Text2Place: Affordance-aware Text Guided Human Placement - Supplementary Document

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

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

Table of Contents

A Ethics statement

We acknowledge that our method can be used to generate images for malicious purposes, but similar to Stable Diffusion, the samples generated by our model can be watermarked. Since our model can be used to generate novel compositions for realistic human placement, it can be used for misleading generations. However, there's a line of research for detecting fake samples from generative models, which we support. We believe the research contributions of this work outweigh the negative impacts.

B Implementation details

We use Stable Diffusion models [5] as a representative T2I model in all of our experiments. We use Stable Diffusion-v1.4 to compute SDS loss (following [**?**]) for the semantic mask generation and Stable Diffusion-XL [5] for the subject conditioned inpainting task due to its superior inpainting capabilities. We perform instance-specific training for the background image to obtain a *semantic mask*. Specifically, we optimize the blob parameters and the foreground image for 1000 iterations using SDS loss [3] with a guidance scale of 200. The learning rate for the foreground image is 0.2, and blob parameters 0.1. We use an AdamW optimizer with no weight decay and a cosine scheduler. For other fixed blob parameters, we set the distance between blobs $r = 0.01$, aspect ratio $a = 2$,

and blob scale $s = 0.6$. Finally, we binarize the soft blob mask to obtain a semantic mask with a threshold value of 0.2. We use standard hyperparameters from hugging face [2, 4] for textual inversion use.

C Analysis on *semantic mask* parameterization.

In this section, we analyze the design choices for parameterization of the *semantic mask* as blobs. We visualize the progression of the semantic mask, foreground image, and the combined image during training. **a) Pixel-wise mask**: We first try a naive baseline where the semantic mask is parameterized as a learnable gray-scale image. This simple parameterization results in a collapse of the mask image, resulting in the white pixels spread over the full image. This obtained mask is not useful for inpainting and cannot generate a person. *semantic masks* resulting in natural placement outputs.
b) Next, we tried parameterizing a mask as a set of Gaussian blobs without tying their centers. During training, the locations of the blobs are updated independently of each other. After convergence, they do not occupy nearby regions to create a continuous *semantic mask* that captures an appropriate human pose. This mask generates the person but changes the background in an unnatural pose. This demonstrates the significance of our design decision to maintain the linkage between the blob centers so that the blobs as a whole can form appropriate human poses and generate plausible semantic masks to place humans.
c) We also tried another baseline approach, where all the blob parameters **s**, **a**, x α, and θ are learnable. This results in generating masks that have a large foreground region, including large background regions. This is primarily because the blob's scale **s** and aspect ratio **a** can increase without any regularization to minimize the SDS loss on the combined image. Hence, when used as a semantic mask for the next stage of inpainting, a person with a relatively large scale is generated, which is inconsistent with the background scene or suffers from significant background distortions. This suggests the importance of keeping some of the blob parameters, like aspect ratio and scale, fixed during optimization to obtain plausible masks for realistic placement.

