, \mathcal{T}_{obj}, C)), \quad (7)$$

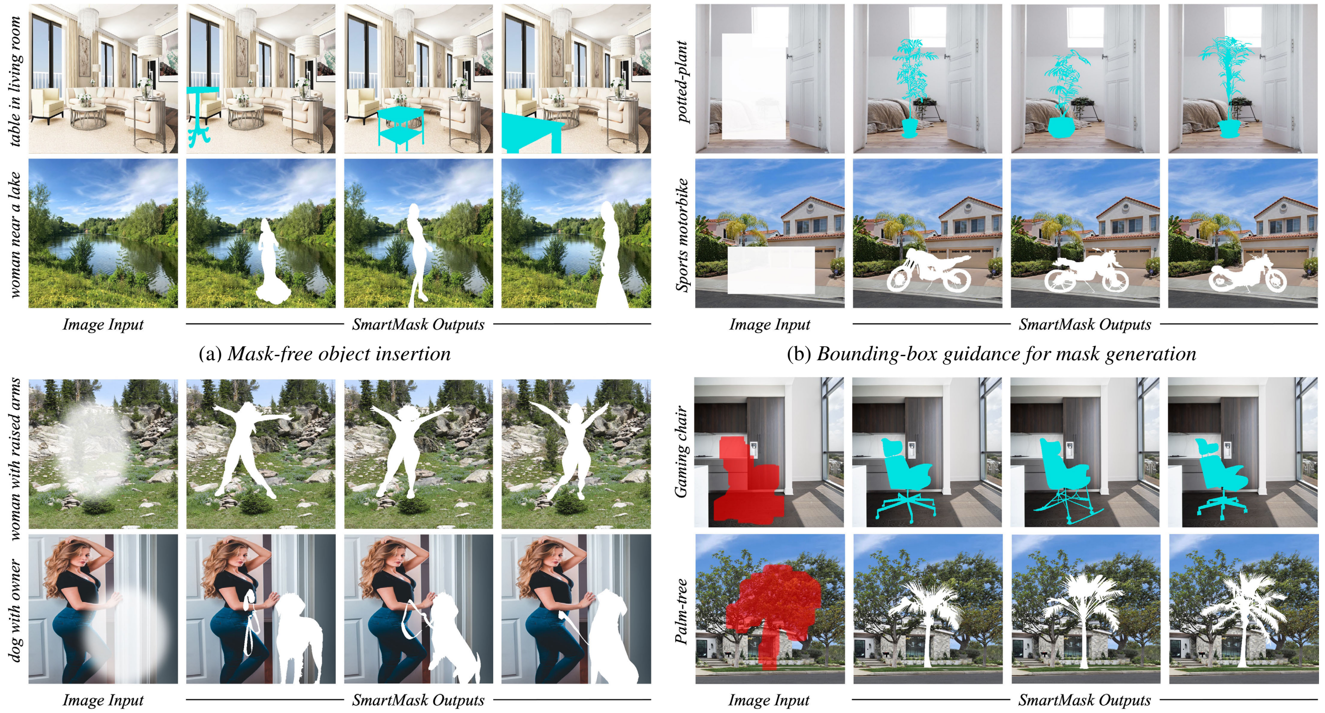
where C represents the caption for the final scene (obtained using ground-truth image $\mathcal{I}$) and provides the model context for placing different scene elements in the image.

4. Experiments

Training Data Collection. As discussed in Sec. 3, we note that a key idea behind *SmartMask* is to model the object insertion problem in semantic space (instead of pixel space), which allows us to leverage semantic amodal segmentation data to obtain large-scale paired training annotations for single or multiple object insertions. However, traditional datasets for semantic amodal segmentations such as COCO-A [47] (2,500 images, 22,163 instances) and KINS [28] (7,517 images, 92,492 instances) though containing fine-grained amodal segmentation annotations may lack sufficient diversity to generalize across different use-cases.

To address this, we curate a new large-scale dataset consisting of fine-grain amodal segmentation masks for different objects in an input image. The overall dataset consists of 32785 diverse real world images and a total of 725897 object instances across more than 500 different semantic classes (*e.g*. man, woman, trees, furniture *etc*.). Each image $\mathcal{I}$ in the dataset consists of a variable number of object instances $\{A_1, A_2, \ldots A_n\}$, $n \in [2, 50]$ and is annotated with an ordered sequence of semantic amodal segmentation maps $\{S_1, S_2 \ldots S_n\}$. The detailed descriptions C_I for each image are obtained using the InstructBLIP [7] model.

