

Mask-Guided, Training-Free Spatial Control for InstructPix2Pix

10-minute presentation notes (CS 5404)

Talk Flow (10 min)

- Motivation & problem (1.5 min)
- Method: Mask-Blended Inference (2 min)
- Implementation details (1.5 min)
- Experiments & results (3 min)
- Discussion, limitations, next steps (1.5 min)
- Wrap-up & Q&A buffer (0.5 min)

Motivation (1.5 min)

- Instruction-following editors (InstructPix2Pix) can bleed edits into background; need spatial control.
- Training mask-aware diffusion is heavy; goal is training-free spatial fidelity using existing checkpoints.
- Chosen base papers: InstructPix2Pix (Brooks et al. 2023) + Segment Anything (Kirillov et al. 2023).

Method: Mask-Blended Inference (2 min)

- Pipeline: SAM click \rightarrow binary mask; run InstructPix2Pix; blend output with original using mask (pixel space).
- Formula: $\tilde{y} = M \odot y_{\text{baseline}} + (1 - M) \odot x_0$.
- Deterministic guidance: no weight changes, just post-process blending; seeds fixed for reproducibility.
- Mask resize with nearest neighbor to preserve crisp boundaries; normalization to $[0,1]$ avoids overflow.

Implementation (1.5 min)

- Code: `submission/code/project_run.py` (inference), `evaluate_metrics.py` (mIoU, CLIP).
- Data: three images in `data/images/` (shirted woman, dog, car); checkpoints via `download_models.sh`.
- Execution: `bash run_pipeline.sh` (installs, runs inference, then metrics). Optional `clean_run.sh` to reset.
- Environment used for final run: Colab, NVIDIA A100 (40 GB), 2 minutes after downloads.

Experiments & Results (3 min)

- Cases (Table in report): shirt recolor (woman), dog to robot, car to hovercraft; 20 denoising steps, guidance 1.5.
- Quantitative (from Colab run):
 - Shirt → red leather: mIoU 0.183, CLIP 0.206
 - Dog → playful robot: mIoU 0.265, CLIP 0.261
 - Car → glowing hovercraft: mIoU 0.173, CLIP 0.221
- Interpretation: dog scores highest (clear silhouette); car shows spatial containment but semantic gap from ambitious styling.
- Qualitative: show rows of (original, mask, baseline, ours). Note background preservation vs. baseline bleed.

Discussion, Limitations, Next Steps (1.5 min)

- Strengths: training-free, quick to reproduce, deterministic seeds, minimal code changes.
- Limitations: blending after sampling (does not steer denoising trajectory); depends on SAM mask quality; heavy checkpoints.
- Next steps: explore latent/attention masking, mask dilations/soft edges, optional human study on spatial fidelity, batch more cases.

Reproducibility Notes

- Commands: `bash download_models.sh`; `bash run_pipeline.sh`.
- Required assets: SAM checkpoint in `models/sam_vit_h_4b8939.pth`; sample images already in repo.
- Results/figures embedded in report under `submission/report/figures/`.
- Code repo: <https://github.com/abrahamyirga/Vision-Project>

Q&A Prompts

- If asked about compute: A100 run; 2 min for three cases after downloads.
- If asked about metrics: automatic mIoU/CLIP provided; human study planned but not executed.
- If asked about training: none performed; method is inference-time only.