Supplementary Material

In Section 7 we include a qualitative comparison with concurrent structure preserving methods, demonstrating how these methods often edit objects unmentioned by the text prompt. In Section 8 we discuss details regarding the MS-COCO ShapePrompts benchmark, in Section 9 we discuss details regarding our annotator evaluation, and in Section 10 we report additional ablations. Finally, we show a variety of additional examples from our method including success and failure cases in Section 11 as well as inferred shape edits, inter-class edits, and outside edits in Section 12.

7. Qualitative Comparison with Concurrent Structure Preserving Methods

We compare our method with concurrent work [2,19,35]. We can see that our method is able to perform better localized edits on a real image. Because these methods lack an explicit shape, they often change irrelevant objects that are not specified in the text prompt. In Row 2, Col 1-3 not only is the horse transformed into a robot but also the man. In Row 5, Col 3, 5 the wall that was present in the real image disappears. In contrast, the variant of our method that uses automatically inferred shapes (thereby requiring the same amount of input as the structure preserving methods) is able to perform edits that only modify the object of interest without disturbing the background. We used the official codebases released by the respective baselines and generated results using their default hyperparameter settings.

8. MS-COCO ShapePrompts Details

Prompts For our MS-COCO ShapePrompts benchmark we design a set of prompts where it is possible to simultaneously synthesize an object that is shape faithful and text aligned (as opposed to prompts entangled with shape, i.e., transforming "chihuahua dog" to "poodle dog" while respecting shape is difficult because poodles are characterized by floppy ears and fluffy fur). While our method is able to perform inter-class edits as seen in Figure 20, we focus our experiments on intra-class edits which make more sense given the shape constraint (e.g. some hyper-specific shapes like the silhouette of an elephant only make sense when edits are done within the object class). For this reason we design prompts for each object class as seen in Figure 10. These prompts were inspired by examples from prior work [1, 8] and a search engine with paired prompts and synthetic images from Stable Diffusion [28].

Shape Faithfulness Metric To measure shape faithfulness we use the pretrained segmentation model MaskFormer [4]. We demonstrate that the model makes meaningful predictions on synthetic images in Figure 11. Even more, the model's predictions are reasonably robust to out-of-distribution variants of the object class, such as "lego truck."

```
prompts = {
    "bear": [
        "stuffed {}",
        "{} wearing sunglasses"
    ],
    "bird": [
        "{} with blue and yellow feathers",
        "{} with iridescent feathers"
    ],
    "cat": [
        "spotted leopard {}",
        "{} wearing a yellow and black tie"
    ],
    "dog": [
        "{} wearing a floral jacket",
        "{} wearing a colorful shirt"
    ],
    "elephant": [
        "{} wearing christmas decorations",
        "holi festival {}"
    ],
    "horse": [
        "futuristic biomechanical robotic {} with synthetic body parts showing",
        "{} covered with gold and diamond chains"
    ],
    "sandwich": [
        "tortilla wrapped {}",
        "{} with peanut butter and jelly filling"
    ],
    "boat": [
        "inflatable {}",
        "{} made of candies"
    ],
    "kite": [
        "origami {} made of paper",
        "glitter {}"
    ],
    "truck": [
        "lego {}",
        "{} with spray paint graffiti"
    ],
}
```

Figure 10. Prompts from the MS-COCO ShapePrompts benchmark.

Figure 11. Synthetic images and their corresponding predicted segmentation and mIoU. Out-of-distribution variants of the object class, such as a truck made of legos, are still segmented correctly.

We use the segmentation model to compute mean intersection over union (mIoU). We compute mIoU on a per-sample basis (i.e., we average the IOU of each object regardless of size) as opposed to a per-pixel basis (which is typically used in semantic segmentation works) since it is equally impor-

Approach	Guidance Scale	KW-mIoU	mIoU (↑)	FID (↓)	CLIP (↑)
Real Images	N/A	86.5	78.6	-	0.16
(1) SD	7.5	30.9	52.5	46.2	0.21
(2) SD + DDIM Inv	7.5	39.8	61.2	42.8	0.21
(3) SD + DDIM Inv + Re-Weight (Ours w/o IOA)	3.5	46.2	59.6	40.6	0.21
(4) SD + DDIM Inv + Re-Weight + Token Inside-Outside Attn	3.5	48.1	62.9	40.6	0.21
(5) SD + DDIM Inv + Re-Weight + Soft Inside-Outside Attn	3.5	51.4	66.2	40.2	0.21
(6) SD + DDIM Inv + Re-Weight + Hard Inside-Outside Attn (Ours)	3.5	**54.8**	**67.6**	**39.0**	0.21

Table 2. Ablations on MS-COCO ShapePrompts (validation set).

