X-Edit: Detecting and Localizing Edits in Images Altered by Text-Guided Diffusion Models

Supplementary Material

G. Overview

This document is structured as follows:

- Sec. H: Additional qualitative results;
- Sec. I: Prediction distributions for original images;
- Sec. J: Qualitative results on out-of-distribution dataset.

H. Additional Qualitative results

In this section we complement Fig. 4 by showing in Fig. 8 some more qualitative results we could not include in the main paper for space constraints.

I. Prediction distributions for original images

In this section, we analyze the distribution of predicted masks generated on original images by two segmentation models: SAM [25] and our finetuned X-Edit method. We already shown in Fig. 5 how our model is better than SAM at predicting blank mask for original images. In Fig. 7 we provide a broader visualization by plotting prediction values over the entire test dataset. We can observe how X-Edit's histogram is concentrated around 0, while SAM's shows a significant spread that translates in non-zero prediction mask displayed in Fig. 5.

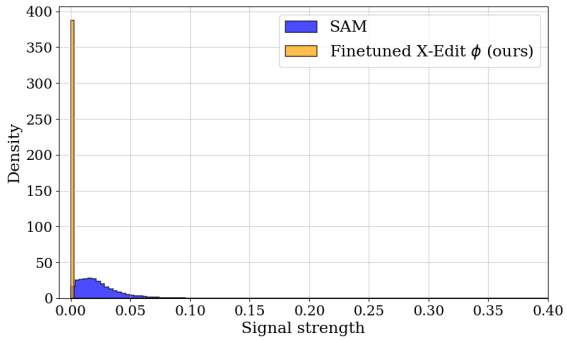

Figure 7. **Comparison of model distributions for original images.** The histograms display the density distributions of predicted mask values for SAM and Finetuned X-Edit ϕ models.

J. Qualitative results on out-of-distribution dataset

In this section, we provide some additional qualitative results on an out-of-distribution dataset we build.

J.1. Out-of-distibution dataset

We build the original part of our out-of-distribution by randomly selecting 100 images from the Flickr30k [62] test split using the Hugging Face Datasets library [28]. We choose each image based on a minimum dimension criterion of 500 pixels to ensure high-quality inputs suitable for editing. Alongside the images, we extract the first caption provided for each image to use as reference text. Besides InstructPix2Pix [5], we select three additional state-of-the-art TGIE methods for editing images:

1. FPE [30] uses self-attention control to guide the diffusion process towards the target prompt
2. MasaCtrl [7] editing allows to set up the mutual self-attention controller with specified steps and layers, thus registering the attention editor within the diffusion pipeline.
3. PnP [53] manipulates internal spatial features and self-attention components during the diffusion process.

By applying these editing methods, we generate two edited versions for each image in the test set. These edited images are used to assess the performance of our models in handling out-of-distribution data and to evaluate their robustness to different types of image alterations.

J.2. Results

Fig. 9 shows some examples of our X-Edit method vs SAM on Flickr30k original images. We notice how our method still produces blank predictions while SAM produces some non-zero maps in some cases.

Fig. 10, Fig. 11, Fig. 12 and Fig. 13 show some examples of our X-Edit method vs SAM [25] and SegFormer [59] on edited images generated by InstructPix2Pix, FPE, MasaCtrl and PnP respectively. We keep the same original images across the figures to enable a clearer comparison, not just between the segmentation methods, but also across the variety of editing outputs with its unique challenges. We can observe how for InstructPix2Pix and FPE our finetuned X-Edit produces masks that are very close to the ground-truth, while SAM tends to produce shallower masks and SegFormer is prone to false positives. This shows X-Edit's strength in precisely identifying edited regions, especially for localized changes like "add fire" or "turn the sand into water" where segmentation closely matches edits. Conversely, SAM's shallow masks fail to capture precise boundaries, and SegFormer's false positives indicate misclassification of unrelated regions. Regarding MasaCtrl