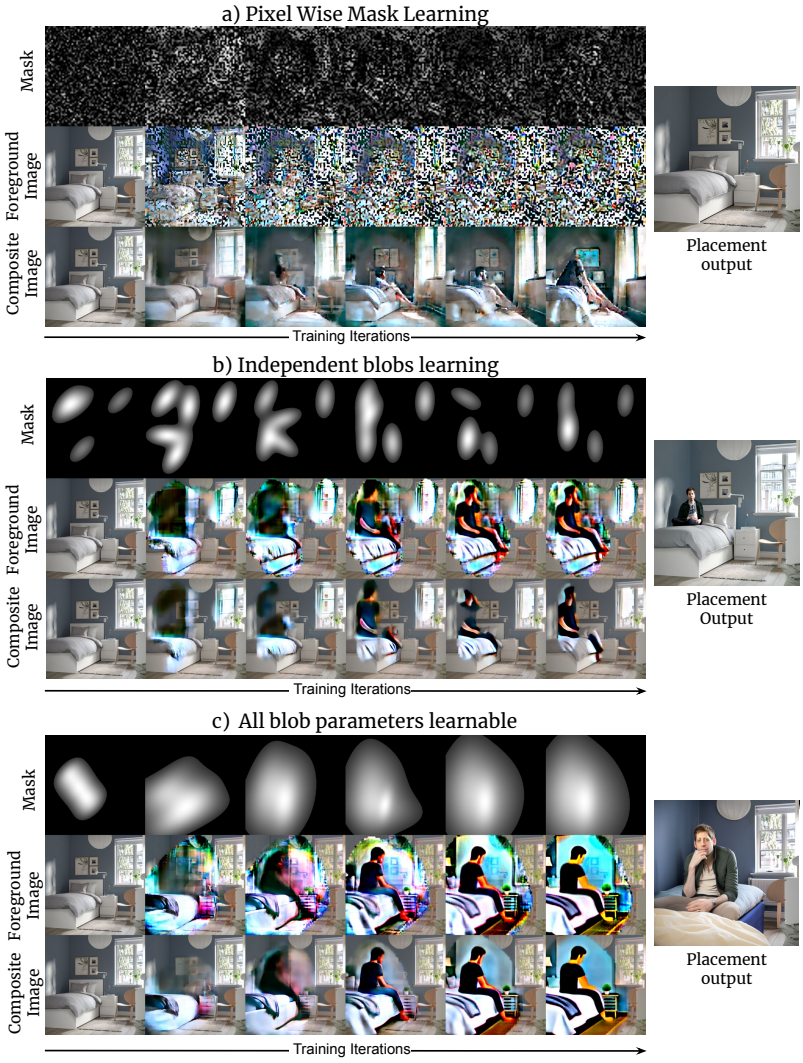

Fig. 1: Ablation over different parameterization for *semantic masks*.

D Ablations

D.1 Semantic mask generation

In this section, we ablate over different design choices for parameterization of *semantic mask*.

Number of blobs. We ablate over the number of Gaussian blobs used to parameterize the *semantic mask* in Table.3b) (main paper). Here, we present qualitative results for the same experiment in Fig. 2. Using less number of blobs results in a crude semantic mask and doesn't capture the required human pose for accurate placement. This results in significant background changes and changes in the person's pose. Having a very high number with a relatively smaller scale and trying to overfit too much to the person's pose. However, we have observed that the inpainting pipeline works best when we have somewhat loose *semantic masks*, we found $n = 5$ achieves the perfect trade-off. We decide the scale based on the number of blobs to give enough space to the mask for generating humans. Specifically, we keep scale values $\{3.0, 1.0, 0.6, 0.43\}$ for number of blobs $\{1, 3, 5, 7\}$ respectively.

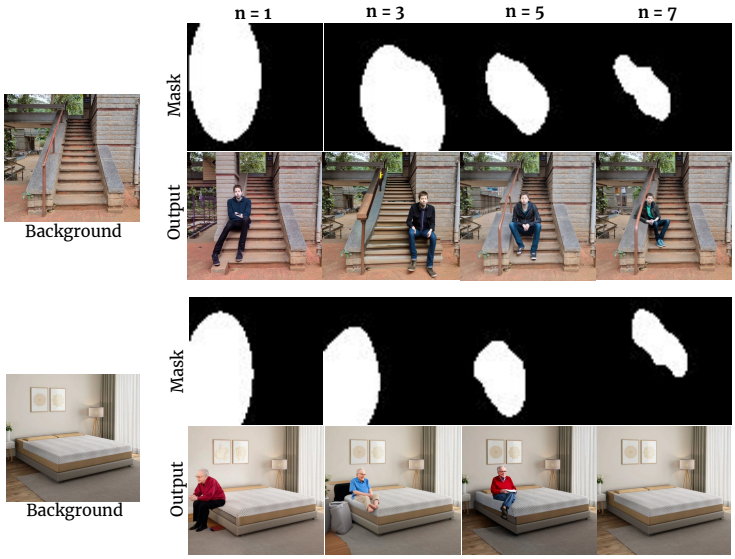

Fig. 2: Ablation over number of blobs used to parameterize *semantic mask*.

Scale of the blobs s. We present qualitative results for the ablation on blob size from Table.3a) (main paper) in Fig. 3. Having a larger scale leads to accurate placement of the person but significantly distorts the background; however, having a small fixed scale for the blob limits the person's size to be inpainted. This results in failure for inpainting where no person is being placed.