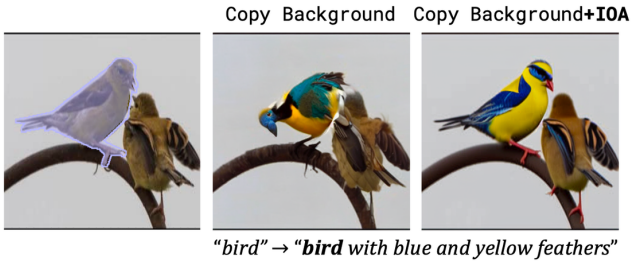

Figure 12. Shape signal from "copy background" is weak in early timesteps. In both examples we only use shape guidance in the first half of generation, where Inside-Outside Attention (+IOA) is able to provide stronger shape signal.

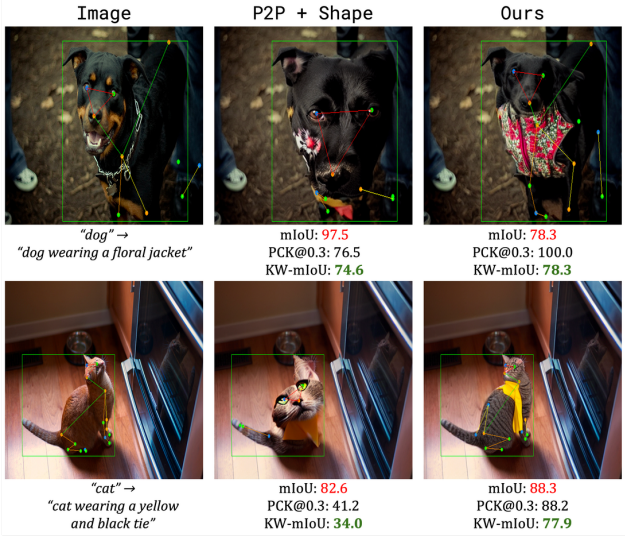

Figure 13. Comparison between mIoU and Keypoint-Weighted mIoU (KW-mIoU). Note that in this example P2P [8] receives a high mIoU score even though the edited objects are incorrectly scaled or cut off. By weighting each sample's mIoU with the percentage of correct keypoints (PCK) to compute the KW-mIoU, we can measure shape faithfulness more reliably.

tant to synthesize both small and large objects in a shape-faithful fashion. We set all pixels outside the mask to a null prediction to compute mIoU only within the edited mask region. We do this because in some settings (e.g. MS-COCO instance masks) the mask may specify one object instance out of multiple to edit, but our segmentation model would identify all instances of the same category, which would result in a diluted mIoU score. Additionally, for all methods that use "copy background" the background should remain identical to the original image.

In addition to mIoU, we introduce a new metric called Keypoint-Weighted mIoU (KW-mIoU). One issue with the standard mIoU metric is that edited objects that are incorrectly scaled or cut off could still get a very high mIoU if they fully occupy the shape (see Figure 13, Col 2). In order to mitigate this issue, we report KW-mIoU for animal classes (horse, dog, cat, elephant) where we weight each sample's mIoU by the percentage of correct keypoints (PCK) as computed between the source and edited images. We report KW-mIoU for animal classes only as we were not able to find a robust object pose estimation model with open-vocabulary capacities and reliable performance. By incorporating pose information, the proposed metric is able to be more sensitive to scale and object parts, and thus measure shape faithfulness more reliably. We use an animal keypoint detection model HRNet [32] pretrained on AP-10K dataset [39] as provided by https://github.com/open-mmlab/mmpose.