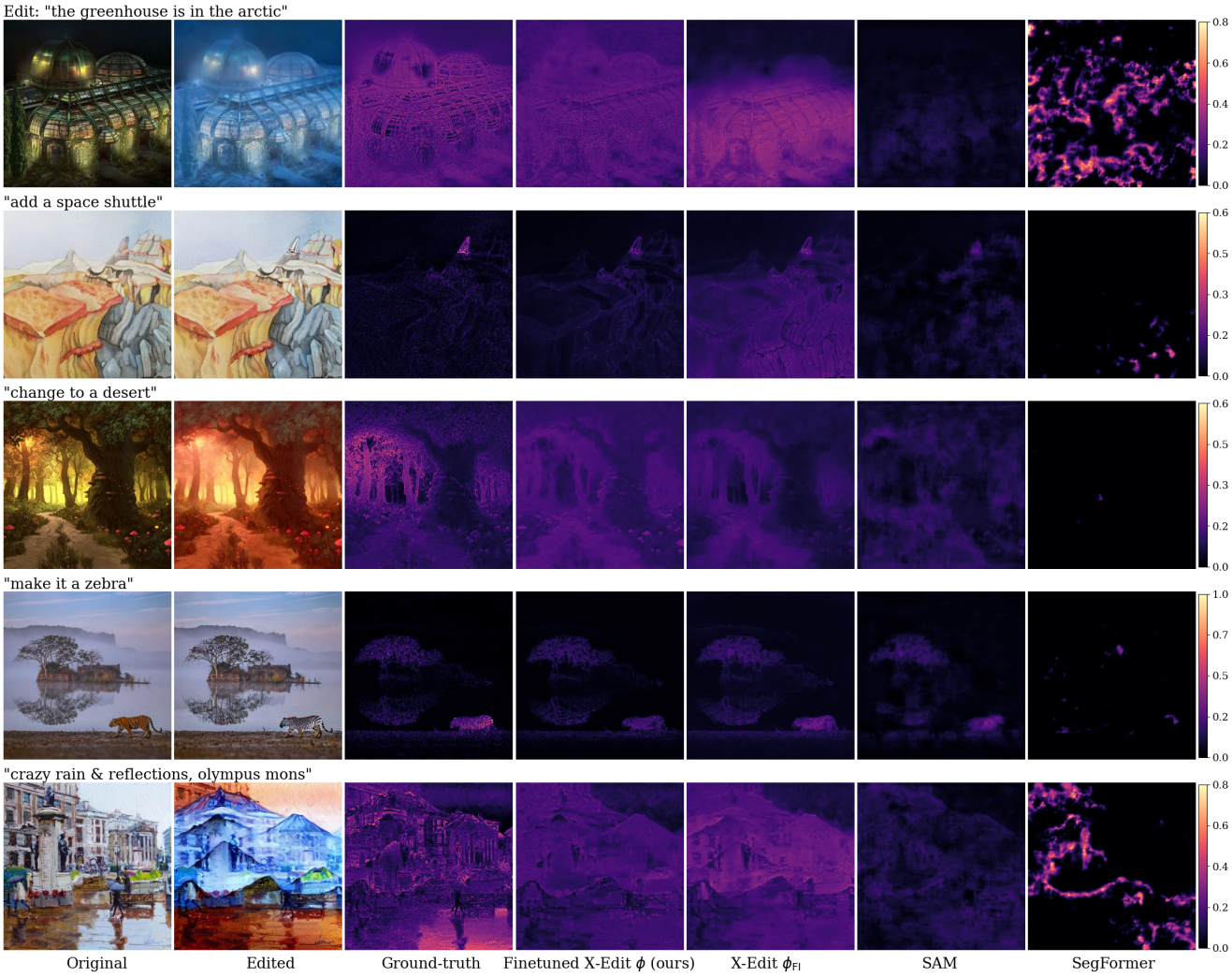

Figure 8. **Additional comparison of predicted masks for edited images.** From left to right: original image, edited image, ground truth mask indicating the edited regions, predicted mask from X-Edit finetuned on ϕ, X-Edit on ϕ_{FI}, SAM and SegFormer. X-Edit finetuned on ϕ (4th column) outperforms other models by more accurately capturing both the shape and placement of edits, demonstrating finer boundary alignment and better preservation of details in complex regions. This improvement highlights X-Edit's effectiveness in maintaining contextual coherence and producing higher-fidelity masks for intricate modifications.

and PnP edits, we notice how X-Edit on ϕ_{FI} seems to be best model, with SAM and SegFormer showing same limitation as before. Edits by MasaCtrl and PnP, involving complex, large-scale changes like pose or style shifts, present additional challenges. X-Edit on ϕ_{FI} effectively identifies the broader impacted regions while maintaining coherence. However, SAM and SegFormer struggle, underestimating edit scopes or misapplying changes to unrelated areas.