SmartMask Training. In order to leverage the rich generalizable prior of T2I diffusion models, we use the weights from publicly available Stable-Diffusion-v1.5 model [30] in order to initialize the weights of the *SmartMask* U-Net model trained in Sec. 3.1. Similar to [5], we modify the architecture of the U-Net model to also condition the output mask predictions on segmentation layout $\mathcal{S}_I$. The *SmartMask* model is trained for a total of 100k iterations with a batch size of 192 and learning rate $1e-5$ using 8 Nvidia-A100 GPUs. During inference, a panoptic semantic segmentation model finetuned on the dataset in Sec. 4 is used for converting real image $\mathcal{I}$ to its semantic layout $\mathcal{S}_I$. ControlNet model trained with SDXL backbone was used to perform precise object insertion with SmartMask outputs. Please refer the supp. material for further training details.



(a) *Mask-free object insertion*

(b) *Bounding-box guidance for mask generation*

(c) *Coarse-spatial guidance for mask generation*

(d) *Scribble-based object insertion*

Figure 4. ***Diverse User Controls***. *SmartMask* is easily adaptable to diverse user-controls for precise mask generation for the target object.

4.1. SmartMask for Object Insertion

Baselines. We compare the performance of our approach on object-insertion with prior works on performing multi-modal image-inpainting using a textual description and coarse bounding box mask. In particular we show comparisons with SD-Inpaint [30], SDXL-Inpaint [27], Blended-Latent Diffusion [30], and Adobe SmartBrush [37]. We also compare the performance of our approach with state-of-the-art commercial inpainting tools by reporting results on recently released Generative-Fill from Adobe Firefly [1].

Evaluation Metrics. Following [37], we report the results for object insertion using *1) Local-FID* [11] which measures the realism of the generated objects, *2) CLIP-Score* [10] which measures the alignment between the textual description and the generated object, and *3) Norm. L2-BG:* which reports the normalized *L2* difference in the background regions before and after insertion, and helps capture the degree to which the background was preserved.

Qualitative Results. Results are shown in Fig. 3. We observe that when performing objectn insertion using a coarse bounding box mask, traditional inpainting methods usually lead to a lot of changes in the background regions around the inserted object (*e.g.* living room details in row-1&2, mountains in row-4 *etc.*). Adobe SmartBrush [37] which is trained to allow better background preservation, shows better performance, however, still suffers from notable changes to background regions. In contrast, by directly predicting a high-fidelity mask for the target object, the proposed approach allows the user to add new objects on the scene with minimal changes to the background image. Furthermore we observe that target object masks are generated in a scene-aware manner, which helps us add new objects while interacting with already existing ones. For instance, when adding *'a man to a couch with table in front'* (Fig. 3), prior works typically replace the couch and table to insert the target object (*'man'*). In contrast, *smartmask* places the *'man sitting on the couch with his leg on the table'*, and provides a more natural way for inserting objects in complex scenes.

Method	Evaluation Criteria		
	Local-FID ↓	CLIP-Score ↑	Norm. L2-BG ↓
SD Inpaint [30]	22.31	0.249	0.374
SDXL Inpaint [27]	21.84	0.235	0.623
Blended L-Diffusion [2]	39.77	0.253	0.451
Adobe SmartBrush [37]	17.94	0.262	0.304
Adobe Gen-Fill* [1]	N/A	**0.268**	0.289
Smartmask (Ours)	19.21	0.261	**0.098**

Table 1. ***Quantitative results for image inpainting***. We observe that in comparison with state-of-art image inpainting methods, our approach leads to better preservation of background regions.

Quantitative Results. In addition to qualitative results, we also report the performance of our approach quantitatively in Tab. 1. We find that similar to results in Fig. 3, the proposed approach shows better background preservation (Norm. L2-BG ↓) while performing comparably in terms of image quality (local-FID) and text-alignment (CLIP-Score).

