Final Project Proposal: Mask-Guided InstructPix2Pix

Abraham Yirga
Missouri University of Science and Technology
aayfn7@umsystem.edu

CS 5404 – Introduction to Computer Vision

Instructor: Dr. Ce Zhou

Abstract

Recent advancements in conditional diffusion models have enabled remarkable progress in text-based image editing. A prominent example, InstructPix2Pix, can modify images based on natural language instructions. However, it can sometimes struggle with spatial precision, causing edits to affect unintended regions or failing to preserve the structure of the background. In this work, I propose an extension, "Mask-Guided InstructPix2Pix," that aims to improve the spatial controllability and precision of instruction-based edits. I hypothesize that by incorporating an object mask as an additional condition, I can constrain the diffusion model's edits to a user-defined region of interest. This will lead to more accurate and predictable results, particularly for complex instructions involving specific objects within a scene. My preliminary plan involves integrating a segmentation mask into the model's architecture and fine-tuning it on a modified dataset to validate the approach.

1. Introduction and Motivation

The ability to edit images using natural language is a long-standing goal in computer vision. Recent breakthroughs in diffusion models have led to powerful tools that are bringing this goal closer to reality. One of the leading methods in this area is InstructPix2Pix [1], a model trained to perform image edits based on human-written instructions.

While groundbreaking, InstructPix2Pix sometimes produces edits that "bleed" outside the intended object or unnecessarily alter the background. For example, an instruction like "make his jacket out of leather" might also change the texture of the shirt or the wall behind the person. This lack of precise spatial control limits its practical applicability for detailed editing tasks.

My motivation is to address this limitation. I propose to enhance InstructPix2Pix by providing the model with explicit spatial guidance in the form of a segmentation mask. This allows a user to specify not only *what* to change via text but also *where* to change it via a mask, leading to more predictable and higher-fidelity image edits.

2. Summary of the Base Paper

The paper "InstructPix2Pix: Learning to Follow Image Editing Instructions" by Brooks et al. [1] proposes a conditional diffusion model for instruction-based image editing. A key challenge for this task is the lack of paired training data (i.e., triplets of input image, instruction, and output image). The authors cleverly solve this by generating a large-scale synthetic dataset. They use a large language model (GPT-3) to create instruction/caption pairs and a pre-trained text-to-image model (Stable Diffusion) combined with Prompt-to-Prompt to generate the corresponding image pairs.

The final InstructPix2Pix model is trained on this dataset of over 450,000 examples. It takes an input image and a text instruction as conditioning and directly generates the edited image in a single forward pass, making it fast and efficient.

3. Problem Statement and Hypothesis

The core problem I address is the lack of fine-grained spatial control in the InstructPix2Pix framework. The model implicitly learns the location of an edit from the instruction, but this can be ambiguous and lead to imperfect results.

My central hypothesis is: **By explicitly conditioning the diffusion model on a user-provided segmentation mask in addition to the text instruction, I can significantly improve the spatial precision of the edits and better preserve the unedited regions of the image.**

4. Proposed Idea and Preliminary Plan

I propose **Mask-Guided InstructPix2Pix**, an instruction-following editor that accepts three aligned inputs: the original image, a natural-language instruction, and a binary mask highlighting the spatial extent of the desired change. The mask acts as an explicit spatial prior that constrains the diffusion trajectory so that edits occur only where the user expects them.

My preliminary plan is structured as follows:

1. **Baseline Reproduction:** Set up the public InstructPix2Pix implementation, reproduce qualitative figures and quantitative scores reported in the paper, and document any deviations.
2. **Mask-Aware Architecture:** Modify the diffusion U-Net to ingest an additional mask channel and explore both early-fusion (concatenation) and attention-based conditioning strategies. I will also add a lightweight mask encoder that produces learned spatial embeddings.
3. **Data Generation:** Build a pipeline that augments the synthetic instruction dataset with segmentation masks derived from the Segment Anything Model (SAM) [2]. Post-process masks to match the textual instruction (e.g., dilating around object boundaries to allow natural transitions).
4. **Training and Ablations:** Fine-tune the modified model on (image, instruction, mask, edited image) quadruplets, then run ablations over mask-noise levels, conditioning strategies, and mask accuracy.
5. **Evaluation:** Compare against the original InstructPix2Pix through automatic metrics and a small human study that judges spatial compliance and instruction fidelity.

5. Methodology and Evaluation Plan

The mask will be concatenated with the RGB channels before encoding, while a parallel branch embeds high-level mask cues into the cross-attention layers used for instruction conditioning. During training I will minimize the standard diffusion loss plus a mask consistency penalty that suppresses changes in unmasked pixels. To increase robustness, I will randomly jitter the mask during training so the model learns to respect soft boundaries.

Evaluation will cover three complementary axes:

- **Spatial Precision:** Mean Intersection over Union (mIoU) between the empirical change map (computed via absolute pixel differences) and the ground-truth mask, alongside foreground/background PSNR to quantify undesired modifications.
- **Instruction Compliance:** CLIP-score between the edited image and the instruction, as well as text-image retrieval accuracy when ranking edits by relevance.
- **User Study:** A survey of 10–15 participants who rate (1) whether the edit stayed within the requested region and (2) whether it satisfied the instruction semantically. I will report mean Likert ratings and inter-rater agreement.

All experiments will run on publicly available datasets and will be released with scripts for reproducibility.

6. Expected Contributions

- A mask-conditioned extension of InstructPix2Pix that demonstrably improves edit localization.
- A reusable data-generation pipeline that pairs synthetic instructions with SAM-derived masks for instruction-driven editing tasks.
- Ablation studies and human evaluations that clarify when spatial conditioning helps or hurts diffusion-based editors.
- An open-source implementation and pretrained weights to accelerate subsequent research.

7. Timeline

- **Week 1 (Nov 7 – Nov 14):** Complete literature review, set up the codebase, and benchmark the original InstructPix2Pix model.
- **Week 2 (Nov 15 – Nov 21):** Implement the automated mask-generation pipeline, validate mask quality, and finalize data preprocessing scripts.
- **Week 3 (Nov 22 – Nov 28):** Integrate mask conditioning into the diffusion model, run ablations over conditioning strategies, and launch fine-tuning jobs.
- **Week 4 (Nov 29 – Dec 2):** Finish training, execute the evaluation suite and human study, create qualitative figures, and draft the final report plus code release.

8. Risks and Mitigation

Automatic masks may occasionally misalign with the textual instruction, which could confuse the model. To mitigate this, I will manually vet a validation subset and discard problematic samples. Computational cost is another risk; if full fine-tuning proves too heavy, I will fall back to low-rank adaptation (LoRA) layers applied to the U-Net while keeping the base weights frozen.

References

[1] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. InstructPix2Pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 1

[2] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2