Shape-Guided Diffusion with Inside-Outside Attention

Dong Huk Park[1*] **Grace Luo**[1*] **Clayton Toste**[1] **Samaneh Azadi**[2]
Xihui Liu[1,3†] **Maka Karalashvili**[4] **Anna Rohrbach**[1] **Trevor Darrell**[1]

[1]UC Berkeley [2]Meta AI [3]The University of Hong Kong [4]BMW Group

[*]*Denotes equal contribution.*

Figure 1. We demonstrate that prior work in local image editing [1, 8, 17, 18] fails to preserve precise object silhouette. We propose *Shape-Guided Diffusion*, a training-free method that uses a novel *Inside-Outside Attention* to respect shape input. Our method can be provided an object mask as input or infer a mask from text.

Abstract

We introduce precise object silhouette as a new constraint in text-to-image diffusion models, which we dub Shape-Guided Diffusion. Our training-free method uses an Inside-Outside Attention mechanism during the inversion and generation process to apply a shape constraint to the cross- and self-attention maps. Our mechanism designates which spatial region is the object (inside) vs. background (outside) then associates edits to the correct region. We demonstrate the efficacy of our method on the shape-guided editing task, where the model must replace an object according to a text prompt and object mask. We curate a new ShapePrompts benchmark derived from MS-COCO and achieve SOTA results in shape faithfulness without a degradation in text alignment or image realism according to both automatic metrics and annotator ratings. Our data and code will be made available at https://shape-guided-diffusion.github.io.

[†]This work was done when Xihui Liu was a postdoc at UC Berkeley.

1. Introduction

What is *shape*? By definition, an object's shape denotes the boundary, outline, or contour that separates it from the external world. As a result, shapes often carry a great deal of semantic meaning. For example, in the bottom row of Figure 1 the silhouette alone reveals that the object is a vehicle oriented rightwards, without the cues of color or texture. Given that object silhouette plays a key role in human visual processing, including object recognition and categorization [26], it follows that shape presents as a powerful cue for representing user intent when interacting with generative models. However, prior work in local image editing [1, 18] typically focuses on the coarsest form of shape input, often in the form of amorphous blobs or "user scribbles" where it is difficult to discern even the object category from the silhouette alone. As a result, these methods often fail when given precise shape inputs. We instead focus on precise object masks, which are easily acquired from off-the-shelf segmentation models. Thus, we consider the task

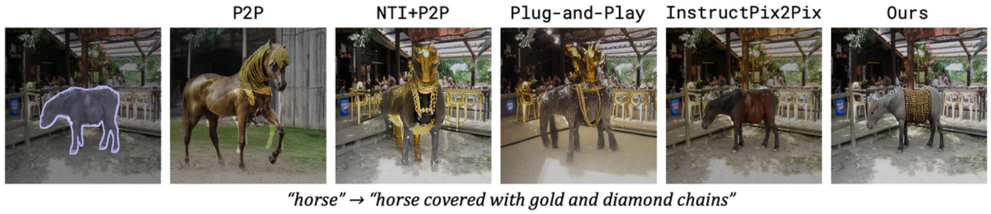

Figure 2. Our work differs from concurrent work in structure-preserving editing [2, 19, 35] in that we *constrain* attention maps such that edits are *localized* to a spatial region. Here we infer our shape constraint from the text prompt, thereby using the same amount of input as other methods.

of *shape-guided editing*, where a real image, text prompt, and object mask are fed to a pre-trained text-to-image diffusion model to synthesize a new object faithful to the the text prompt and the mask's shape.

Our method is motivated by the observation that diffusion models often contain spurious attentions that weakly associate object and background pixels, which makes it difficult to produce an edit that preserves a given shape boundary. To overcome this issue, we delineate the object (inside) and background (outside), with a novel Inside-Outside Attention mechanism that modifies the cross- and self- attention maps such that a token or pixel referring to the object is constrained to attend to pixels inside the shape, and vice versa. We apply this mechanism not only to the generation process to perform shape sensitive edits, but we also apply it to the inversion process to better preserve information about the source object before editing.