4.2. Evaluating Mask Controllability and Quality

A key advantage of *SmartMask* is the ability to generate high-quality masks for target object in a controllable man-

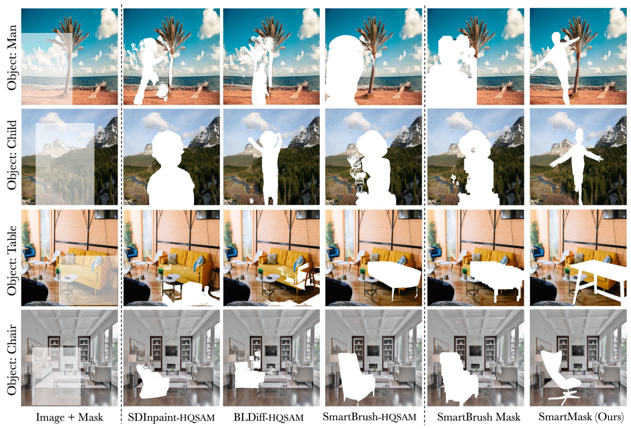


Figure 5. ***Comparing output mask quality*** with different Inpaint + HQSAM (middle) methods and SmartBrush mask output (right).

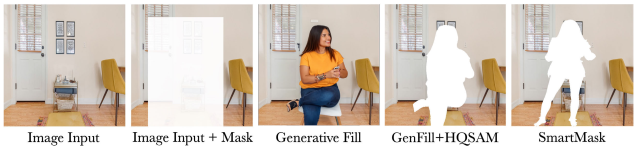


Figure 6. ***Limitation of Inpaint + HQSAM***. In addition to poor mask quality errors (Fig. 5), we observe that *Inpaint+HQSAM* can lead to scene-unaware masks (*e.g. mask for woman sitting in air*) .

ner. In this section, we evaluate the performance of *SmartMask* in terms of 1) user control, & 2) output mask quality.

1) User Control. As shown in Fig. 4, we observe that *SmartMask* allows the user to control the output object mask in four main ways. *1) Mask-free insertion:* where the model automatically suggests diverse positions and scales for the target object (*e.g*., *table, woman* in Fig. 4a). *2) Bounding-box guidance:* which allows user to specify the exact bounding box for object insertion (*potted plant, motorbike* in Fig. 4b). *3) Coarse spatial guidance:* (Fig. 4c) providing a precise bounding box can be challenging for cases with complex object insertions *e.g*., *dog with owner*. *SmartMask* allows the user to only specify a coarse location for the target object, and the model automatically adjusts the object placement (*i.e. dog with head near woman's hand*) to capture object interactions. *3) User-scribbles:* Finally, the user may also control the output shape by providing coarse scribbles. The *smartmask* model can use this as guidance to automatically predict the more finegrain-masks for the target object (*e.g*., palm-tree, gaming chair in Fig. 4d).

2) Output Mask Quality. We also report results on the quality of generated masks by showing comparisons with the mask-prediction head of the SmartBrush model [37]. Furthermore, we also show comparisons combining standard inpainting methods with HQSAM [13, 16]. To this end, we first use the provided bounding-box mask to inpaint the target object. The user-provided bounding-box and the inpainted output are then used as input to the HQSAM model [13] to obtain target object-mask predictions.

Method	User Study Results		
	Win ↑	Draw	Lose ↓
SDInpaint + HQSAM [13, 30]	92.04%	5.12%	2.84%
Blended L-Diff + HQSAM [2, 13]	88.91%	8.33%	2.75%
SmartBrush + HQSAM [13, 37]	63.89%	27.78%	8.34%
SmartBrush Mask [37]	91.67%	2.78%	5.56%

Table 2. ***User study results***. For evaluating generated mask quality. We observe that *SmartMask* generates higher quality masks as compared to SmartBrush and various Inpaint+HQSAM methods.