9. Annotator Evaluation Details

Our annotator evaluation included 25 total people spread across 5 evaluations (Ours vs. Blended Diffusion, Ours vs. SD-Inpaint, Ours vs. SDEdit + Shape, Ours vs. P2P + Shape, Ours vs. P2P vs. P2P + Shape). We asked each annotator to rate 100 samples, where they were told that they would be "rating AI-edited images, where the goal is to edit one object according to a text prompt while maintaining its shape." Each sample was formatted as pictured in Figure 14, in the grid the first column ("Original") corresponds to the original image, the second column ("A") corresponds to an edited image, and the third column ("B") corresponds to another edited image. To help annotators judge faithfulness in addition to the first row ("Full Image") we also provide the second row ("Masked Object") which masks the full image according to the shape of the original object. Along the metrics of shape faithfulness, text alignment, and image realism we asked annotators to rate whether synthetic image A or B performed better, or whether they "Tie." We define the metrics using the instructions seen in Figure 15.

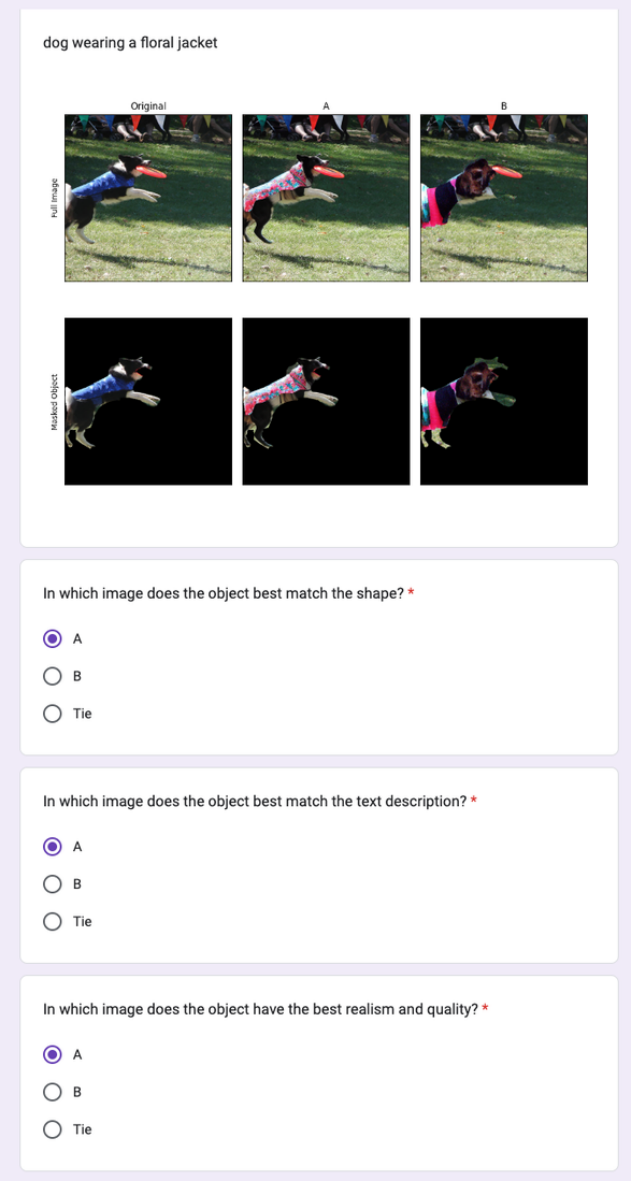

Figure 14. Screenshot of our annotator evaluation. People were asked to compare images edited by our method versus a baseline in anonymized and randomized order and rate (1) shape faithfulness, (2) text alignment, (3) image realism. In our final evaluation comparing Ours vs. P2P vs. P2P + Shape they also rated (4) best overall edit.

We also gave annotators the option to mark whether one of the synthetic image makes no substantial edit to the original object (i.e. the image copies and recolors the same object), which would be unfairly marked as having better shape faithfulness and image realism at the cost of text alignment under our evaluation standard. We removed these marked samples from our final comparison, resulting in the removal of less than 10% of samples from the 2500 total samples (from 25 annotators rating 100 images each) across all evaluations.