To summarize, our contributions include the following:
(1) We identify a limitation in prior image editing methods where the shape of the original object is not preserved and provide empirical insights on why this issue exists.
(2) Unlike existing mask-based editing adaptations (e.g., copying the background or finetuning the model to use mask input), we introduce a training-free mechanism that applies a shape constraint on the attention maps at inference time. To the best of our knowledge, we are the first work to explore *constraining* attention maps during inversion, which allows us to discover inverted noise that better preserves shape information from a real image.
(3) Our method achieves SOTA results in shape faithfulness on our MS-COCO ShapePrompts benchmark, and is rated by annotators as the best editing method 2.7x more frequently than the most competitive baseline. We demonstrate diverse editing capabilities such as object edits, background edits, and simultaneous inside-outside edits.

2. Related Work

Diffusion Models Diffusion models [29] define a Markov chain of diffusion steps that slowly adds random noise to data then learn a model to reverse this process. Variants include Denoising Diffusion Probabilistic Models (DDPM) [9], Denoising Diffusion Implicit Models (DDIM) [30], and score-based models [31]. Recently, diffusion models [20, 24, 25, 27] have shown impressive performance on text-guided image synthesis. Our work focuses on adapting these diffusion models towards text-guided local editing according to a text prompt and object mask.

Global and Local Image Editing Various works have extended generative models towards image editing. For text-guided global editing, StyleCLIP [22] adapts StyleGAN [11] and DiffusionCLIP [13] adapts diffusion models to edit entire images according to a text prompt. Blended Diffusion [1] proposes a method for local editing constrained to a mask by copying an appropriately noised version of the source image's background at each diffusion timestep. While this "copy background" technique can be generally combined with other methods to enable local editing in diffusion models, we demonstrate that this method alone is insufficient for preserving object shape, and we further improve shape faithfulness with our proposed method.

Structure Preserving Image Editing Various works in image-to-image translation can also preserve structure during editing. To accomplish this, some works copy random seeds [38], finetune model weights [12, 36], copy attention maps [8], or condition on a partially noised version of the source image [17]. There also exist a few concurrent works [2,19,35] that were developed independently and at the same time as our work, where we display a conceptual comparison in Figure 2 with additional examples in the Supplemental. (a) While these methods are often able to mimic the general style and structure of the source image, they struggle to perform a local edit where the background is left undisturbed. (b) This phenomenon occurs because these methods solely rely on the textual grounding capabilities of the diffusion model and *copy* attention maps, which are often noisy and entangle object and background pixels (see Figure 5). On the other hand, we *constrain* these attention maps according to a shape, which can be derived from a user or from a more reliable automatic grounding module such as a segmentation model, thereby incorporating more accurate spatial localization in a training-free fashion. (c) Due to this entanglement, works such as Prompt-to-Prompt (P2P) [8] often significantly drift when editing *real* images (see Figure 1). Our Inside-Outside Attention mechanism mitigates this drift when applied to the inversion process (see Figure 5).

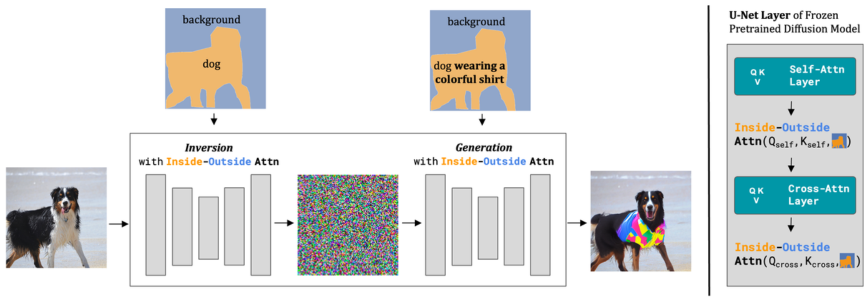

Figure 3. Shape-Guided Diffusion. Our method takes a real image, source prompt ("dog"), edit prompt ("dog wearing a colorful shirt"), as well as an optional object mask, and outputs an edited image. We infer the object mask from the source prompt if it is not provided using a shape inference function, e.g., a segmentation model. Left: we modify a frozen pretrained text-to-image diffusion model during both the inversion and generation processes. Right: we show a detailed view of one layer in the U-Net, where Inside-Outside Attention constrains the self- and cross-attention maps according to the mask.

