Text2Place: Affordance-aware Text Guided Human Placement

Rishubh Parihar, Harsh Gupta, Sachidanand VS, and R. Venkatesh Babu

Vision and AI Lab, Indian Institute of Science, Bangalore
Project Page

Abstract. For a given scene, humans can easily reason for the locations and pose to place objects. Designing a computational model to reason about these affordances poses a significant challenge, mirroring the intuitive reasoning abilities of humans. This work tackles the problem of realistic human insertion in a given background scene termed as **Semantic Human Placement**. This task is extremely challenging given the diverse backgrounds, scale, and pose of the generated person and, finally, the identity preservation of the person. We divide the problem into the following two stages **i)** learning *semantic masks* using text guidance for localizing regions in the image to place humans and **ii)** subject-conditioned inpainting to place a given subject adhering to the scene affordance within the *semantic masks*. For learning semantic masks, we leverage rich object-scene priors learned from the text-to-image generative models and optimize a novel parameterization of the semantic mask, eliminating the need for large-scale training. To the best of our knowledge, we are the first ones to provide an effective solution for realistic human placements in diverse real-world scenes. The proposed method can generate highly realistic scene compositions while preserving the background and subject identity. Further, we present results for several downstream tasks - scene hallucination from a single or multiple generated persons and text-based attribute editing. With extensive comparisons against strong baselines, we show the superiority of our method in realistic human placement.

Keywords: Spatial relations · Human inpainting

1 Introduction

Given a background scene, humans can easily visualize how persons can interact with the scene in multiple ways. For, e.g., for a living room, one can imagine a person sitting on the sofa, or walking towards the door. To understand this relationship between humans and scenes, J.J. Gibson coined the term *affordances* [22] which points to the interaction between objects and the environment. Designing computational models for this task is extremely challenging and is crucial for common sense visual understanding. Earlier methods for human affordance predictions are constrained by the specific datasets [12, 18, 66]. To generalize affordance predictions, a recent method [35] trained with a large



Fig. 1: (Top): Proposed approach for text-based placement of humans. Given a background image, we predict the plausible semantic region compatible with the text prompt to place humans. Next, given a few subject images, we perform subject-conditioned inpainting to realistically place humans in appropriate poses following the scene affordances. **(Bottom)**: Our method enables **a)** realistic human placements at diverse locations and poses and several downstream applications **b)** scene hallucination by generating compatible scenes for the given pose of the human **c)** text-based editing of the human and **d)** placing multiple persons.

dataset of human videos to place humans in a given bounding box. However, this formulation only models the *local* human affordance in the given bounding box but cannot reason about the *global* human affordance, such as where a person can sit or stand. In this work, we aim to learn generalized local and global human affordances *without needing large-scale training*. Specifically, we guide the affordances through action prompts (e.g., *'a person sitting on sofa'*).

We model the human affordance of a scene as *semantic masks* covering plausible image regions to place humans. Training a model to predict the *semantic masks* for *unseen* objects (e.g. person) is challenging primarily due to the unavailability of datasets. Typical image description datasets have annotations for what is present in the scene, such as object locations or scene captions, and do not provide information regarding the *affordances*. Existing approaches [35, 54] for affordance prediction adapt these datasets by inpainting objects to create synthetic datasets of image pairs *with/without* objects. However, these methods are extremely expensive to train and are limited to the objects on which the model is trained. To facilitate a more flexible setting, we build upon large text-to-image [55] (T2I) generative models. These models have rich priors for object-scene composition; which are implicit in the generation process. *Can we*

use these priors from T2I models to obtain text-guided affordances for a given scene? To this end, we propose an insightful approach that uses score distillation sampling (SDS) [49] from T2I models to optimize *semantic masks* for learning human affordances. Specifically, given a background image and an action prompt (e.g., *'a person sitting on the sofa'*), we use SDS loss to optimize parametric *semantic mask* to localize regions to place humans. Naively parameterizing *semantic masks* in the pixel space leads to collapse where the optimized mask covers the full image. To regularize the semantic masks, we propose a novel blob-based parameterization that constrains the semantic mask to a suitable region according to the required human pose.

The obtained *semantic mask* is used to perform subject-conditioned inpainting for scene completion. Given a few images of the subject, we project it into the text token embedding space of pretrained T2I model [55] using Textual Inversion [20]. Next, we use the learned token embedding and the action prompt to place the subject using the inpainting pipeline [55]. Large T2I models have excellent inpainting capability and can adjust the person's pose given a *semantic mask* for realistic scene completions (Fig. 1). We call this problem for placing persons following affordances as **Semantic Human Placement (SHP)**. To the best of our knowledge, we provide the first solution to solve SHP for realistic human placement in in-the-wild scenes.

We present extensive results for SHP on diverse indoor and outdoor scenes. Further, we show multiple downstream applications of *scene hallucination* conditioned on the single or multiple generated persons. Subject conditioned inpainting with T2I models enables text-based editing of the generated person. Additionally, our method generalizes to place diverse objects beyond humans in the scene. We compare our method against several baselines, including recent Vision Language Models models [1, 39], and achieve superior scene completion with background preservation evident in quantitative evaluations and user study. In summary, our primary contributions are:

1. A novel problem formulation of **Semantic Human Placement (SHP) for** realistically placing a given subject in background scenes.
2. Method to effectively parametrize semantic masks and learning masks using distillation from text-to-image models without large-scale training.
3. A subject conditioned inpainting pipeline to generate identity-preserving realistic human placements from a few subject images.
4. Demonstrate the efficacy of our models on downstream tasks of person and scene hallucination, composing multiple persons and placing other objects.

2 Related Work

Object and scene affordances. Inpired by the earlist notion of afforances [22], several works have been proposed to perform affordance prediction [11, 13, 19, 24, 25, 30, 38, 50, 67]. These works have focused on modeling human-object [7, 23, 34, 74, 90] and human-scene affordances [6, 37, 64]. An interesting framework is to

learn human affordances from videos [19, 50, 67]. In an insightful work [66] learn human affordances from a large video dataset of sitcom shows. Specifically, they learn plausible human poses for each of the sitcom scenes, which limits their applicability in learning the affordances of diverse real scenes. Brooks [5] train a generative model to learn pose-conditioned human and scene generation to generate realistic compositions. Further, they allow for diverse scene generations that are compatible with a given human pose. More recently, Kulal [35] trained a large diffusion model to place humans in a defined bounding box with appropriate local affordances. For learning object-scene, Ramrakhya [54] curated a large-scale dataset of image pairs with/without the object of interest to train object affordances for a given scene. A concurrent work SmartMask [60] performs a large-scale training of diffusion model to predict object masks. All the above methods, although learning good affordances, either rely on large-scale training, require guidance as a bounding box, or do not generalize to novel object categories. Our work aims to learn local and global human affordances in diverse scenes without a need for large-scale training and also to support affordance for novel objects.