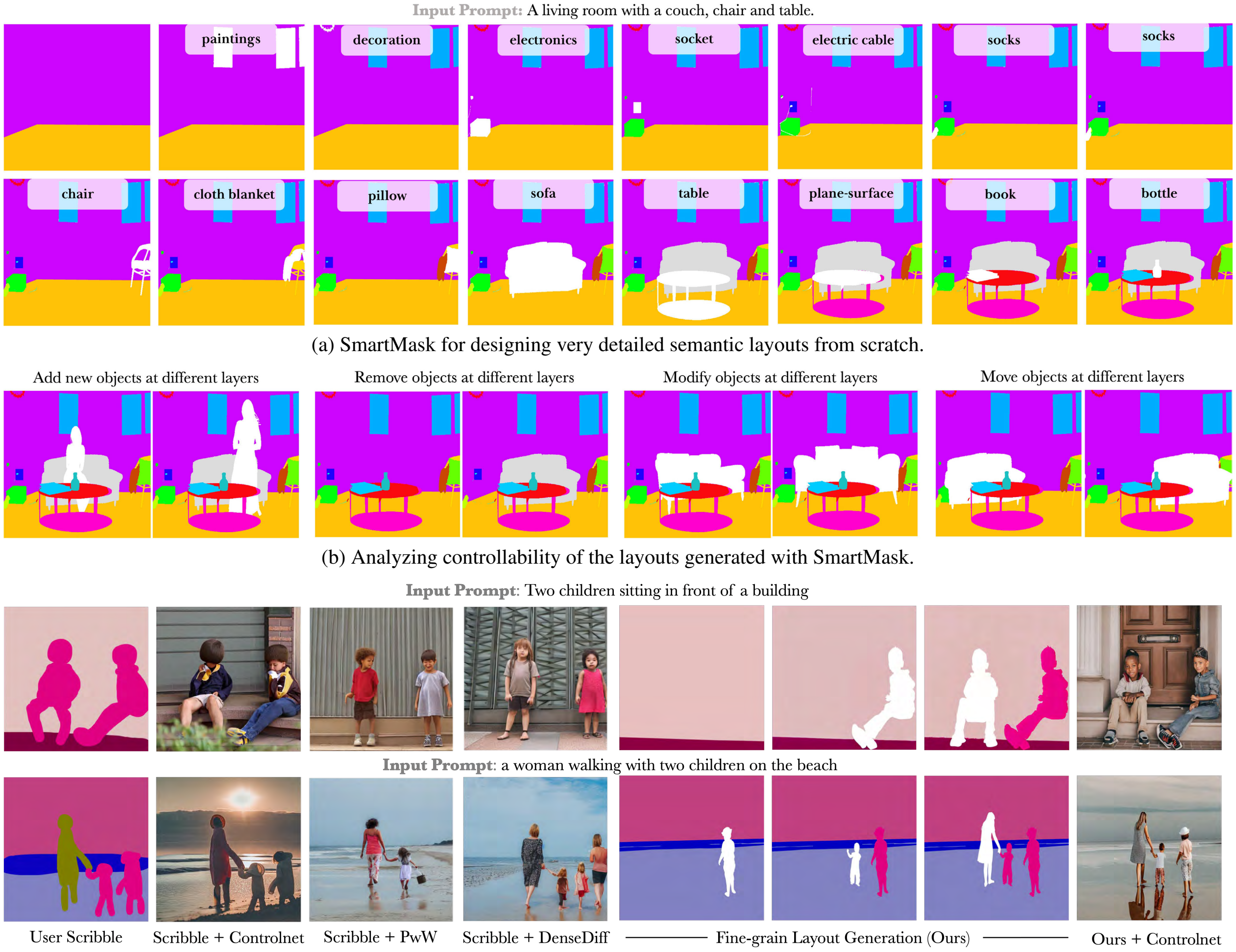
Results are shown in Fig. 5. We observe that as compared to outputs of SmartBrush [37] mask-prediction head, *SmartMask* generates higher-quality masks with fewer artifacts. Similarly, while using HQSAM on inpainting outputs helps achieve good mask quality for some examples (*e.g*. child in row-2), the HQSAM generated masks (or the inpainted image) often have accompanying artifacts which limits the quality of the output masks. In addition to poor mask quality errors, we also observe that Inpaint+HQSAM can lead to scene-unaware masks (Fig. 6). This occurs because prior inpainting methods typically add additional objects in the background when performing object insertion. For instance, when inserting *woman in a living room* in Fig. 6), we observe that Adobe Gen-Fill [1] adds an additional chair on which the woman is sitting. Extracting only the object mask for such inpainted outputs can lead to scene-unaware masks where *'the woman appears floating in the air'* as the chair was not present in the original image.