Image Inpainting Image inpainting is the task of infilling the missing regions of an image. Researchers have proposed dilated convolution [10], partial convolution [15], gated convolution [41], contextual attention [40], and co-modulation [43] for GAN-based image inpainting. Lugmayr *et al.* [16] recently proposed a diffusion-based model for free-form image inpainting. There exist variants of GLIDE [20] and Stable Diffusion [18, 25] finetuned for text-conditional inpainting. However, these methods were trained with free-form masks without semantic meaning. There exist a few training-based methods that use object masks, none of which are publicly available. Make-a-Scene [6] trained an auto-regressive transformer conditioned on full segmentation maps of a scene. Shape-guided Object Inpainting [42] trained a GAN and Imagen Editor [37] trained Imagen [27] with object masks for inpainting. In contrast, we apply our model on top of an open-source text-to-image diffusion model at inference time. Because our method is training-free, it is more flexible and can be applied towards tasks beyond object editing, such as background editing or simultaneous inside-outside editing, as discussed in Section 5.2.

3. Shape-Guided Diffusion

We present Shape-Guided Diffusion, a *training-free* method that enables a pretrained text-to-image diffusion model to respect shape guidance. Our goal is to locally edit image x_{src} given text prompts $\mathcal{P}_{src}$ and $\mathcal{P}_{edit}$ and optional object mask m (inferred from $\mathcal{P}_{src}$ if not provided), so that edited image x_{edit} is faithful to both $\mathcal{P}_{edit}$ and m. We introduce Inside-Outside Attention to explicitly constrain the cross- and self-attention maps during both the inversion (image to noise) and generation (noise to image) processes. An overview of our method can be found in Figure 3 and Alg. 1. We build upon Stable Diffusion (SD), a Latent Diffusion Model (LDM) [25] that operates in low-resolution latent space. LDM latent space is a perceptually equivalent downsampled version of image space, meaning we are able to apply Inside-Outside Attention in latent space via downsampled object masks. For the rest of this paper, when we denote "pixel", "image", or "noise", we are referring to these concepts in LDM latent space.

Algorithm 1 Shape-Guided Diffusion

Input: A diffusion model DM with autoencoder $\mathcal{E}, \mathcal{D}$, real image x_{src}, a source prompt $\mathcal{P}_{src}$, an edit prompt $\mathcal{P}_{edit}$, and either a binary object mask m or a shape inference function InferShape$(\cdot)$.
Hyperparameters: Classifier-free guidance scale w_g.
Output: An edited image x_{edit} that differs from x_{src} only within the mask region m.

```
if m is not provided then
    m ← InferShape(x_src, P_src)
end if
[z̄_0, ..., z̄_T] ~ InsideOutsideInversion(z|E(x_src), P_src, m, DM)
z_T ← z̄_T
for all t from T to 1 do
    InsideOutsideAttention(DM, P_edit, m)
    z_cond ← DM(z_t, P_edit)
    z_uncond ← DM(z_t, ∅)
    z_{t-1} ← z_uncond + w_g * (z_cond − z_uncond)
    z_{t-1} ← z_{t-1} ⊙ m + z̄_{t-1} ⊙ (1 − m)
end for
x_edit ← D(z_0)
```

3.1. Inside-Outside Attention

LDMs contain both cross-attention layers used to produce a spatial attention map for each textual token and self-attention layers used to produce a spatial attention map for each pixel. We postulate that prior methods often fail because of spurious attentions – attentions that seek to edit the object compete with those that seek to preserve the background because they are not well localized (see Figure 5). Hence, we manipulate the cross-attention map such that the inside tokens are responsible for editing a distinct, non-overlapping spatial region compared with the outside tokens (e.g., "dog","shirt", etc. may only edit the dog and "back-