We selected edited images to showcase diverse challenges, evaluating how X-Edit and other models predict edited regions. Examples include additive edits (*e.g.*, "add fire"), background or foreground modifications (*e.g.*, "turn the grass into pool"), stylistic changes (*e.g.*, "make it an Andy Warhol painting"), and targeted edits (*e.g.*, "turn the straw hat into a red hat"), each testing specific capabilities like seamless integration, spatial coherence, style adaptation, and precise localization. These scenarios are crucial for testing models' ability to accurately identify changed regions. The range of chosen edits enables the assessment of models' robustness in capturing edits while maintaining accuracy. Importantly, the examples highlight both successes and failures. For instance, FPE identifies edited regions well in context-aware tasks but struggles with stylistic transformations, introducing artifacts and failing to preserve spatial coherence. Similarly, MasaCtrl excels in structural edits and pose adjustments but can distort facial features or intro-

duce unintended changes, especially in close-ups. For example, edits involving faces or abstract changes sometimes show exaggerated or inaccurate deformations.

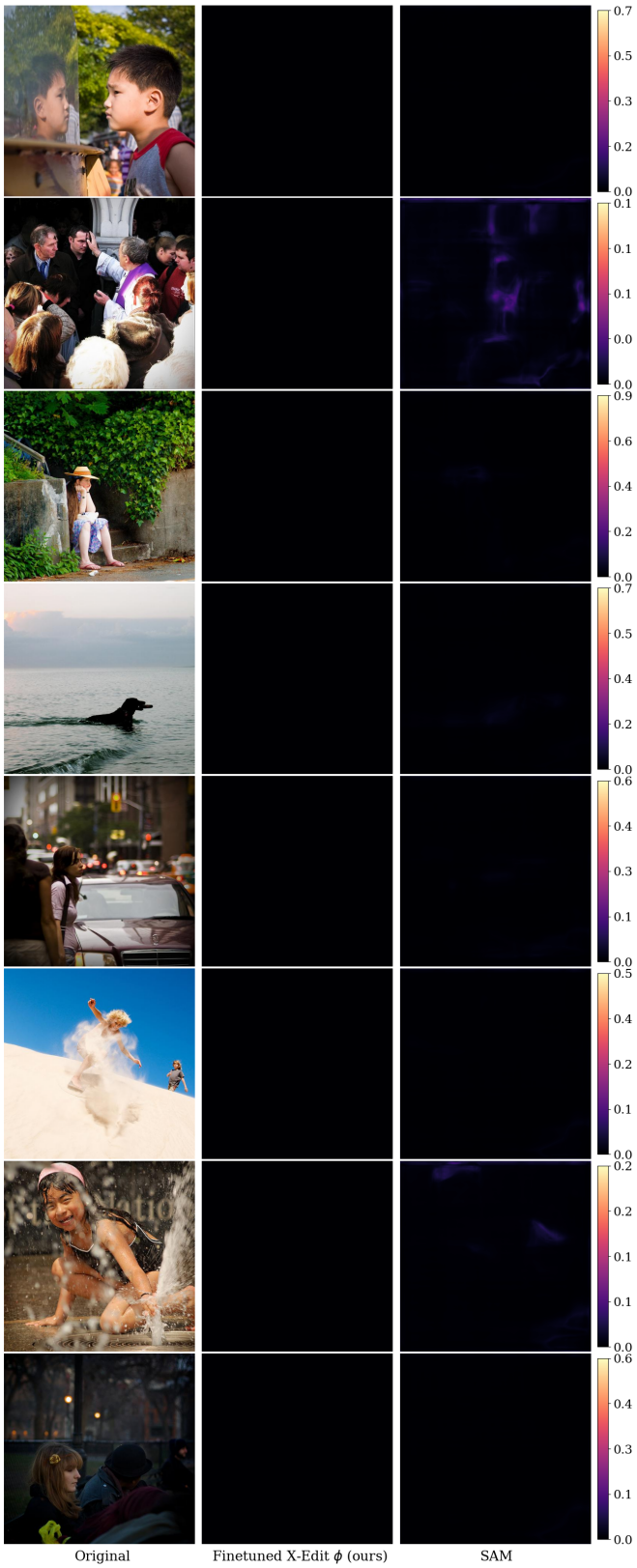

Figure 9. **Comparison of predicted masks for original images on out-of-distribution flickr30k images.**