Inpainting. The task for inpainting is to fill a given masked region with plausible pixels to facilitate realistic image completion. Earlier attempts for inpainting explore local image features [2, 4, 15, 44]. Later, several works were designed to perform inpainting with large-scale training of CNNs [72, 77, 78, 84, 85] and transformers [3, 17, 79]. Recent works leverage the exceptional generation quality of diffusion models to perform inpainting [42, 57, 71]. The closest to our work is guided inpainting by leveraging rich priors learned in large text-to-image diffusion models [55]. A seminal work [71] performs a reference image-guided inpainting instead of using the text prompts. Later, several works were proposed that advanced the idea of reference-guided inpainting with single [70, 73, 83] or multiple subject images [41] when given with an inpainting mask. In this work, we condition the inpainting with text along with the reference image without needing an inpainting mask.

Diffusion Models and personalization. Diffusion models have become state-of-the-art for image generation and enable several downstream editing applications [14, 27, 28, 61, 69]. These models scale extremely well when trained with large-scale image-text pairs for conditional generation [52, 53, 55, 58]. In contrast to GAN-based editing methods [29, 45–47, 59], which are limited to the data domains they are trained on, text-to-image models enable exceptional editing capabilities for in-the-wild real images [32, 43, 48] and enable various controls in the generation process [8, 27, 63, 81]. The text-to-image models can be personalized for a given subject with few input images. In this line, popular works like Dreambooth [56], Custom Diffusion [36] and Textual-Inversion [20] have proven effective in embedding concepts within text-to-image models with test-time tuning. Following this, there are several works [9, 21, 65, 76, 80, 87] that aim to improve personalization in terms of training efficiency and identity preservation. For simplicity, we use Textual inversion [20] for learning a given subject for subject-conditioned inpainting. To distill the rich representations learned from

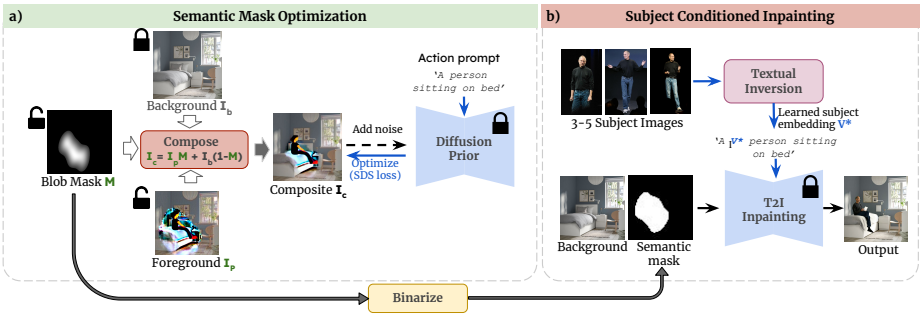


Fig. 2: Our approach consists of two stages: **a) Semantic Mask Optimization.** Given a background image $\mathcal{I}_b$, we initialize a blob mask $\mathcal{M}$ parameterized as Gaussian blobs and a foreground person image $\mathcal{I}_p$. These two images are combined to form a composite image $\mathcal{I}_c$, which is used to compute SDS loss with the action prompt. During optimization, only $\mathcal{M}$ and $\mathcal{I}_p$ are getting updated via $\mathcal{I}_c$. After training $\mathcal{M}$ converge to a plausible human placement region, which is then used for inpainting. **b) Subject conditioned inpainting.** Given a few subject images, we perform Textual Inversion to obtain its token embedding $\mathbf{V}*$. Next, we use the inpainting pipeline of T2I models to perform personalized inpainting of the subject.

text-to-image models to other domains, Score Distillation Sampling (SDS) is proposed. In the seminal work on dreamfusion [49], SDS loss is used to optimize NeRF [68] for text-to-3D generation. Later, several modifications to the original SDS loss are proposed [26,31,49] to improve high-frequency details. As our goal is to optimize semantic mask instead of generation, we use SDS loss as it works well in capturing low-level features.

3 Method

The goal of **Semantic Human Placement (SHP)** is to predict human affordances in a given background scene to place a given subject realistically. Our approach for SHP involves two stages, as shown in Fig. 2. Firstly, we learn human affordances defined as *semantic mask* in the background image with text guidance locating the regions where a person can be placed plausibly (Sec. 3.1). Next, we perform subject-conditioned inpainting using the *semantic mask* and a few subject images for a realistic scene composition (Sec. 3.2).

3.1 Semantic mask generation

Predicting semantic masks for objects (humans) that are not present in the scene is extremely challenging. The popular image description datasets only describe the objects present in the image. Existing methods [38,54] adapt these datasets by inpainting objects to create a paired dataset for object affordance and perform large-scale training to predict affordance. We take a radical approach to learning affordance using knowledge from T2I [55] models, *eliminating the need for data curation and expensive training*. T2I models have excellent object-scene compositional reasoning implicit in the generation process. An effective way

to distill this knowledge is to use Score Distillation Sampling (SDS) loss [49]. SDS and its variations have proven extremely effective for realistic text-to-3D generation by optimizing underlying 3D representation [10, 49, 62].

The success of SDS loss raises a natural question: *Can we use the SDS loss to learn the semantic mask for object placements?* If yes, *what could be an effective semantic mask representation for optimization?* Directly learning the semantic mask in the pixel space could easily lead to collapse covering all the image pixels (supplementary). We propose a novel *semantic mask* representation, parameterized as a set of interconnected Gaussian blobs. The proposed representation is extremely efficient, with only a few learnable parameters and fast rendering. Furthermore, it's highly compatible with learning through SDS loss.

Training. For effectively learning the parameterized *semantic mask* $\mathcal{M}$ on a background image $\mathcal{I}_b$, we create a framework involving a learnable foreground person image $\mathcal{I}_p$ initialized as a copy of $\mathcal{I}_b$. The major function of the $\mathcal{I}_p$ is to aid the learning of *semantic mask* parameters. Specifically, at each training iteration, we combine $\mathcal{I}_p$ and $\mathcal{I}_b$ using mask $\mathcal{M}$ to obtain a combined image $\mathcal{I}_c$ which will be used to compute SDS loss. The intuition is that as the training progresses, $\mathcal{I}_p$ will generate a person, and $\mathcal{M}$ will converge to the person's location and shape. This will, in turn, generate a person in $\mathcal{I}_c$ after composition, as shown in Fig. 3. This formulation enables the learning of semantic plausible regions for human placement in $\mathcal{M}$. To compute SDS loss, a noisy version of $\mathcal{I}_c$ is passed through the T2I model with an action prompt (e.g.,*'a person sitting on a bed'*). The gradients of the loss are computed with respect to $\mathcal{I}_c$ and are further backpropagated to compute gradients and update the image $\mathcal{I}_p$ and parameters of $\mathcal{M}$. Once trained, we binarize the optimized *semantic mask* $\mathcal{M}$ and use it for subject-conditioned inpainting on the background image. We note that the optimized mask is not pixel-perfect due to the Gaussian blob parameterization. However, the downstream inpainting based on T2I models expects a coarse semantic mask to allow room for adapting the foreground during inpainting Fig. 5. Further, this enables us to generate pose variations of the subjects and text-based editing of the placed person, which is not possible with a pixel-perfect mask Fig. 11.