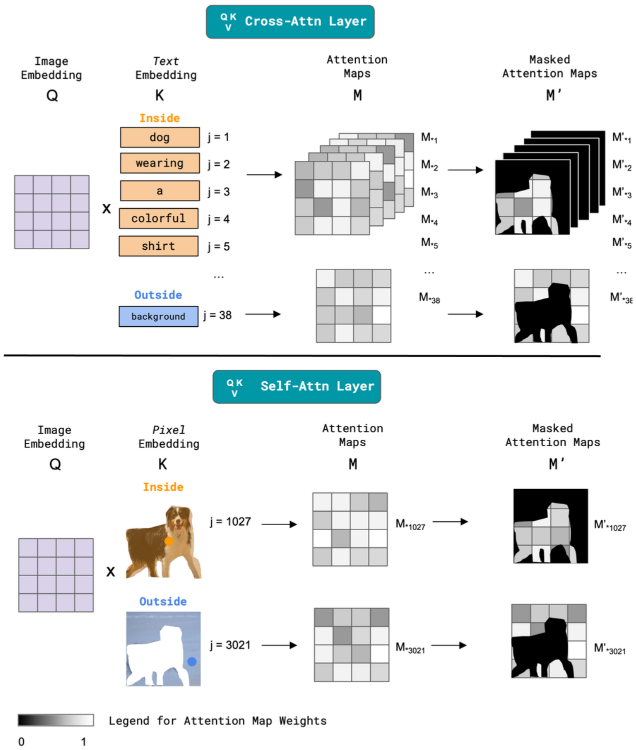

Figure 4. Inside-Outside Attention. We modify both the cross- and self-attention maps. Here j refers to token/pixel indices and M_{*j} denotes the attention map corresponding to the j-th index. Cross-Attn Layer (top): depending on whether the text embedding refers to the inside or outside the object, we constrain the attention map M according to the object mask or the inverted object mask to produce M'. Self-Attn Layer (bottom): we perform a similar operation on the inside and outside pixel embeddings.

ground" may only edit the remaining scene). Since self-attention layers heavily influence how pixels are grouped to form coherent objects, we apply a similar manipulation to the self-attention map to further ensure that the desired object is contained within the boundaries of the input mask.

An overview of Inside-Outside Attention is given in Figure 4 and our algorithm is defined as follows (also see Alg. 2). For one forward pass at each timestep during inversion or generation, we go through all layers of the diffusion model DM and manipulate the cross- and self-attention maps M. We denote the dimensions of M as $\mathbb{R}^{HW \times d_\tau}$ and $\mathbb{R}^{HW \times HW}$ for each cross- and self-attention map, respectively, where H is the image height, W is the image width, HW is the number of pixels in the flattened image, and d_τ is the number of tokens. We also downsample m according to the resolution of the cross- or self-attention layer. For the cross-attention map, we determine column indices J_{in} and J_{out} based on whether the token refers to the object or the background. For the self-attention map, we determine column indices J_{in} and J_{out} based on whether the pixel belongs inside or outside the object as defined by mask m. Finally, we compute the new constrained attention maps $M'_{*j_{in}} = \{M_{*j_{in}} \odot m \mid \forall j_{in} \in J_{in}\}$ and $M'_{*j_{out}} = \{M_{*j_{out}} \odot (1-m) \mid \forall j_{out} \in J_{out}\}$.

Algorithm 2 Inside-Outside Attention

Input: A diffusion model DM, a binary object mask m, a prompt $\mathcal{P}$.
Output: An edited diffusion model where the attention maps M are masked according to m and $\mathcal{P}$ for one forward pass.

for all $l \in$ layers(DM) **do**
  if type(l) is CrossAttention
    $J_{in} \leftarrow \{j \mid j$th token refers to object$\}$
    $J_{out} \leftarrow \{j \mid j$th token refers to background$\}$
  elif type(l) is SelfAttention
    $J_{in} \leftarrow \{j \mid j$th pixel belongs inside object$\}$
    $J_{out} \leftarrow \{j \mid j$th pixel belongs outside object$\}$
  $M'_{*j_{in}} = M_{*j_{in}} \odot m \quad \forall j_{in} \in J_{in}$
  $M'_{*j_{out}} = M_{*j_{out}} \odot (1-m) \quad \forall j_{out} \in J_{out}$
end for

3.2. Inside-Outside Inversion

To edit real images, we use DDIM inversion [20, 30] to convert the source image to inverted noise. However, we observe that using inversion with a text-to-image diffusion model often results in a shape-text faithfulness trade-off. While running generation with the fully conditional or unconditional model can reconstruct the real image, using non-zero levels of classifier-free guidance can completely drift from the real image, as seen in the bottom row of Figure 5. We propose applying Inside-Outside Attention to mitigate this trade-off. Similar to how prior work can associate tokens to *entire images* [7, 36], with Inside-Outside Attention we can associate tokens to *specific spatial regions*. As seen in the top row of Figure 5, applying our mechanism during inversion and generation allows one to both reconstruct and edit the real image with classifier-free guidance. We also visualize the cross-attention maps for the token "dog," where with our mechanism its effect is constrained to the silhouette and leaves the chair in the background unaffected, whereas without our mechanism its effect leaks into the background and morphs the dog while removing the chair.