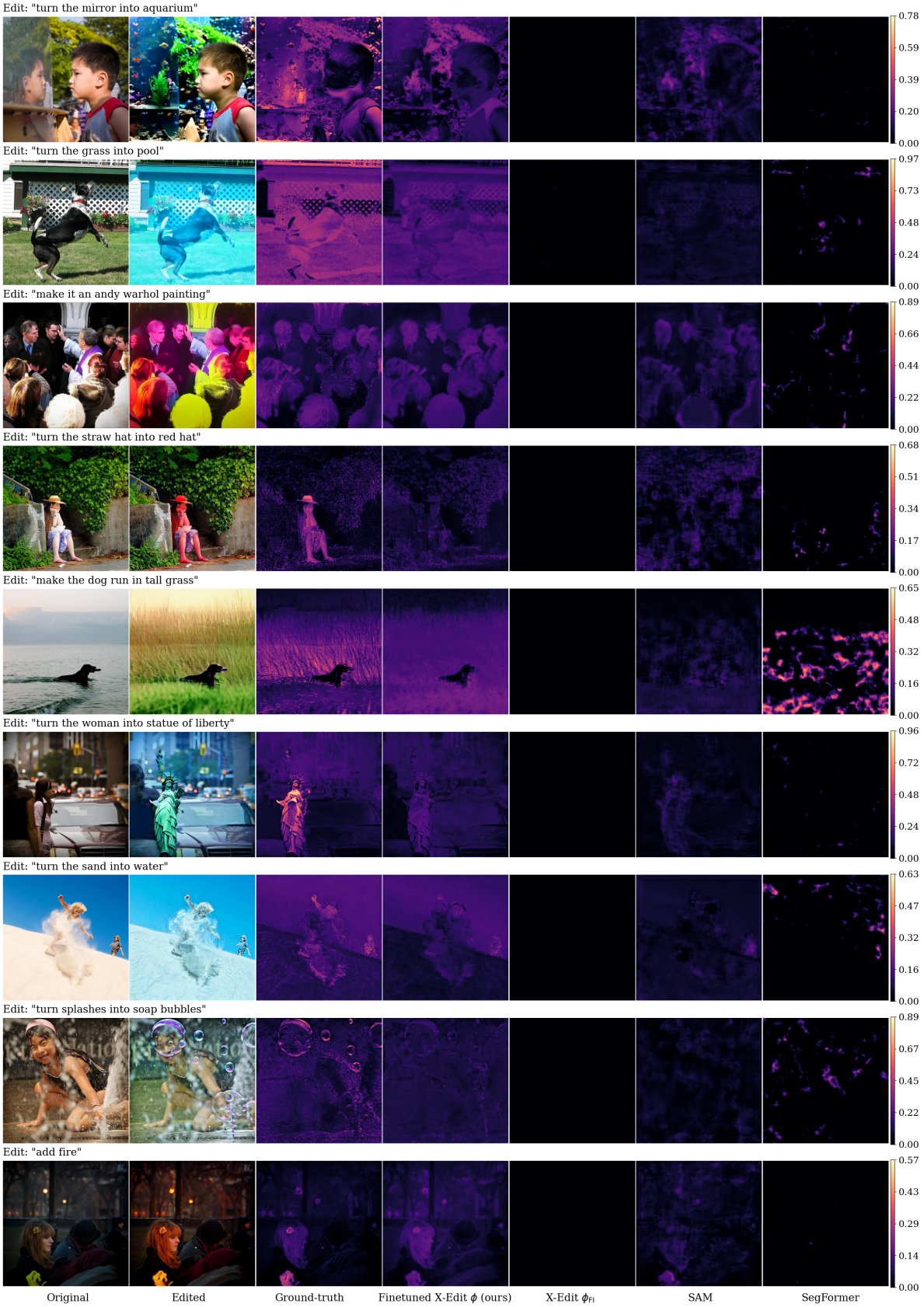

Figure 10. **Comparison of predicted masks for edited images with InstructPix2Pix method on out-of-distribution Flickr30k images.**