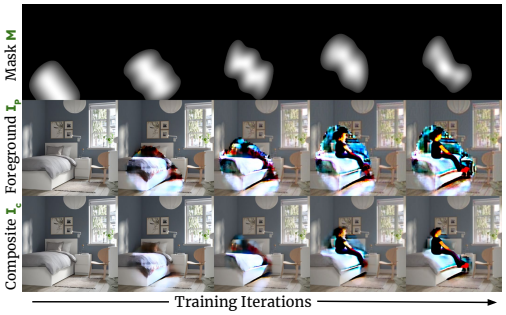


Fig. 3: Training progression, of semantic mask $\mathcal{M}$ and person image $\mathcal{I}_p$ with SDS loss with prompt *'A person sitting on bed'*. $\mathcal{M}$ starts to converge at the appropriate person region following prompt.

Mask parametrization. For effectively learning of *semantic mask* using SDS loss, we parameterize $\mathcal{M}$ as a set of $\mathbf{K}$ interconnected Gaussian blobs (Fig. 4). Following [16], we use simple ellipsoid representation for blobs by defining

center location $\mathbf{x} \in [0,1]^2$, scale $\mathbf{s} \in \mathbb{R}$, aspect ratio $\mathbf{a} \in \mathbb{R}$, rotation angle $\theta \in [-\pi/2, \pi/2]$. We tie the blobs by keeping a fixed distance $\mathbf{r} \in \mathbb{R}$ between the center of consecutive blobs. Precisely, the center location $\mathbf{x}_i$ for the i^{th} $(i > 1)$ blob is computed using the previous blob center $\mathbf{x}_{i-1}$ using Eq. 1.

$$\mathbf{x}_i = \mathbf{x}_{i-1} + \begin{bmatrix} \mathbf{r} \cdot \cos(\alpha_i) \\ \mathbf{r} \cdot \sin(\alpha_i) \end{bmatrix} \quad (1)$$

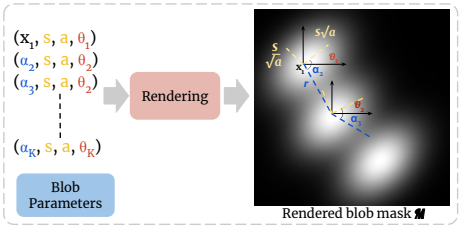


Fig. 4: Blob mask parameterization.

where $\alpha_i \in \mathbb{R}$ is the relative angle between the blob centers from $\mathbf{x}$-axis. Finally, the total blob parameters defining *semantic mask* are the following:

$$\gamma_1 \equiv (x_1, s, a, \theta_1) \quad and \quad \gamma_i \equiv (\alpha_i, s, a, \theta_i) \quad for\ i \in \{2, K\} \quad (2)$$

Next, we use parameters γ_i to create a mask image $\mathcal{M}_i \in [0,1]^{H \times W}$ for i^{th} blob. We perform this operation in a differentiable manner to enable optimization of the parameters. Specifically, we compute the Mahanobolis distance between every pixel coordinate location $\mathbf{x}_{grid}$ from the i^{th} blob center $(\mathbf{x}_i)$ to assign the intensity value at $\mathbf{x}_{grid}$

$$D^m(\mathbf{x}_{\text{grid}}, \mathbf{x}_i) = (\mathbf{x}_{\text{grid}} - \mathbf{x}_i)^T (R \cdot \Sigma \cdot R^T)^{-1} (\mathbf{x}_{\text{grid}} - \mathbf{x}_i) \quad (3)$$

$$\mathcal{M}_i[\mathbf{x}_{grid}] = \exp(-0.5 * D^m(\mathbf{x}_{grid}, \mathbf{x}_i)) \quad (4)$$

where $\Sigma = c. \begin{bmatrix} {\delta_x}^2 & 0 \\ 0 & {\delta_y}^2 \end{bmatrix}$, R is the rotation matrix based on θ_i , $c = 0.02$ controls the sharpness of each blob and $\delta_x = s/\sqrt{a}$, and $\delta_y = s\sqrt{a}$. Finally, we obtain aggregated *semantic mask* $\mathcal{M}$ by taking the pixel-wise mean of the individual blob masks $\mathcal{M}_i$. During training, we only optimize the first blob center location $\mathbf{x}_1$, all rotation angles θ_i, and relative angles α_i. This provides enough flexibility for the mask to reach the appropriate location and adjust based on the required human pose. Empirically, we have observed fixing $\mathbf{s}$, $\mathbf{a}$, and $\mathbf{r}$ to a constant for all the blobs results in better convergence of the mask (suppl. material).

3.2 Subject conditioned inpainting

Given a few unposed images of a subject, we aim to place them in the background image following affordances using the optimized *semantic mask* $\mathcal{M}$. Inpainting humans is extremely hard, as the person's pose should be appropriately changed based on the location of the mask (e.g. if we are inpainting a person on a chair, a sitting pose is more suitable). Naively, following a reference-based inpainting [71] does not consider these person-scene affordances and results in unnatural compositions (Fig. 9). We leverage rich human-scene priors from the T2I [55] models to perform plausible inpainting of humans respecting the local affordances. The text-guided inpainting pipeline of T2I models generates plausible outputs adapting the person's pose according to the background. We ask: *How can we adapt*

this inpainting pipeline for subject conditioned inpainting? We propose to pass the subject-specific knowledge in T2I through the input text conditioning used for inpainting (Fig. 2b)). Specifically, we learn a token embedding $\mathbf{V}*$ representing our subject from a few input images using Textual Inversion [20]. Next, we use $\mathbf{V}*$ along with the inpainting prompt (e.g., *'A* $\mathbf{V}*$ person sitting on a sofa') to condition the inpainting pipeline of T2I model. This simplistic framework for conditioning generates realistic human placements that follow the scene affordances, capitalizing on the rich object-scene priors from the T2I models (Fig. 2). Further, our inpainting can benefit from improved text-based personalization methods in T2I models.