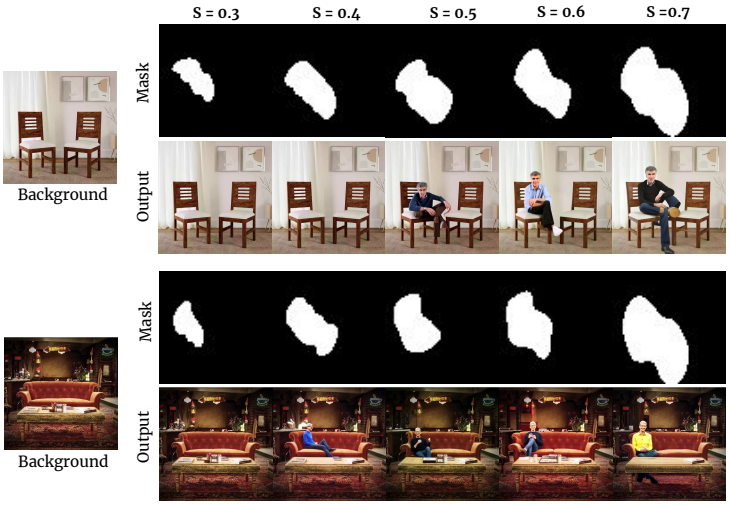

Fig. 3: Ablation over the scale of blob **s** in *semantic masks*.

Dilation size. In our early experiments, we found that using an exact mask of a person for inpainting doesn't generate the person as shown in Fig.5 (main paper). To this end, we dilated the obtained *semantic mask* after binarization to provide flexibility to the inpainting pipeline. The dilation effect is similar to the blob's scale parameter **s**. We ablate over the kernel size used for dilation in Fig. 4. As we increase the kernel size, the $\%Person$ increases, suggesting successful human placement; however, it increases background distortion. This is also evident in the qualitative results in Fig. 4. Having a significantly large kernel size also generates large humans that look unnatural.

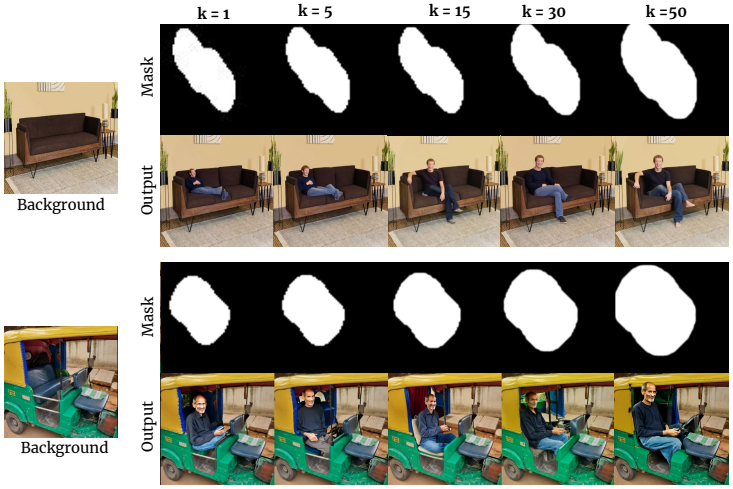

Fig. 4: Ablation over the kernel size used dilation of *semantic mask*.

Due to this, we always dilate our mask obtained in semantic mask generation before we use it for the inpainting task. However, too much dilation could cause a lot of background changes. So we performed ablation studies on the dilation kernel size, and from Tab. 1, we find that a kernel size of 15 seems to be performing best with low background change.

Table 1: Ablation over dilation kernel size.

Kernel Size (k)	LPIPS ↓	CLIP-sim ↑	% Person ↑
1	0.0869	0.2699	77.9
3	0.0882	0.2718	81.5
5	0.0904	0.2736	90.6
15	0.0992	0.2735	93.1
30	0.1135	0.2749	96.8

D.2 Subject conditioned inpainting

We ablate over different inpainting methods to generate plausible human placement results in Fig. 5. For our inpainting pipeline based on Textual Inversion [1], we use different numbers of training images. Additionally, we used only single-face crop images instead of full-body images to analyze the impact of providing full-body images. Using five images significantly improved the inpainting results with consistent identity in the generation. Using a face image generates decent results, but it has varied body shapes over the generations. We also compared with reference-based inpainting pipeline Paint-by-Example [6], which generates unnatural outputs with inconsistent identities. Our simplistic method for conditional inpainting generates significantly better human placement results that naturally blend wells in the scene.