3.3. Method Summary

In summary, we make the observation that object shape can be better preserved if spurious attentions are removed, and we propose the novel inference-time mechanism Inside-Outside Attention. Our method Shape-Guided Diffusion uses Inside-Outside Attention to constrain the attention maps during both inversion and generation, which we depict in Figure 3. The Shape-Guided Diffusion algorithm can

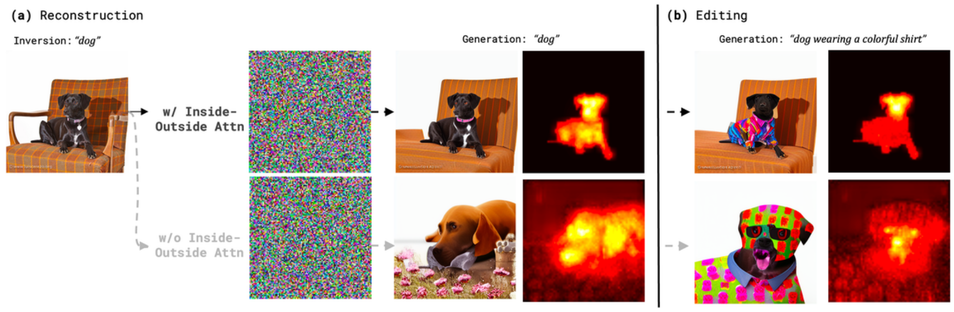

Figure 5. Spurious attentions and classifier-free guidance limits shape preservation. Inside-Outside Attention (top) preserves the shape relationship between the object and background by associating tokens to specific spatial regions. We demonstrate this property when reconstructing (left) and editing (right) a real image with classifier-free guidance. We also depict the cross attention map for the token "dog" averaged all attention heads and timesteps.

be defined as follows (also see Alg. 1). If the mask is not provided, we use the shape inference function InferShape$(\cdot)$ to identify $\mathcal{P}_{src}$ in the image. For our experiments we use an off-the-shelf segmentation model [4], but any method for textual grounding could also be used with our method. We run Inside-Outside Inversion on the conditional diffusion model driven by the prompt $\mathcal{P}_{src}$ (e.g., "dog") to get inverted noise $\bar{z}_T$. We then set our initial noise z_T to $\bar{z}_T$. For each sampling step, we apply Inside-Outside Attention for both the conditional and unconditional diffusion models using mask m and $\mathcal{P}_{edit}$ (e.g., "dog wearing a colorful shirt"). We mix the predictions of both models using the original formulation of Ho *et al*. [9], which applies classifier-free guidance to the conditional prediction (Line 10, Alg. 1). In early experiments we found this design choice leads to higher text alignment without a loss in other metrics. Finally, we copy the real image's background found during the inversion process $\bar{z}_{t-1} \cdot m$ to form the edited image prediction z_{t-1}. This ensures the edited image x_{edit} and the original image x_{src} only differ within the mask region m.

4. MS-COCO ShapePrompts

Benchmark We evaluate our approach on MS-COCO images [14]. We filter for object masks with an area between [2%, 50%] of the image, following prior work in image inpainting [33]. Our test set derived from MS-COCO val 2017 contains $1,149$ object masks spanning 10 categories covering animal, vehicle, food, and sports classes. We create a validation set with $1,000$ object masks in the same fashion derived from MS-COCO train 2017. For each category we design a few prompts that add clothing or accessories (e.g., "floral shirt" or "sunglasses"), manipulate color (e.g., "iridescent", "with spray paint graffiti"), switch material ("lego", "paper"), or specify rare subcategories ("spotted leopard cat", "tortilla wrapped sandwich"). More information about the prompts can be found in the Supplemental.
Metrics Since we aim to synthesize an image faithful to the input shape, we use mean Intersection over Union (mIoU) as a metric. Specifically, we compute the proportion of pixels within the masked region correctly synthesized as the desired object class, as determined by a segmentation model [4] trained on COCO-Stuff [3]. Since animal object masks are particularly fine-grained, and mIoU does not capture a full picture of degenerate cases (e.g., if the edit replaces a cat's full body with a cat's head), we also compute a keypoint-weighted mIoU (KW-mIoU) for the animal classes. Specifically, we weight each sample's mIoU by the percentage of correct keypoints when comparing the source vs. edited image, as determined by an animal keypoint detection model [39]. We also report FID scores as a metric for image realism, which measures the similarity of the distributions of real and synthetic images using the features of an Inception network [21, 34]. Finally, we report CLIP [23] scores as a metric for image-text alignment, which measures the similarity of the text prompt and synthetic image using the features of a large pretrained image-text model. More information on metrics can be found in the Supplemental.