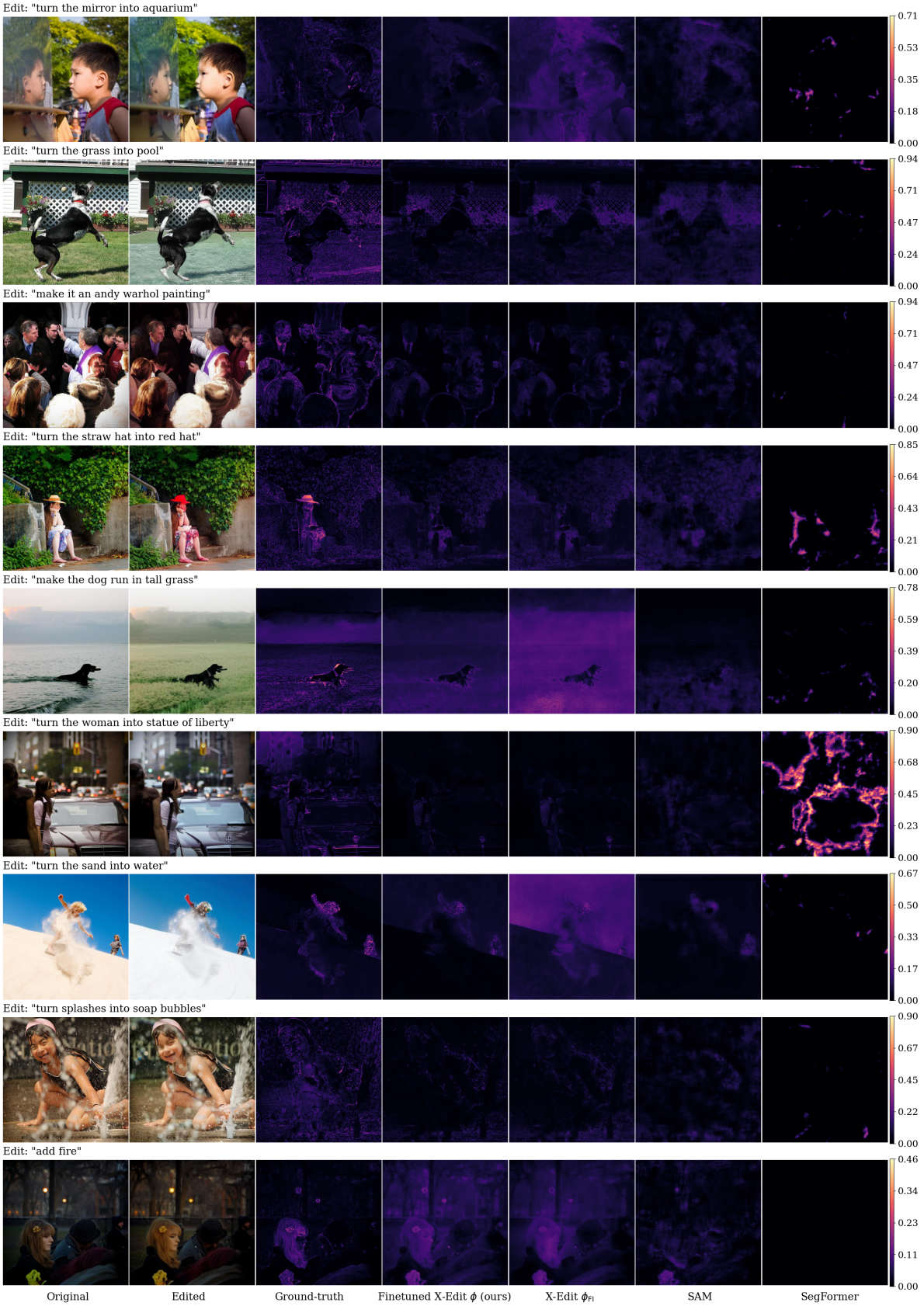

Figure 11. **Comparison of predicted masks for edited images with out-of-distribution FPE method on out-of-distribution Flickr30k images.**