How accurate should the semantic mask be? We analyze the sensitivity of our inpainting pipeline on the preciseness of *semantic mask* $\mathcal{M}$ in Fig. 5. Given a background image, we first run our method to place a person in the scene to obtain source image. Next, we compute the segmentation mask of the person using SAM [33]. The obtained person mask is compatible with the background and can be used for inpainting. We ablate over the coarseness of the mask by iteratively dilating the segmentation mask and performing inpainting. Using the original mask for inpainting is too stringent and does not place the person as the T2I inpainting pipeline is stochastic. On increasing the dilation, inpainting is successful but lead to background distortions. We conclude that using a pixel-perfect *semantic mask* hurts inpainting and a coarse mask is needed for realistic inpainting results. This further motivates our blob parameterization which creates a coarse mask *semantic mask* for placing humans.

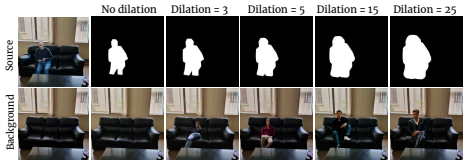


Fig. 5: Pixel-perfect mask is not effective for T2I inpainting.

4 Experiments

We perform extensive experiments to evaluate our method and provide detailed ablations. We use Stable Diffusion models as our representative T2I models in all the experiments. Given a background image, we perform test time optimization to obtain the semantic mask for human placement. Specifically, we optimize the blob parameters for the mask for 1000 iterations using SDS loss [49] with a guidance scale of 200. Next, we perform subject-conditioned inpainting by performing textual inversion with $3-5$ subject images. Further details about implementation are provided in the supplementary. In this section, we first discuss datasets and metrics (Sec. 4.1) followed by experiments on placing persons (Sec. 4.2), downstream applications (Sec. 4.4) and finally ablation studies (Sec. 4.5).

4.1 Dataset and metrics

Dataset. For evaluation, we collected a dataset consisting of 30 in-the-wild indoor and outdoor background images from the web, along with a few manually

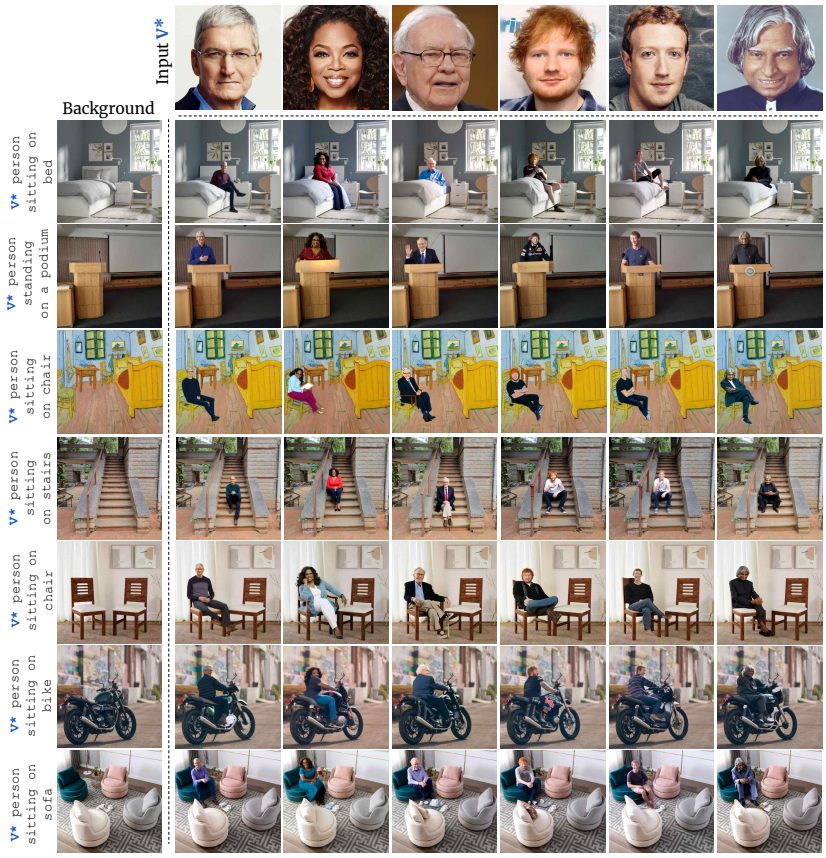


Fig. 6: Realistic human placement in diverse indoor and outdoor scenes.

captured images in a university. For personalization, we select 15 celebrity images, including scientists, actors, sports persons, and tech executives, and collect $4-5$ images for each subject. For placement, we use action-based text prompts such as *'a person sitting on sofa'*, based on the background image. Additional details about the prompts are provided in the supplementary.

Metrics. We evaluate realistic human placement on the following three aspects:

a) Text Alignment - The generated image after placing a human should follow the given action prompt for desired text-based placement. We compute the cosine similarity between the CLIP [51] image embedding of the generated image and CLIP text embedding of the provided prompt. Higher CLIP similarity (**CLIP-sim**) signifies that human placement follows the given prompt.

b) Person generation - Inpainting model fails to generate a person if the inpainting mask is incorrectly placed (e.g., a person sitting on a window). To quantify correct person generation during inpainting, we detect humans in the generated images by applying SAM [33] on the output and checking for an instance of the 'person' class. We aggregate this metric as the percentage of images where a person is detected (**% Person**), noting the accuracy of placement.



Fig. 7: Our method enables several interfaces for realistic human placement.

c) Background Preservation - To quantify the background preservation, we first segment out the human from the generated output using SAM [33] and compute LPIPS [82] in the *non-human* regions between the background and inpainted image. For perfect placement, the background should be preserved during inpainting with lower LPIPS scores.

We want to highlight that there is an inherent tradeoff between these metrics: Higher CLIP-sim and % Person could be attained by using a large inpainting mask, but this results in significant background distortions (higher LPIPS). Hence, we aim to balance these metrics well for desired results.

4.2 Placing humans

We present our qualitative results for Semantic Human Placement in Fig. 6. Our method generates highly realistic placement results while preserving the subject's identity and background. The obtained semantic masks properly localize the region in the background corresponding to the given action prompt. Also, observe the mask shape changes according to the desired human pose, which is compatible with the background (e.g., sitting pose in backgrounds 1 and 3 and bike riding pose in background 6). Further, the proposed Text2Place enables diverse human placement interfaces as shown in Fig. 7.

Placing multiple persons sequentially (Fig. 7a)) can be easily done by iteratively optimizing for two semantic masks. While optimizing for the semantic mask for the second subject, we add additional loss to minimize the overlap between the two masks. This helps generate disjointed semantic masks that are needed to place multiple persons.