5. Experiments

In Section 5.1 we evaluate our method on the shape-guided editing task where it must replace an object given a (real image, text prompt, object mask) triplet from MS-COCO ShapePrompts. We also evaluate on the same task with masks inferred from the text and ablate the use of our Inside-Outside Attention mechanism. In Section 5.2 we present additional results beyond object editing.
Baselines For our baselines, we compare against the local image editing method Blended Diffusion [1], the inpainting method SD-Inpaint [18], and the structure preserving methods SDEdit [17] and P2P [8]. Blended Diffusion, built on top of a Guided Diffusion [5] backbone, uses mask input by copying the source image's background at each timestep and text input by applying classifier guidance with CLIP [23]. SD-Inpaint, built on top of a Stable Diffusion [25] backbone, finetunes the model with an extra U-Net channel to use mask input and applies classifier-free guidance to use text input. SDEdit partially noises then denoises the source image and P2P copies cross attention maps to

Figure 6. Comparison to prior work. We compare our results with Blended Diffusion [1], SD-Inpaint [18], SDEdit [17], and P2P [8] on MS-COCO images. Our method is able to generate realistic edits that are faithful to both the input shape and text prompt. + Shape denotes a variant of the structure preserving method adapted for local image editing using the "copy background" method from [1].

preserve structure, and they apply classifier-free guidance to use text input. For the structure preserving methods we use implementations built on top of a Stable Diffusion backbone, and in some experiments we adapt them to use mask input by applying the "copy background" method from [1].

Experimental Setup For all baselines we use the default hyperparameters provided by their respective repositories. For sampling we use a standard DDIM scheduler for 50 inversion and generation steps. When using Inside-Outside Attention on cross-attention layers, we evenly divide the maximum number of text tokens excluding the <bos> token, resulting in 38 "inside" tokens and 38 "outside" tokens. The attentions for the <bos> token are zeroed out.

5.1. Comparison to Prior Work

MS-COCO Shape We first experiment with object masks provided by MS-COCO as our shape guidance. In Figure 6, we depict real images (first row) and edits made by Blended Diffusion (second row), SD-Inpaint (third row), and SDEdit + Shape (fourth row), P2P + Shape (fifth row), Ours (sixth row). Prior works demonstrate a variety of failure modes in shape-guided editing, where an object may be transformed into a new shape, removed completely, severely downscaled, or fail to respect the text prompt. On the other hand, our method is able to simultaneously respect the shape and the prompt without a compromise in image realism. As seen in Table 1, our method outperforms the local editing and in-

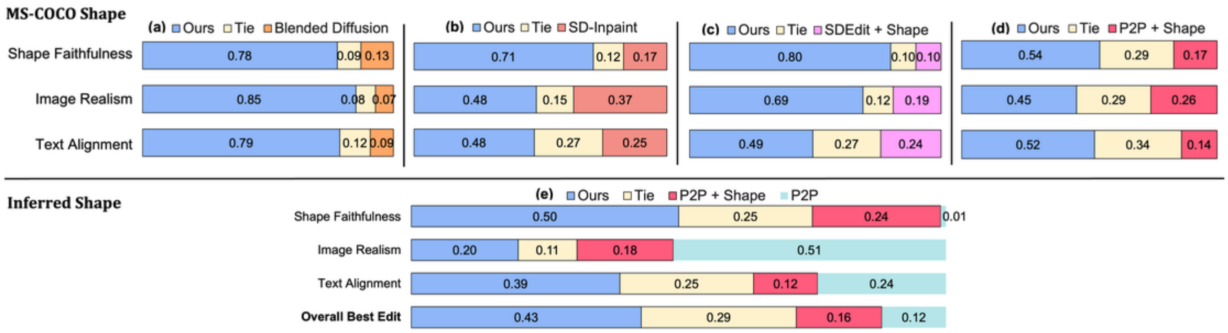

Figure 7. Annotator evaluation on MS-COCO ShapePrompts (100-sample subset of test set). Columns (a, b, c, d): we asked people to rate edits performed by our method vs. a baseline, where the two edits were presented as anonymized and in randomized order. Rows (shape faithfulness, image realism, text alignment): annotators selected the superior edit along these three axes. Each bar denotes the percentage of samples where the superior edit was "Ours", "Tie", or a baseline. In (e) we use the same procedure, except we presented three anonymized edits, ours vs. two baselines. Annotators were additionally asked to select the "overall best edit." We provide further details in the Supplemental.