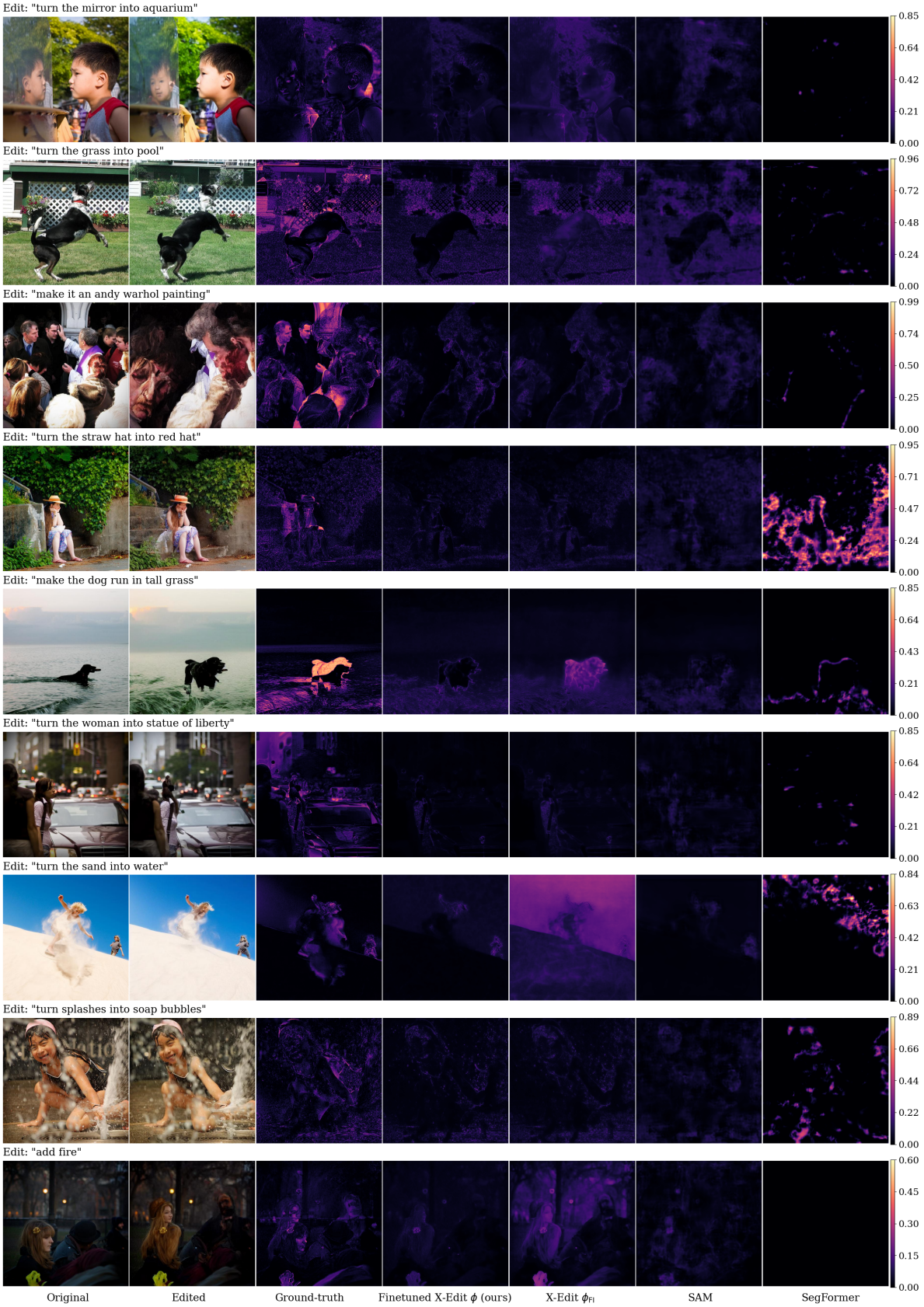

Figure 12. **Comparison of predicted masks for edited images with MasaCtrl method on out-of-distribution Flickr30k images.**