1. Shape
Definition: The edited object matches the scale and silhouette of the original object, and it is not cut off, zoomed out, or zoomed in. If the edited objects are similar in shape faithfulness, you may tie-break by judging the pose.
Heuristic: The edited object fills the masked version of the image in the "Masked Object" row.

2. Text
Definition: The edited object matches the text prompt.
Heuristic: The edited object matches what you were imagining given the text prompt.

3. Realism
Definition: The edited image looks realistic and lacks visual artifacts or logical flaws.
Heuristic: If you saw it on the internet you would believe the image is a real photograph / created by a human artist.

4. Best Edit
Definition: The edited image is the best along all three axes.
Heuristic: The edited image is what you would expect if a human were to edit the original image in Photoshop, i.e. most of the content is similar to the original image except the object which is modified according to the text prompt.

Figure 15. Metric definitions given as reference in the annotator evaluation.

10. Additional Ablations

We report an ablation study in Table 2. (1) uses a standard guidance scale of 7.5 and copies the background of the real image onto the prediction at each timestep as done in Blended Diffusion [1], (2) uses DDIM inverted noise during the generation process, (3) re-weights cross-attention maps based on the change between $\mathcal{P}_{src}$ and $\mathcal{P}_{edit}$ as done in P2P [8][1], (4) applies Inside-Outside Attention only to the cross-attention layers (Token Inside-Outside Attn), (5) applies Inside-Outside Attention with a hard mask for the cross-attention and soft mask for the self-attention layers (Soft Inside-Outside Attn), and (6) applies Inside-Outside Attention with a hard mask for both the cross- and self-attention layers (Hard Inside-Outside Attn), the design used in our final method. Comparing (1) and (2), we show that using DDIM inverted noise helps in both mIoU and FID. For (3), we empirically find that higher guidance scale, when used in combination with DDIM inversion, makes the model rely less on the the inverted noise, resulting in less realistic editing. However, we find that simply lowering the guidance scale leads to degradation in text faithfulness. Using cross-attention re-weighting mitigates this issue and allows us to achieve better image realism with similar performance in shape and text faithfulness. In (4), when we apply the Inside-Outside Attention mechanism only to the cross-attention layers we observe a small boost in mIoU with the same FID score, and when we apply it to both the cross- and self- attention layers in (5) we observe a more significant boost of 6.6 points in mIoU and 0.4 points in FID from (3). Comparing (5) and (6) we find that using a hard mask for Inside-Outside Attention performs better than using a soft mask, as seen by the further boost of 1.4 points in mIoU and 1.2 points in FID. Our final method that

[1]When re-weighting, we use a constant scalar upweighting of 2.5 as determined by hyperparameter sweeps in early experiments.

combines DDIM inversion, re-weighting, and hard Inside-Outside Attention achieves the best performance in mIoU and FID with scores of 67.6 and 39.0 respectively without a degradation in CLIP score. In Figure 12 we also demonstrate that our Inside-Outside Attention Mechanism is able to provide stronger shape signal than "copy background." Specifically, "copy background" provides a weaker shape cue because its signal is centered around how well the edit blends with the copied background at each timestep, which is harder to determine at early and noisy timesteps.

11. Success / Failure Cases

Success Cases In Figure 18 we show edits made by our method for each prompt in the MS-COCO ShapePrompts benchmark. We demonstrate that our method is able to handle partially occluded masks and maintain relationships between the object and background, as seen in the case of the "inflatable boat" where the edit maintains the position of the man and dog on that boat. We demonstrate that our method is able to add accessories while simultaneously respecting the input shape, as seen by the "elephant wearing christmas decorations" where an ear is converted to a santa hat. Finally, our method is seamlessly able to edit material ("lego truck", "boat made of candies", "origami kite made of paper") and color ("truck with spray paint graffiti", "bird with iridescent feathers", "holi festival elephant").