The above findings are also reflected in a quantitative user study (Tab. 2), where human subjects are shown a pair of object mask suggestions (ours vs baselines discussed above), and asked to the select the mask suggestion with the higher quality. As shown in Tab. 2, we observe that *SmartMask* outputs are preferred by majority of human subjects over SmartBrush mask [37] and Inpaint + HQSAM outputs.

4.3. SmartMask for Semantic Layout Design

In addition to object insertion, we also find that when used iteratively along with the visual-instruction tuning based planning model from Sec. 3.3, *SmartMask* forms a convenient approach for designing detailed semantic layouts with a large number of fine-grain objects (*e.g*. humans, furniture *etc*.). Results are shown in Fig. 7a. We observe that given a sequence of user provided scene elements (*e.g*. painting, sofa, chair *etc*.), *SmartMask* generates the entire scene layout from scratch. Furthermore, unlike static layouts generated by a panoptic segmentation model, *SmartMask* generated layouts allow the user greater control over the details of each scene element. Since each object in the final layout is represented by a distinct object mask, the final layouts are highly controllable and allow for a range of custom operations such as adding, removing, modifying or moving objects through simple layer manipulations. (refer Fig. 7b).

Controllable S2I Generation. Layout to image gener-


Input Prompt: A living room with a couch, chair and table.
paintings
decoration
electronics
socket
electric cable
socks
socks
chair
cloth blanket
pillow
sofa
table
plane-surface
book
bottle
Add new objects at different layers
Remove objects at different layers
Modify objects at different layers
Move objects at different layers
Input Prompt: Two children sitting in front of a building
Input Prompt: a woman walking with two children on the beach
User Scribble
Scribble + Controlnet
Scribble + PwW
Scribble + DenseDiff
Fine-grain Layout Generation (Ours)
Ours + Controlnet

(a) SmartMask for designing very detailed semantic layouts from scratch.

(b) Analyzing controllability of the layouts generated with SmartMask.

(c) Using SmartMask generated layouts for better quality layout-to-image generation.

Figure 7. ***Fine-grained layout design.*** We observe that SmartMask when used iteratively, allows the user to generate very detailed layouts from scratch (a). The generated layouts are highly controllable and allow for custom variations through simple layer manipulations (b).

ation methods *e.g*., ControlNet [41] enable the generation of controllable image outputs from user-scribble based semantic segmentation maps or layouts. However, generating the user-desired layouts with coarse scribbles can itself be quite challenging for scenes with objects that require finegrain details for best description (*e.g*. humans, chairs *etc*.). As shown in Fig. 7c, we find that this can lead to image outputs with either deformity artifacts (child in row-1) or incorrect description (woman and children in row-2) when using ControlNet [41]. A similar problem is also observed in other coarse-scribble based S2I methods such as DenseDiffusion [14] and Paint-with-Words (PwW) [4], which provide coarse control over object position but are unable to control finegrain details such as pose, action *etc*. of the target object. *SmartMask* helps address this problem by allowing any novice user to generate controllable (Fig. 7b) fine-grain layouts from scratch, which can allow users to better leverage existing S2I methods [41] for higher quality layout-to-image generation (refer Fig. 7c).

5. Conclusion

In this paper, we present *SmartMask* which allows a novice user to generate scene-aware precision masks for object insertion and finegrained layout design. Existing methods for object insertion typically rely on a coarse bounding box or user-scribble input which can lead to poor background preservation around the inserted object. To address this, we propose a novel diffusion based framework which leverages semantic amodal segmentation data in order to learn to generate fine-grained masks for precise object insertion. When used along with a ControlNet-Inpaint model, we show that the proposed approach achieves superior object-insertion performance, preserving background content more effectively than previous methods. Additionally, we show that *SmartMask* provides a highly controllable approach for designing detailed layouts from scratch. As compared with user-scribble based layout design, we observe that the proposed approach can allow users to better leverage existing S2I methods for higher quality layout-to-image generation.