Diverse placement with text guidance, (Fig. 7b)) is achieved by providing different location prompts during semantic mask optimization.

Table 1: Baseline comparison

Method	LPIPS ↓	CLIP-sim ↑	% Person ↑
GracoNet [86]	0.1090	0.2601	53.48
TopNet [88]	0.1162	0.2617	67.3
LLaVA [39]	0.1296	0.2501	20.91
GPT4V [1]	0.1059	0.2615	64.18
Ours (center)	**0.0845**	0.2613	55.52
Ours	0.0934	**0.2726**	**88.55**

Placing kids. Our notable blob mask representation enables the placement of smaller subjects by initializing with a relatively small blob size. This enables the realistic placement of kids in diverse scenes with background preservation. (Fig. 7c).

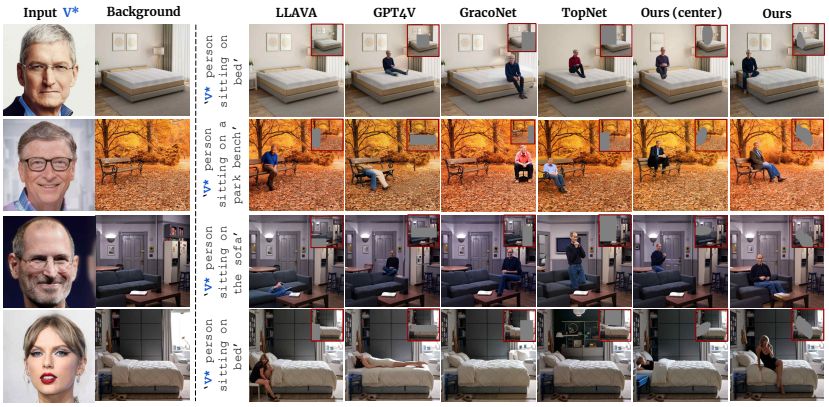


Fig. 8: Baseline Comparison: Notably, LLaVa generates masks with semantically incorrect locations, while GPT4v produces excessively large mask sizes. GracoNet and TopNet predict bounding boxes at the correct location but of inaccurate size, whereas our method accurately determines the optimal location and size for person insertion.

4.3 Comparisons

Human placement comparisons. We compare our approach against the following baselines: Object placement baselines **1) TopNet** [89] and **2) GracoNet** [86] trained on large-scale object placement dataset OPA [40]. The OPA dataset contains significant human placement examples, making these baselines suitable for comparison. Additionally, we compare with state-of-the-art VLMs - **3) GPT4V [1]** and **4) LLAVA** [39]. These models have excellent multimodal reasoning capabilities. For our experiments, we provide them with a background image and a prompt *'predict a bounding box location for placing a human on the chair'*. From all the above baselines, the obtained bounding box coordinates are then used to create a semantic mask for inpainting in the next stage. To analyze the importance of the shape of the semantic mask vs the location, we compare it against baseline - **5) Ours (center)**, where we parameterize the semantic mask as a single blob with a learnable center. The scale s is fixed to be 0.9, and the orientation angle is randomly selected. The comparison results are present in Fig. 8 and Tab. 1.

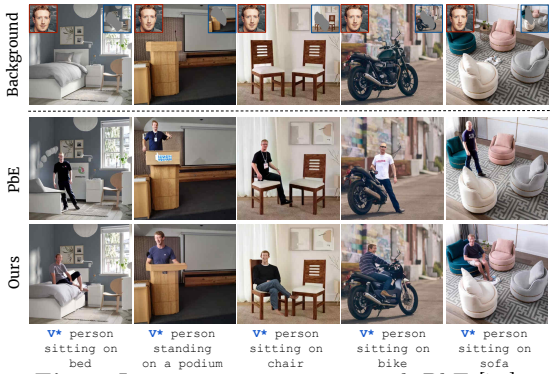


Fig. 9: Inpainting comparison with PbE [71]

Analysis. GracoNet and TopNet are able to localize the bounding box correctly but struggle with generating the appropriate size of the box, resulting in unnatural placements. Likewise, in ours (center), the mask converges to the appropriate location for placement; however, it cannot place the person following local affordance due to the inappropriate shape of the semantic mask. This underscores the

importance of semantic mask shape for forming realistic poses during placement. Bounding box masks generated by LLAVA and GPT4V are not able to cover the appropriate region to place a human. LLAVA mostly generates boxes on the left or bottom of the background image, resulting in no human placement or placement with unnatural poses. GPT4V generates masks at appropriate locations but fails to capture the shape needed to place humans. This results in unnatural human placement results. The same is evident from quantitative results, where GPT4V achieves the second-highest %-Person score in comparison. Our method generates a person in most cases as both the semantic mask has an appropriate shape and location that follows the action prompts. Quantitatively, ours (center) achieves the best LPIPS score, suggesting the best background preservation; however, it has a lower CLIP-sim than ours, suggesting that the person is not appropriately placed following the text.

Inpainting comparison. We compare our inpainting method with a state-of-the-art reference-based inpainting method Paint-by-example (PbE) [71] in Fig. 9. PbE is trained to inpaint a reference subject image given with an inpainting mask. For comparison, we use the semantic mask obtained from our method for both the inpainting methods. As PbE does not allow for text conditioning, it mostly generates a standing person. Moreover, the identity and the pose of the placed subject are inconsistent in general-purpose PbE.

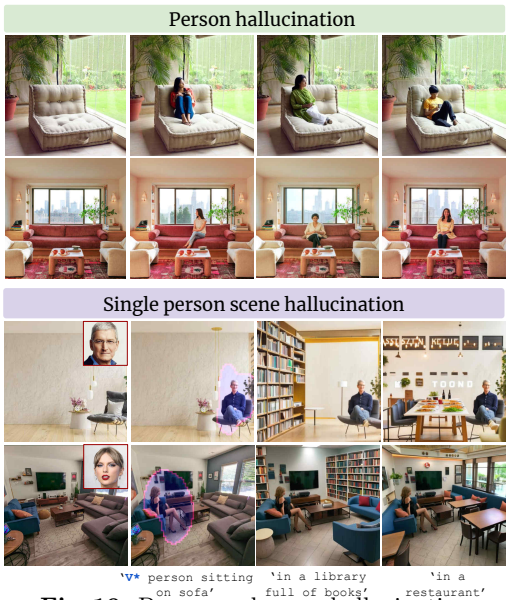


Fig. 10: Person and scene hallucination

4.4 Applications

Person hallucination. Given a background scene, we can hallucinate new persons by passing action prompts without subject conditioning (e.g., *'a person sitting on sofa'*) (Fig. 10). Our inpainting pipeline generates realistic outputs with humans in diverse poses consistent with the background, follows the text prompts, and preserves the subject's identity.