Failure Cases We also show failure cases of our method in Figure 19. Sometimes the shape is inherently difficult, such as the case with (a) uncommon pose (i.e., our method repositions an elephant sitting on its hind legs to standing), (b) uncommon perspective (i.e., our method transforms a close-up of a dog's eyes to a dog's entire face and converts its hair into fabric to obey the "floral jacket" in the prompt), or (c) multi-part mask inputs (i.e., our method only edits the front half of the truck into a lego material). Since we use DDIM inversion (d) ghosting can occur where remnants of the original object (e.g. a mouth or ear) can appear in the synthesized object. Since we localize the attention maps (e) global context can be ignored (i.e., our method edits a colorful dog into a black-and-white photo or creates artifacts at the boundary between a dog's legs and water). Finally, our method may produce strange (f) accessory placement (i.e., our method places a bowtie on the cat's arm because its neck is not visible in the original image).

12. Additional Editing Results

Inferred Shape Edits We show additional examples of edits made by our method with an inferred shape as input in Figure 17. Our method is able to handle a wide array of inferred shapes including those with multiple instances, noise, and occlusions.

Inter-Class Edits We show additional examples of inter-class edits in Figure 20, including converting from cat to dog, dog to cat, or sheep to cow.

Outside Edits. We show additional examples of outside edits in Figure 21, including transforming the background to different locations, seasons, or times of day.

Spurious Attentions and Classifier-Free Guidance In the main text we discuss how our Inside-Outside Attention mechanism is able to better perform reconstruction and editing with classifier-free guidance by removing spurious attentions. We additionally show our method vs. P2P [8] in the same setting in Figure 22. P2P exhibits spurious attentions where the token "dog" not only attends to the dog but also the background, causing the shape of the original dog to diverge completely.