References

[1] Adobe. Adobe firefly – generative ai for everyone, 2023. 2, 5, 6, 7

[2] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM Transactions on Graphics (TOG)*, 42(4):1–11, 2023. 5, 6, 7

[3] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022. 2

[4] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 2, 8

[5] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 5

[6] Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. *arXiv preprint arXiv:2304.03373*, 2023. 2

[7] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 5

[8] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12873–12883, 2021. 2

[9] Arnab Ghosh, Richard Zhang, Puneet K Dokania, Oliver Wang, Alexei A Efros, Philip HS Torr, and Eli Shechtman. Interactive sketch & fill: Multiclass sketch-to-image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1171–1180, 2019. 2

[10] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 6

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[12] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 2

[13] Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan Liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment anything in high quality. In *NeurIPS*, 2023. 7

[14] Yunji Kim, Jiyoung Lee, Jin-Hwa Kim, Jung-Woo Ha, and Jun-Yan Zhu. Dense text-to-image generation with attention modulation. In *ICCV*, 2023. 2, 8

[15] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9404–9413, 2019. 4

[16] 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. *arXiv preprint arXiv:2304.02643*, 2023. 7

[17] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5549–5558, 2020. 2

[18] Donghoon Lee, Sifei Liu, Jinwei Gu, Ming-Yu Liu, Ming-Hsuan Yang, and Jan Kautz. Context-aware synthesis and placement of object instances. *Advances in neural information processing systems*, 31, 2018. 2

[19] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023. 2

[20] Chen-Hsuan Lin, Ersin Yumer, Oliver Wang, Eli Shechtman, and Simon Lucey. St-gan: Spatial transformer generative adversarial networks for image compositing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9455–9464, 2018. 2

[21] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 3, 4, 5

[22] 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. 2

[23] Xihui Liu, Guojun Yin, Jing Shao, Xiaogang Wang, et al. Learning to predict layout-to-image conditional convolutions for semantic image synthesis. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[24] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[25] Li Niu, Qingyang Liu, Zhenchen Liu, and Jiangtong Li. Fast object placement assessment. *arXiv preprint arXiv:2205.14280*, 2022. 2

[26] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019. 2

[27] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2, 6

[28] Lu Qi, Li Jiang, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Amodal instance segmentation with kins dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3014–3023, 2019. 3, 5

[29] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu,

and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2

[30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 2, 5, 6, 7

[31] Jaskirat Singh, Stephen Gould, and Liang Zheng. High-fidelity guided image synthesis with latent diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5997–6006. IEEE, 2023. 2

[32] Jaskirat Singh and Liang Zheng. Divide, evaluate, and refine: Evaluating and improving text-to-image alignment with iterative vqa feedback. *Advances in Neural Information Processing Systems*, 36:70799–70811, 2023. 2

[33] Vadim Sushko, Edgar Schönfeld, Dan Zhang, Juergen Gall, Bernt Schiele, and Anna Khoreva. You only need adversarial supervision for semantic image synthesis. *arXiv preprint arXiv:2012.04781*, 2020. 2

[34] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2

[35] 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*, pages 461–470, 2019. 2

[36] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022. 2

[37] Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. Smartbrush: Text and shape guided object inpainting with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22428–22437, 2023. 2, 5, 6, 7

[38] Shiyuan Yang, Xiaodong Chen, and Jing Liao. Uni-paint: A unified framework for multimodal image inpainting with pretrained diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3190–3199, 2023. 2

[39] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022. 2

[40] Yu Zeng, Zhe Lin, and Vishal M Patel. Shape-guided object inpainting. *arXiv preprint arXiv:2204.07845*, 2022. 2

[41] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2, 8

[42] Lingzhi Zhang, Tarmily Wen, Jie Min, Jiancong Wang, David Han, and Jianbo Shi. Learning object placement by inpainting for compositional data augmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 566–581. Springer, 2020. 2

[43] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. 5

[44] Siyuan Zhou, Liu Liu, Li Niu, and Liqing Zhang. Learning object placement via dual-path graph completion. In *European Conference on Computer Vision*, pages 373–389. Springer, 2022. 2

[45] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5104–5113, 2020. 2

[46] Sijie Zhu, Zhe Lin, Scott Cohen, Jason Kuen, Zhifei Zhang, and Chen Chen. Topnet: Transformer-based object placement network for image compositing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1838–1847, 2023. 2

[47] Yan Zhu, Yuandong Tian, Dimitris Metaxas, and Piotr Dollár. Semantic amodal segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1464–1472, 2017. 3, 5