Scene hallucination. Another interesting application is to conditionally generate a scene compatible with the given pose of the person (Fig. 10). To this end, we first place a human subject in the background using the predicted semantic mask. Next, we invert the semantic mask and perform outpainting of the region covered by the subject using the same T2I-based inpainting pipeline. We can conditionally generate diverse scenes by providing text prompts specifying the background (e.g., '*a library full of books*').

Applications of subject conditioned inpainting. The proposed pipeline for inpainting enables various downstream applications to control the generated image, as shown in Fig. 11.

• *Firstly*, we can perform text-based editing of the generated person by changing the text prompt for inpainting (e.g., *'person -> person wearing a hat'*). As the inpainting is bound to modify only the semantic mask region, we achieve highly disentangled and localized editing of the generated subject.

• *Secondly*, as we are using T2I diffusion model-based inpainting, which is inherently stochastic, we can generate diverse *plausible* poses of the person from the same semantic mask and the action prompt. Note that our mask parameterization as a set of blobs allows us to generate diverse subject poses.

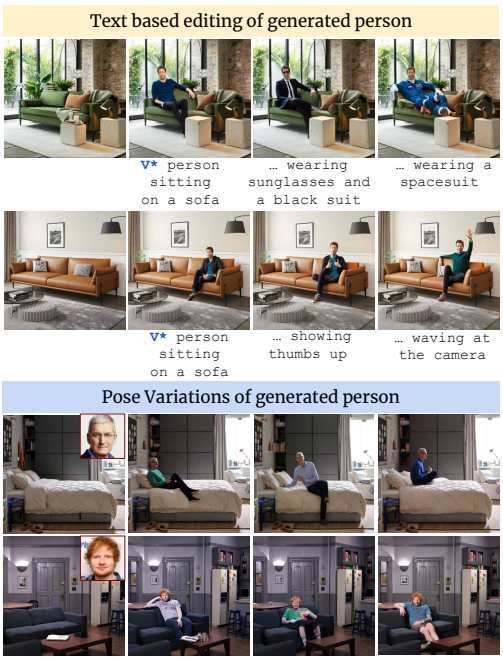


Fig. 11: Applications of inpainting

Additional Conditionings. Our method can be integrated seamlessly with other conditioning approaches such as pose-conditioned ControlNet [81] and IP-adapter [75] for explicit conditioning on face identity (Fig. 12).

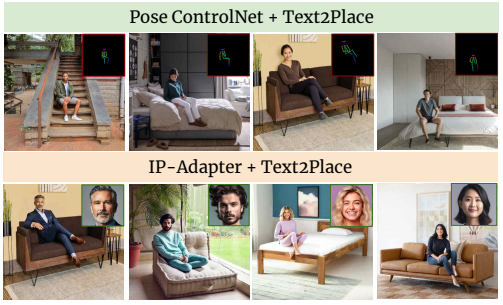


Fig. 12: Additional Conditioning

Placing objects beyond humans. perform these semantic edits requiring a *loose* semantic mask, which would be challenging with a pixel-perfect semantic mask. Our method can easily be adapted to generate placements of various objects (Fig. 13). The obtained placements look highly realistic, with accurate placement of objects showing our approach's generalization without additional training.

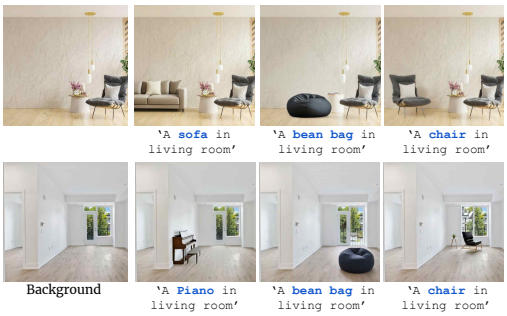


Fig. 13: Placing objects beyond humans.

4.5 Ablations

We ablate over the following blob parameters. Qualitative results for ablations are provided in the supplementary.

Scale of blob (s). is a crucial hyperparameter that controls the size of the blob; having a large **s** value results in significant background change, and a small **s** restricts the full body generation during the inpainting task. Empirically, we found **s** =0.6 gives the best results from Table 2a).

Number of blobs. Using a single blob doesn't give freedom to generate the full body image of a person in different poses, so we use n number of blobs and fix the distance between each adjacent blob as discussed in Section 3.3, and the number n is a hyperparameter in our method, having fewer blobs restrict the mask diversity, but more blobs can cause bigger masks, leading to more background change. From Table 2b), we can see that n=5 gives a good balance between diversity and background change.

Table 2: Ablation of blob parameters.

(a) Ablation over blob scale **s**

Scale	LPIPS ↓	CLIP-sim ↑	% Person ↑
0.3	0.0537	0.2594	41.1
0.4	0.0806	0.2663	69.0
0.5	0.0858	0.2712	81.5
0.6	**0.0904**	**0.2736**	**90.6**
0.7	0.1074	0.2729	96.0

(b) Ablation over # blobs

#blobs	LPIPS ↓	CLIP-sim ↑	% Person ↑
1	0.1318	0.2780	93.0
3	0.1305	0.2797	94.9
5	0.0904	0.2736	90.6
7	0.0780	0.2749	75.0

5 Discussion

Limitations. While our approach excels in many aspects, it does have limitations (Fig. 14. Notably, it may not effectively handle the placement of relatively small objects in a scene. This is because our blob mask representation occupies a significant image region. Using larger masks for inpainting results in changing the background or placing objects of inappropriate scale. Furthermore, our method lacks precise control over the generated person's poses, as tighter masks may hinder inpainting using T2I models shown in Fig. 5.

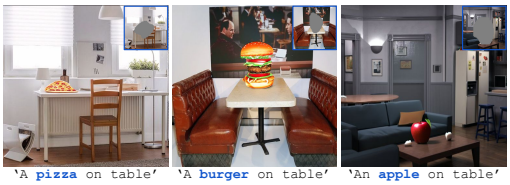


Fig. 14: Failure cases.

Conclusion. In this work, we propose the first method to learn text-conditioned human affordances for realistic human placement for in-the-wild scenes. We leverage rich priors learned in text-to-image (T2I) diffusion models to optimize plausible *semantic masks* for human placement. The obtained masks are then used to perform subject-conditioned inpainting to place a given subject. Semantic masks are parameterized with a novel blob representation that adapts to the human pose during optimization with score distillation sampling in T2I models. The proposed method enables several downstream applications, including human and scene hallucination, placing multiple persons in a scene, and text-based editing of the generated subject. In summary, we propose a robust approach to perform text-guided placement of realistic humans in scenes without large pertaining.