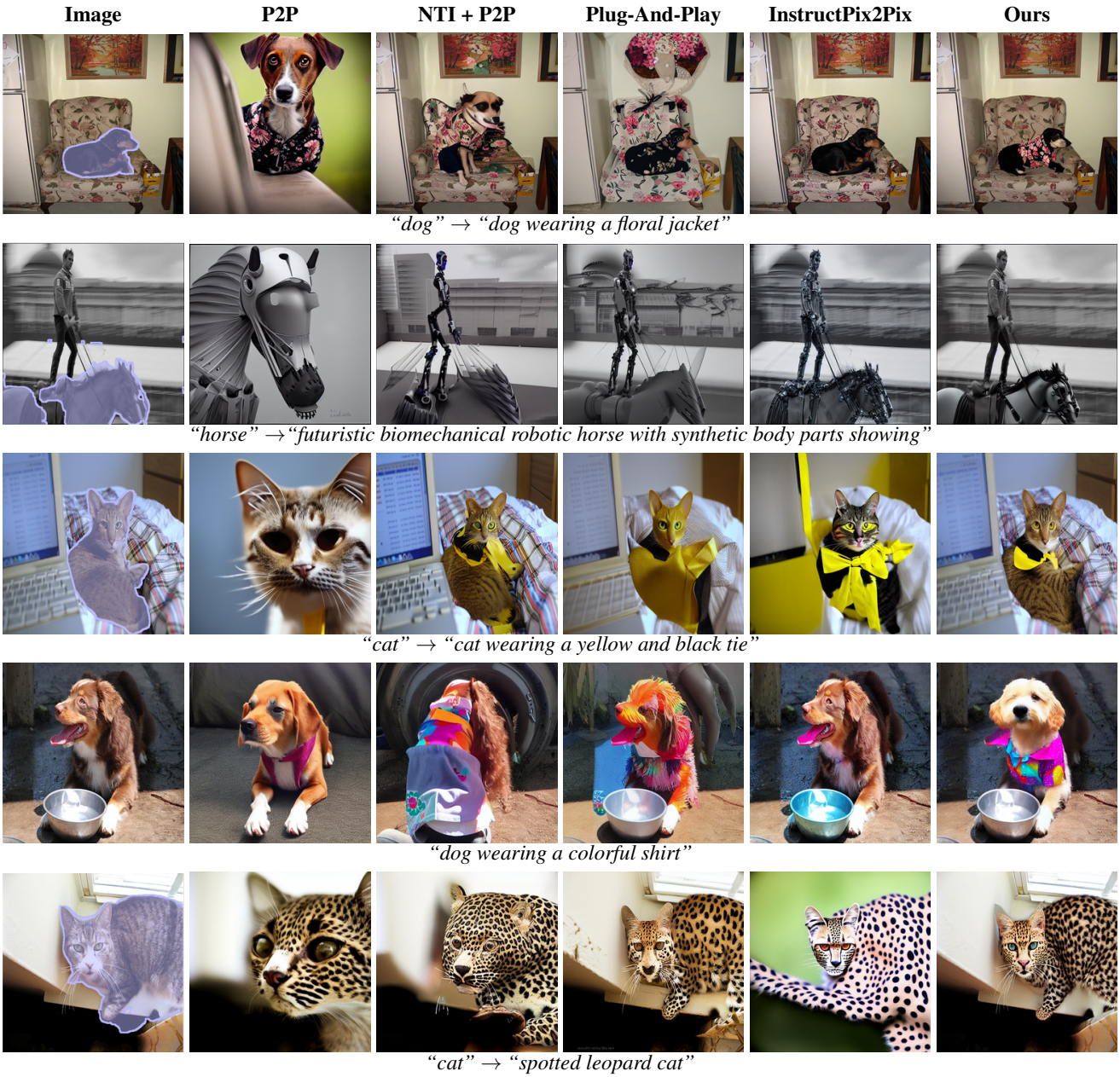

Figure 16. Qualitative examples comparing our method with concurrent work for structure preserving editing. We compare against P2P [8], NTI + P2P [19], Plug-and-Play [35], and InstructPix2Pix [2]. Here, we use the variant of our method that uses inferred shape, which requires the same amount of input (real image and text prompts) as the structure preserving methods.