Approach	KW-mIoU (↑)	mIoU (↑)	FID (↓)	CLIP (↑)
Real Images	83.3	76.3	-	0.15
MS-COCO Shape				
Blended Diffusion [1]	23.3	41.8	46.2	0.20
SD-Inpaint [18]	38.5	51.7	43.7	0.19
SDEdit + Shape [17]	31.0	49.9	45.1	**0.21**
P2P + Shape [8]	46.9	63.3	**39.6**	0.20
Ours (w/o IOA)	43.8	55.3	41.5	0.21
Ours	**53.3**	**63.6**	40.2	**0.21**
Inferred Shape				
P2P [8]	24.2	64.6	97.5	**0.26**
P2P + Shape [8]	37.7	54.0	51.1	0.21
Ours (w/o IOA)	33.0	46.0	56.8	0.22
Ours	**43.0**	**54.9**	**49.5**	0.22

Table 1. Automatic evaluation on MS-COCO ShapePrompts (test set). MS-COCO Shape uses object masks provided by MS-COCO, and Inferred Shape uses object masks inferred from the text. Ours w/o IOA denotes our method without Inside-Outside Attention.

Figure 8. Our method can handle challenging cases present in automatically inferred object masks such as overlapping instances or reflection.

painting baselines [1, 18] across the board, with at least a 15 point improvement in KW-mIoU. Comparing with the structure preserving baselines [8, 17], we achieve at least a 6 point improvement in KW-mIoU with comparable FID and CLIP scores. We hypothesize that our method is able to significantly outperform the baselines in shape faithfulness because our constraint operates on attention maps, whereas "copy background" operates on model outputs, which can be less meaningful at early timesteps when outputs are close to pure noise. See the Supplemental for further discussion.

We also conducted an evaluation with annotator ratings. We created four evaluations corresponding to each baseline, each of which contained 100 samples comparing an edit made by our method vs. the baseline in an anonymized and randomized fashion. For each sample, we asked five people to select the superior edit along the axes of shape faithfulness, image realism, and text alignment. As seen in the top row of Figure 7, annotators confirm that our method outperforms the baselines in shape faithfulness, with our method selected as superior at least 54% of the time (3.2x the most competitive baseline P2P + Shape). For image realism and text alignment, our method was selected as superior at least 48% of the time (1.3x and 1.9x the most competitive baseline SD-Inpaint).

Inferred Shape Next, we demonstrate that our method also works on automatically inferred masks (Figure 8). We compare against our most competitive baseline, vanilla P2P and P2P adapted for local image editing using the inferred mask (P2P + Shape). P2P often produces edits that look nothing like the source image (see Figure 1), so annotators rate it as the worst overall image *editing* method (see Figure 7). In fact, the method often produces simple images with a prominent single object, which significantly deviates from the distribution of complex, multi-object MS-COCO images. As a result, as seen in Table 1 these types of images achieve the worst FID scores with unusually high CLIP scores because these types of images tend to maximize text-

Figure 9. Additional editing results. Our method can perform intra- or inter-class edits on the same image, outside edits, and simultaneous inside-outside edits.

image alignment scores, at the cost of faithfully preserving content in the real image. In contrast, our method is rated by annotators as the best image editing method for 43% of samples, 2.7x more than the most competitive baseline P2P + Shape (see Figure 7). We also outperform P2P + Shape in all automatic metrics in Table 1.

Ablations In Table 1 we ablate our Inside-Outside Attention mechanism (Ours w/o IOA vs. Ours). The mechanism is a critical component of our method, providing a 9.5 point and 10 point increase in KW-mIoU in the MS-COCO Shape and Inferred Shape settings respectively. Ours w/o IOA performs better than all baselines on all metrics, except P2P + Shape (only P2P and our method use inversion), demonstrating how DDIM inversion is another critical component. In the Supplemental we also ablate the effect of DDIM inversion, guidance scale hyperparameters, and a soft vs. hard shape constraint on the self-attention maps.