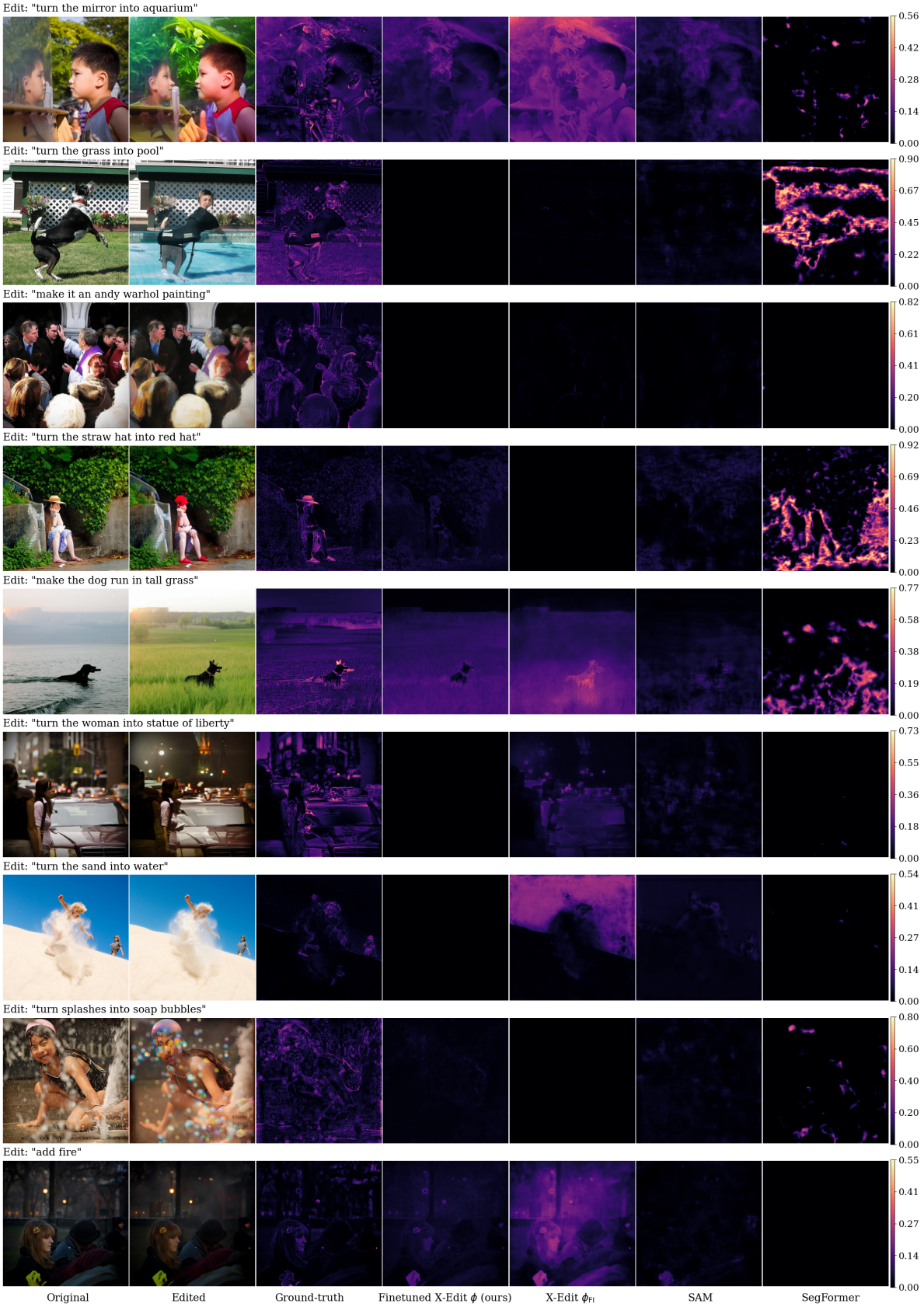

Figure 13. **Comparison of predicted masks for edited images with PnP method on out-of-distribution Flickr30k images.**