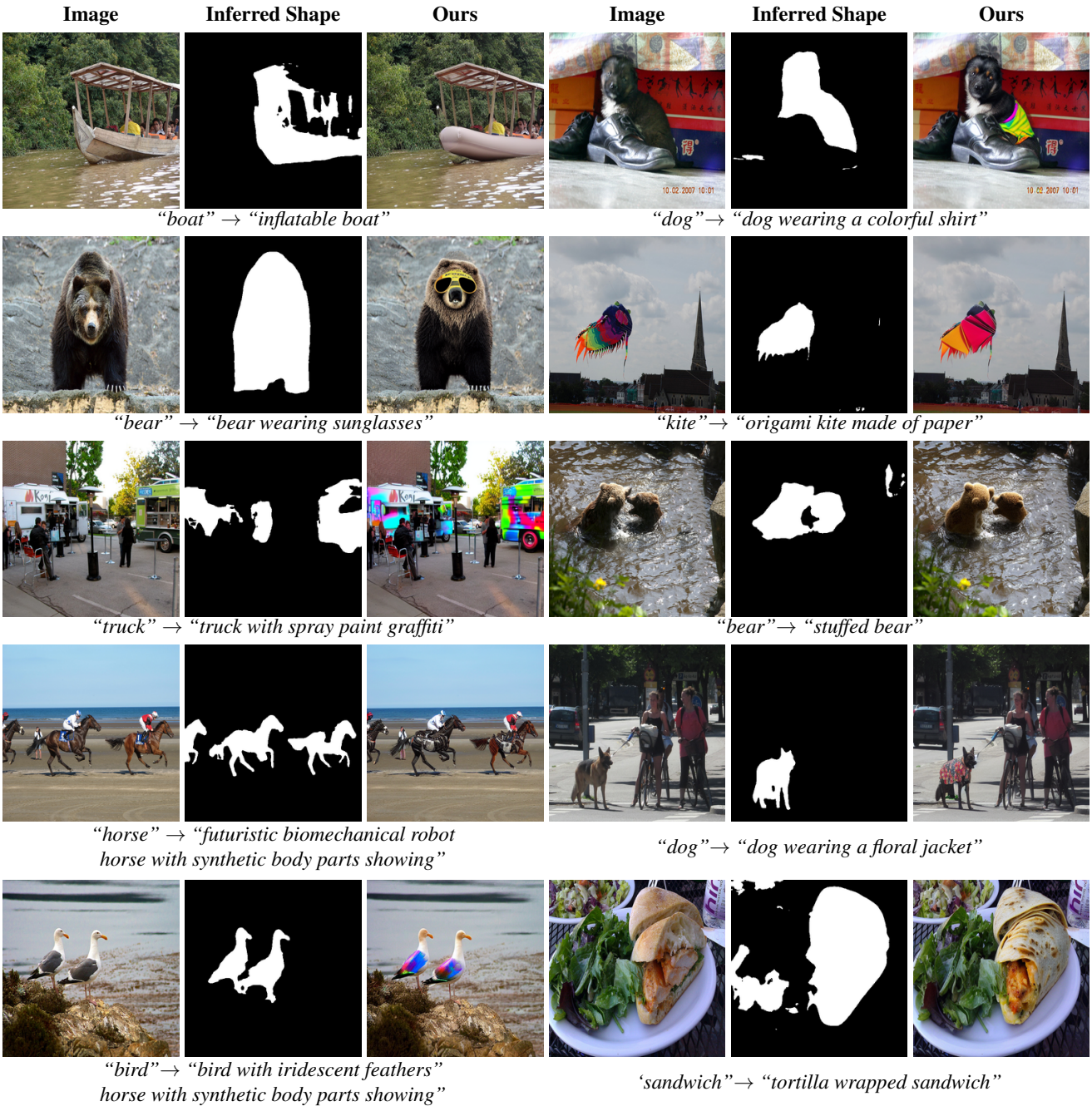

Figure 17. Additional examples generated by our method with masks automatically inferred from the text (predicted by MaskFormer [4]). Depending on the inferred shape, our method is able to edit multiple instances and handle complex shapes caused by noisy mask predictions or severe occlusions.