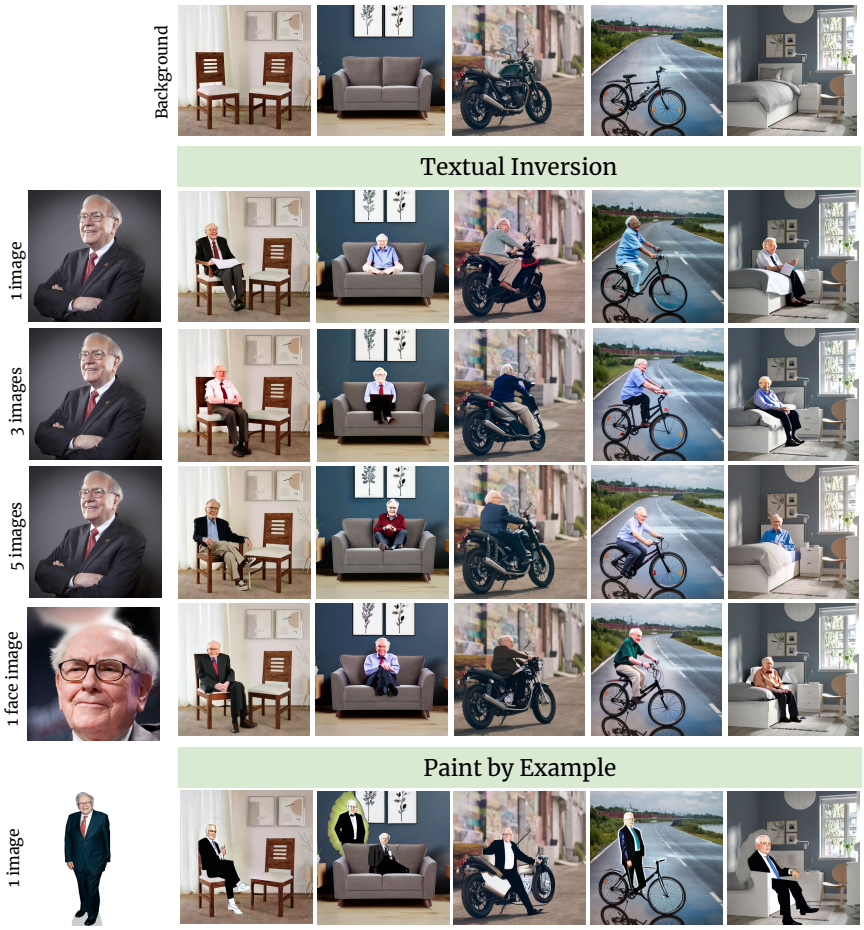

Fig. 5: Ablation over various inpainting configurations.

E Additional Results.

We present additional results for Semantic Human Placement, scene and human hallucination and text-based editing from our proposed method in Fig. 6, 7, 8.