5.2. Additional Editing Results

In Figure 9, we demonstrate additional capabilities of our method beyond object editing. (a) Our method is able to perform both intra- and inter- class edits on the same image, including adding accessories to a cow or transforming it into a sheep. (b) Our method is able to perform outside edits, including changing the time of day or location. Because we invert the image prior to editing, our method sometimes maintains structures from the real image, for example transforming the cabinet into a landmass in both edited images. (c) Our method is able to perform simultaneous edits with one prompt for the inside region and another for the outside region. Since our method delineates edits on the object vs. background, although every pixel in the image is transformed we can maintain the object-background relation from the source scene. In contrast, it is not obvious how to adapt structure preserving methods for this simultaneous editing setting, since with "copy background" they require one region (e.g., the background) to remain identical to the source image to enforce locality.

6. Conclusion

In this work, we present Shape-Guided Diffusion, a training-free method for utilizing precise object silhouette as a constraint in text-to-image diffusion models, which can be user provided or automatically inferred from the text. While prior work fails to respect shape inputs, our novel Inside-Outside Attention mechanism removes spurious attentions and localizes object vs. edits. We evaluate our method on our newly proposed MS-COCO ShapePrompts benchmark on the shape-guided editing task, where the goal is to edit an object given an input mask and text prompt. We show that our method significantly outperforms the baselines in shape faithfulness without a degradation in text alignment or image realism.

References

[1] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *CVPR*, 2022. 1, 2, 5, 6, 7, 11, 13

[2] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, November 2022. 2, 11, 15

[3] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *CVPR*, 2018. 5

[4] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *NeurIPS*, 2021. 5, 11, 16

[5] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. 5

[6] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. *arXiv preprint arXiv:2203.13131*, 2022. 3

[7] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 4

[8] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 1, 2, 5, 6, 7, 11, 12, 13, 14, 15, 20

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2, 5

[10] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 36(4):1–14, 2017. 3

[11] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2

[12] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023. 2

[13] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *CVPR*, 2022. 2

[14] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context. In *ECCV*, 2014. 5

[15] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *ECCV*, 2018. 3

[16] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, 2022. 3

[17] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. 1, 2, 5, 6, 7

[18] Runway ML. Stable diffusion inpainting. `https://huggingface.co/runwayml/stable-diffusion-inpainting`, 2022. 1, 3, 5, 6, 7

[19] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. *arXiv preprint arXiv:2211.09794*, 2022. 2, 11, 15

[20] 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. In *ICML*, 2022. 2, 3, 4

[21] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *CVPR*, 2022. 5

[22] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *ICCV*, 2021. 2

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 5

[24] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021. 2

[25] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 3, 5

[26] Eleanor Rosch, Carolyn B Mervis, Wayne D Gray, David M Johnson, and Penny Boyes-Braem. Basic objects in natural categories. *Cognitive Psychology*, 8:382–439, 1976. 1

[27] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 2, 3

[28] Sharif Shameem. Lexica. `https://lexica.art/`, 2022. 11

[29] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 2

[30] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 2, 4

[31] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021. 2

[32] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on*

computer vision and pattern recognition, pages 5693–5703, 2019. 12

[33] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. *arXiv preprint arXiv:2109.07161*, 2021. 5

[34] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 5

[35] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. *arXiv preprint arXiv:2211.12572*, 2022. 2, 11, 15

[36] Dani Valevski, Matan Kalman, Yossi Matias, and Yaniv Leviathan. Unitune: Text-driven image editing by fine tuning an image generation model on a single image. *arXiv preprint arXiv:2210.09477*, 2022. 2, 4

[37] Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J. Fleet, Radu Soricut, Jason Baldridge, Mohammad Norouzi, Peter Anderson, and William Chan. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. *arXiv preprint arXiv:2212.06909*, 2023. 3

[38] Chen Henry Wu and Fernando De la Torre. Unifying diffusion models' latent space, with applications to cyclediffusion and guidance. *arXiv preprint arXiv:2210.05559*, 2022. 2

[39] Hang Yu, Yufei Xu, Jing Zhang, Wei Zhao, Ziyu Guan, and Dacheng Tao. Ap-10k: A benchmark for animal pose estimation in the wild. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 5, 12

[40] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *CVPR*, 2018. 3

[41] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *ICCV*, 2019. 3

[42] Yu Zeng, Zhe Lin, and Vishal M Patel. Shape-guided object inpainting. *arXiv preprint arXiv:2204.07845*, 2022. 3

[43] Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I Chang, and Yan Xu. Large scale image completion via co-modulated generative adversarial networks. 2021. 3