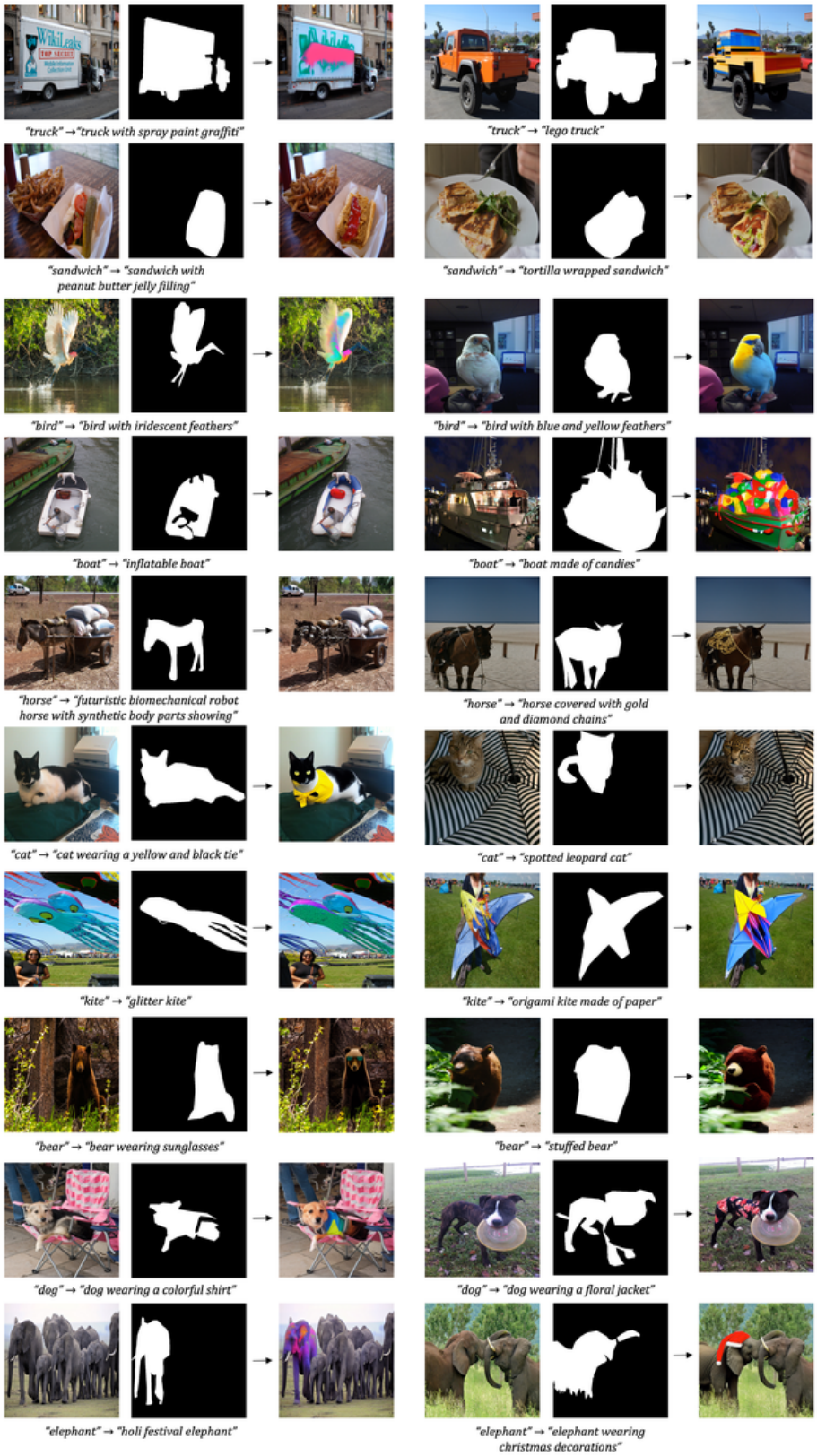

Figure 18. Examples of success cases from our method that demonstrate its ability to handle partially occluded masks, add accessories, transform materials, or recolor objects.

Figure 19. Examples of failure cases from our method that relate to (a) uncommon pose, (b) uncommon perspective, (c) multi-part mask, (d) ghosting, (e) global context, (f) accessory placement.

Figure 20. Additional examples of inter-class edits.

Figure 21. Additional examples of outside edits from our method where we transform the background to various locations (New York City, London), seasons (winter, autumn), and times of day (sunset, night) for various objects (truck, boat, cat).

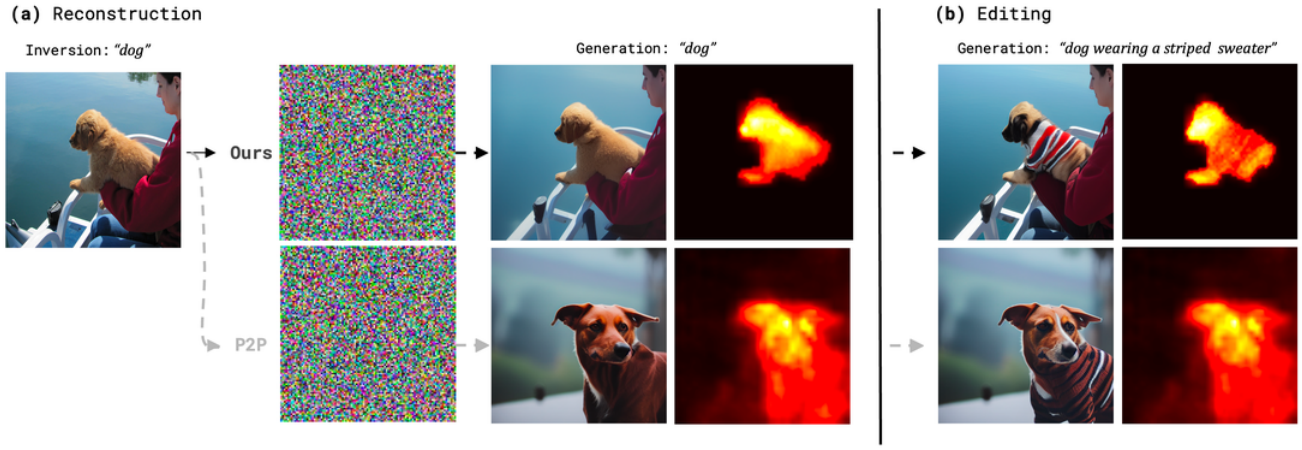

Figure 22. Spurious attentions and classifier-free guidance also affects P2P [8]. We compare our method (top) and P2P (bottom) for reconstructing (left) and editing (right) an image with corresponding cross attention maps for the token "dog" averaged over all layers and timesteps.