Fig. 6: Semantic Human Placement with diverse human poses

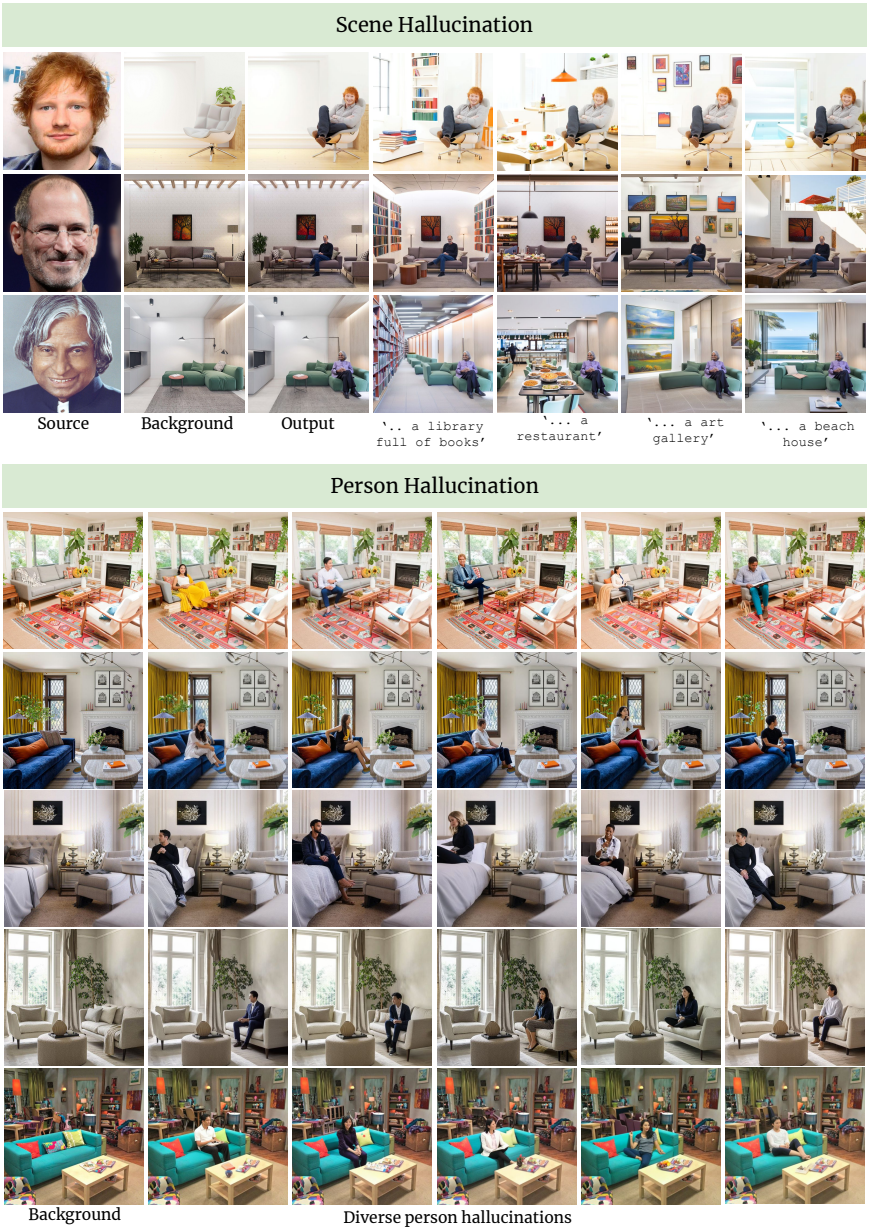

Fig. 7: Downstream applications - scene and person hallucinations

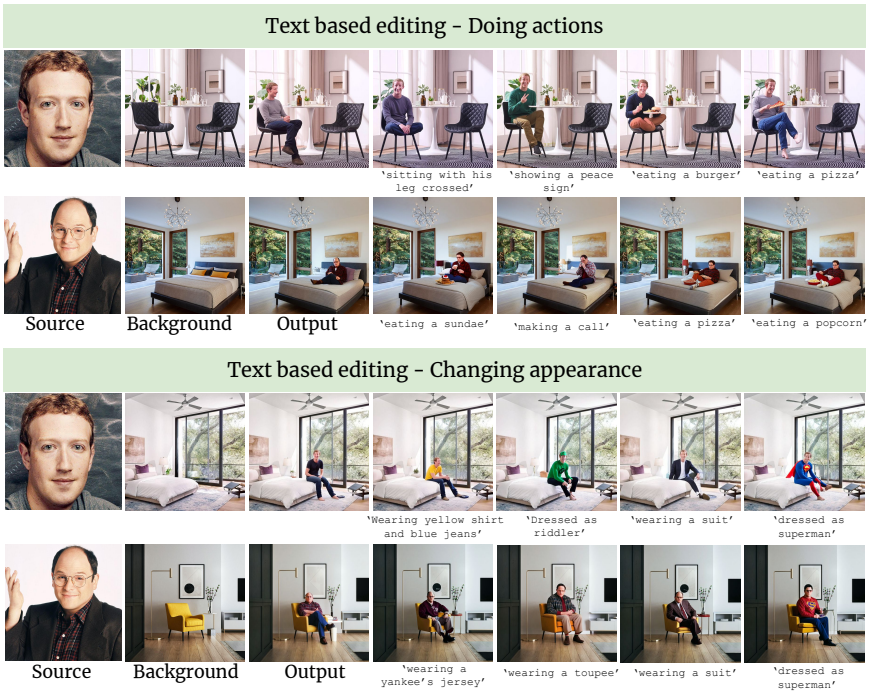

Fig. 8: Downstream applications - text-based editing of the generated person in the scene

F Prompts used for Placing Person

- A person sitting on a bed
- A person sitting on a sofa
- A person sitting on a chair
- A person sitting on a bench
- A person sitting in a car
- A person sitting on stairs
- A person sitting in an auto rickshaw
- A person riding a motorbike
- A person riding a cycle
- A person playing pool
- A person walking on a road
- A person sitting there
- A person running there
- A person walking there
- A person riding a cycle
- A person walking on a sidewalk
- A person standing on a podium
- A person standing in a room
- A person standing outdoor
- A person standing near the Eiffel Tower
- A person standing in front of the Taj Mahal

References

1. 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)
2. HuggingFace: Diffusers. `https://github.com/huggingface/diffusers` (2022)
3. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022)
4. Roettger, T.: Xl-textual-inversion. `https://github.com/oss-roettger/XL-Textual-Inversion/blob/main/XL_Inversion.ipynb` (2023)
5. 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)
6. 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)