Acknowledgements. We thank Tejan Karmali for regular discussions and Abhijnya Bhat for reviewing the draft and providing helpful feedback. This work was partly supported by PMRF from Govt. of India (Rishubh Parihar) and Kotak IISc AI-ML Centre.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Ballester, C., Bertalmio, M., Caselles, V., Sapiro, G., Verdera, J.: Filling-in by joint interpolation of vector fields and gray levels. IEEE transactions on image processing **10**(8), 1200–1211 (2001)
3. Bar, A., Gandelsman, Y., Darrell, T., Globerson, A., Efros, A.: Visual prompting via image inpainting. Advances in Neural Information Processing Systems **35**, 25005–25017 (2022)
4. Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B.: Patchmatch: A randomized correspondence algorithm for structural image editing. ACM Trans. Graph. **28**(3), 24 (2009)
5. Brooks, T., Efros, A.A.: Hallucinating pose-compatible scenes. In: European Conference on Computer Vision. pp. 510–528. Springer (2022)
6. Cao, Z., Gao, H., Mangalam, K., Cai, Q.Z., Vo, M., Malik, J.: Long-term human motion prediction with scene context. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16. pp. 387–404. Springer (2020)
7. Cao, Z., Radosavovic, I., Kanazawa, A., Malik, J.: Reconstructing hand-object interactions in the wild. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12417–12426 (2021)
8. Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., Cohen-Or, D.: Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. ACM Transactions on Graphics (TOG) **42**(4), 1–10 (2023)
9. Chen, X., Huang, L., Liu, Y., Shen, Y., Zhao, D., Zhao, H.: Anydoor: Zero-shot object-level image customization. arXiv preprint arXiv:2307.09481 (2023)
10. Chen, Z., Wang, F., Liu, H.: Text-to-3d using gaussian splatting. ArXiv (2023)
11. Chuang, C.Y., Li, J., Torralba, A., Fidler, S.: Learning to act properly: Predicting and explaining affordances from images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 975–983 (2018)
12. Delaitre, V., Fouhey, D.F., Laptev, I., Sivic, J., Gupta, A., Efros, A.A.: Scene semantics from long-term observation of people. In: Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12. pp. 284–298. Springer (2012)
13. Delaitre, V., Fouhey, D.F., Laptev, I., Sivic, J., Gupta, A., Efros, A.A.: Scene semantics from long-term observation of people. In: Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12. pp. 284–298. Springer (2012)
14. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in Neural Information Processing Systems **34**, 8780–8794 (2021)
15. Efros, A.A., Leung, T.K.: Texture synthesis by non-parametric sampling. In: Proceedings of the seventh IEEE international conference on computer vision. vol. 2, pp. 1033–1038. IEEE (1999)

16. Epstein, D., Park, T., Zhang, R., Shechtman, E., Efros, A.A.: Blobgan: Spatially disentangled scene representations. In: European Conference on Computer Vision. pp. 616–635. Springer (2022)
17. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021)
18. Fouhey, D.F., Wang, X., Gupta, A.: In defense of the direct perception of affordances. arXiv preprint arXiv:1505.01085 (2015)
19. Fouhey, D.F., Wang, X., Gupta, A.: In defense of the direct perception of affordances. arXiv preprint arXiv:1505.01085 (2015)
20. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)
21. Gal, R., Arar, M., Atzmon, Y., Bermano, A.H., Chechik, G., Cohen-Or, D.: Encoder-based domain tuning for fast personalization of text-to-image models. ACM Transactions on Graphics (TOG) **42**(4), 1–13 (2023)
22. Gibson, J.J.: The ecological approach to the visual perception of pictures. Leonardo **11**(3), 227–235 (1978)
23. Gkioxari, G., Girshick, R., Dollár, P., He, K.: Detecting and recognizing human-object interactions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8359–8367 (2018)
24. Grabner, H., Gall, J., Van Gool, L.: What makes a chair a chair? In: CVPR 2011. pp. 1529–1536. IEEE (2011)
25. Gupta, A., Satkin, S., Efros, A.A., Hebert, M.: From 3d scene geometry to human workspace. In: CVPR 2011. pp. 1961–1968. IEEE (2011)
26. Hertz, A., Aberman, K., Cohen-Or, D.: Delta denoising score. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2328–2337 (2023)
27. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-or, D.: Prompt-to-prompt image editing with cross-attention control. In: The Eleventh International Conference on Learning Representations (2022)
28. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)
29. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)
30. Jiang, Y., Koppula, H., Saxena, A.: Hallucinated humans as the hidden context for labeling 3d scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2993–3000 (2013)
31. Katzir, O., Patashnik, O., Cohen-Or, D., Lischinski, D.: Noise-free score distillation. arXiv preprint arXiv:2310.17590 (2023)
32. Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6007–6017 (2023)
33. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
34. Koppula, H.S., Gupta, R., Saxena, A.: Learning human activities and object affordances from rgb-d videos. The International journal of robotics research **32**(8), 951–970 (2013)

35. Kulal, S., Brooks, T., Aiken, A., Wu, J., Yang, J., Lu, J., Efros, A.A., Singh, K.K.: 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)
36. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1931–1941 (2023)
37. Lee, J., Chai, J., Reitsma, P.S., Hodgins, J.K., Pollard, N.S.: Interactive control of avatars animated with human motion data. In: Proceedings of the 29th annual conference on Computer graphics and interactive techniques. pp. 491–500 (2002)
38. Li, X., Liu, S., Kim, K., Wang, X., Yang, M.H., Kautz, J.: Putting humans in a scene: Learning affordance in 3d indoor environments. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12368–12376 (2019)
39. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744 (2023)
40. Liu, L., Liu, Z., Zhang, B., Li, J., Niu, L., Liu, Q., Zhang, L.: Opa: object placement assessment dataset. arXiv preprint arXiv:2107.01889 (2021)
41. Lu, L., Zhang, B., Niu, L.: Dreamcom: Finetuning text-guided inpainting model for image composition. arXiv preprint arXiv:2309.15508 (2023)
42. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11461–11471 (2022)
43. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. arXiv preprint arXiv:2108.01073 (2021)
44. Osher, S., Burger, M., Goldfarb, D., Xu, J., Yin, W.: An iterative regularization method for total variation-based image restoration. Multiscale Modeling & Simulation **4**(2), 460–489 (2005)
45. Parihar, R., Dhiman, A., Karmali, T., Babu, R.V.: Everything is there in latent space: Attribute editing and attribute style manipulation by stylegan latent space exploration. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 1828–1836 (2022)
46. Parihar, R., Magazine, R., Tiwari, P., Babu, R.V.: We never go out of style: Motion disentanglement by subspace decomposition of latent space. arXiv preprint arXiv:2306.00559 (2023)
47. Park, T., Zhu, J.Y., Wang, O., Lu, J., Shechtman, E., Efros, A., Zhang, R.: Swapping autoencoder for deep image manipulation. Advances in Neural Information Processing Systems **33**, 7198–7211 (2020)
48. Parmar, G., Kumar Singh, K., Zhang, R., Li, Y., Lu, J., Zhu, J.Y.: Zero-shot image-to-image translation. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–11 (2023)
49. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022)
50. Quadflieg, S., Koldewyn, K.: The neuroscience of people watching: how the human brain makes sense of other people's encounters. Annals of the New York Academy of Sciences **1396**(1), 166–182 (2017)
51. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from

natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
52. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)
53. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International Conference on Machine Learning. pp. 8821–8831. PMLR (2021)
54. Ramrakhya, R., Kembhavi, A., Batra, D., Kira, Z., Zeng, K.H., Weihs, L.: Seeing the unseen: Visual common sense for semantic placement. arXiv preprint arXiv:2401.07770 (2024)
55. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)
56. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: 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)
57. Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., Norouzi, M.: Palette: Image-to-image diffusion models. In: ACM SIGGRAPH 2022 Conference Proceedings. pp. 1–10 (2022)
58. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022)
59. Shen, Y., Gu, J., Tang, X., Zhou, B.: Interpreting the latent space of gans for semantic face editing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9243–9252 (2020)
60. Singh, J., Zhang, J., Liu, Q., Smith, C., Lin, Z., Zheng, L.: Smartmask: Context aware high-fidelity mask generation for fine-grained object insertion and layout control. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6497–6506 (2024)
61. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
62. Tang, J., Ren, J., Zhou, H., Liu, Z., Zeng, G.: Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. arXiv preprint arXiv:2309.16653 (2023)
63. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1921–1930 (2023)
64. Wang, J., Xu, H., Xu, J., Liu, S., Wang, X.: Synthesizing long-term 3d human motion and interaction in 3d scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9401–9411 (2021)
65. Wang, Q., Bai, X., Wang, H., Qin, Z., Chen, A.: Instantid: Zero-shot identity-preserving generation in seconds. ArXiv (2024)
66. Wang, X., Girdhar, R., Gupta, A.: Binge watching: Scaling affordance learning from sitcoms. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2596–2605 (2017)
67. Wang, X., Girdhar, R., Gupta, A.: Binge watching: Scaling affordance learning from sitcoms. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2596–2605 (2017)

68. Wang, Z., Wu, S., Xie, W., Chen, M., Prisacariu, V.A.: NeRF−−: Neural radiance fields without known camera parameters. arXiv preprint arXiv:2102.07064 (2021)
69. Wu, C.H., De la Torre, F.: Unifying diffusion models' latent space, with applications to cyclediffusion and guidance. arXiv preprint arXiv:2210.05559 (2022)
70. Xie, S., Zhao, Y., Xiao, Z., Chan, K.C., Li, Y., Xu, Y., Zhang, K., Hou, T.: Dreaminpainter: Text-guided subject-driven image inpainting with diffusion models. arXiv preprint arXiv:2312.03771 (2023)
71. Yang, B., Gu, S., Zhang, B., Zhang, T., Chen, X., Sun, X., Chen, D., Wen, F.: Paint by example: Exemplar-based image editing with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18381–18391 (2023)
72. Yang, C., Lu, X., Lin, Z., Shechtman, E., Wang, O., Li, H.: High-resolution image inpainting using multi-scale neural patch synthesis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6721–6729 (2017)
73. Yang, S., Chen, X., Liao, J.: Uni-paint: A unified framework for multimodal image inpainting with pretrained diffusion model. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 3190–3199 (2023)
74. Yao, B., Fei-Fei, L.: Modeling mutual context of object and human pose in human-object interaction activities. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 17–24. IEEE (2010)
75. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)
76. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. ArXiv **abs/2308.06721** (2023)
77. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Generative image inpainting with contextual attention. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5505–5514 (2018)
78. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Free-form image inpainting with gated convolution. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4471–4480 (2019)
79. Yu, Y., Zhan, F., Wu, R., Pan, J., Cui, K., Lu, S., Ma, F., Xie, X., Miao, C.: Diverse image inpainting with bidirectional and autoregressive transformers. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 69–78 (2021)
80. Yuan, G., Cun, X., Zhang, Y., Li, M., Qi, C., Wang, X., Shan, Y., Zheng, H.: Inserting anybody in diffusion models via celeb basis. arXiv preprint arXiv:2306.00926 (2023)
81. Zhang, L., Agrawala, M.: Adding conditional control to text-to-image diffusion models pp. 3836–3847 (2023)
82. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
83. Zhang, X., Guo, J., Yoo, P., Matsuo, Y., Iwasawa, Y.: Paste, inpaint and harmonize via denoising: Subject-driven image editing with pre-trained diffusion model. arXiv preprint arXiv:2306.07596 (2023)
84. Zhao, S., Cui, J., Sheng, Y., Dong, Y., Liang, X., Chang, E.I., Xu, Y.: Large scale image completion via co-modulated generative adversarial networks. arXiv preprint arXiv:2103.10428 (2021)

85. Zheng, H., Lin, Z., Lu, J., Cohen, S., Shechtman, E., Barnes, C., Zhang, J., Xu, N., Amirghodsi, S., Luo, J.: Image inpainting with cascaded modulation gan and object-aware training. In: European Conference on Computer Vision. pp. 277–296. Springer (2022)
86. Zhou, S., Liu, L., Niu, L., Zhang, L.: Learning object placement via dual-path graph completion. In: European Conference on Computer Vision. pp. 373–389. Springer (2022)
87. Zhou, Y., Zhang, R., Sun, T., Xu, J.: Enhancing detail preservation for customized text-to-image generation: A regularization-free approach. arXiv preprint arXiv:2305.13579 (2023)
88. Zhu, S., Lin, Z., Cohen, S., Kuen, J., Zhang, Z., Chen, C.: Topnet: Transformer-based object placement network for image compositing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1838–1847 (2023)
89. Zhu, S., Lin, Z., Cohen, S., Kuen, J., Zhang, Z., Chen, C.: Topnet: Transformer-based object placement network for image compositing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1838–1847 (2023)
90. Zhu, Y., Fathi, A., Fei-Fei, L.: Reasoning about object affordances in a knowledge base representation. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II 13. pp. 408–424. Springer (2014)