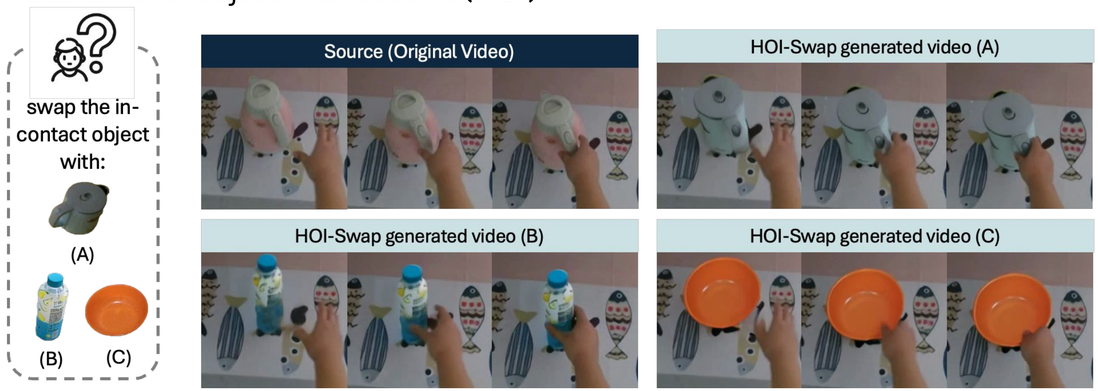
HOI-Swap: Swapping Objects in Videos with Hand-Object Interaction Awareness

Zihui (Sherry) Xue[1,2], Mi Luo[1], Changan Chen[1], Kristen Grauman[1,2]
[1] UT Austin [2] FAIR, Meta AI

NEURAL INFORMATION PROCESSING SYSTEMS

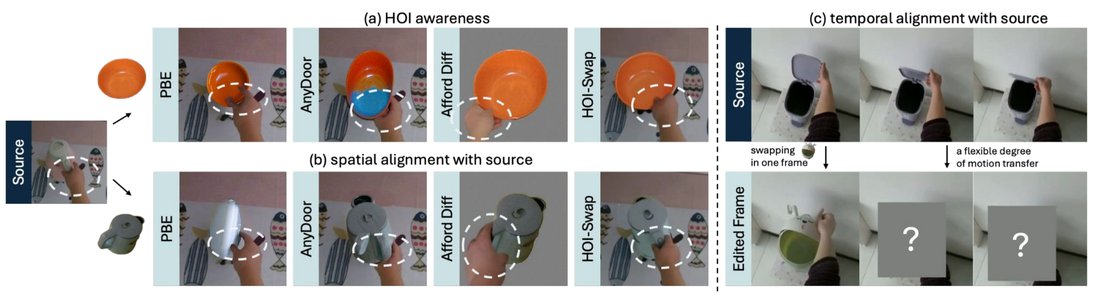
See our website for data, code & qualitative videos →

Key idea: we present HOI-Swap that seamlessly swaps the in-contact object in videos using a reference object image, producing precise video edits with natural hand-object interactions (HOI).

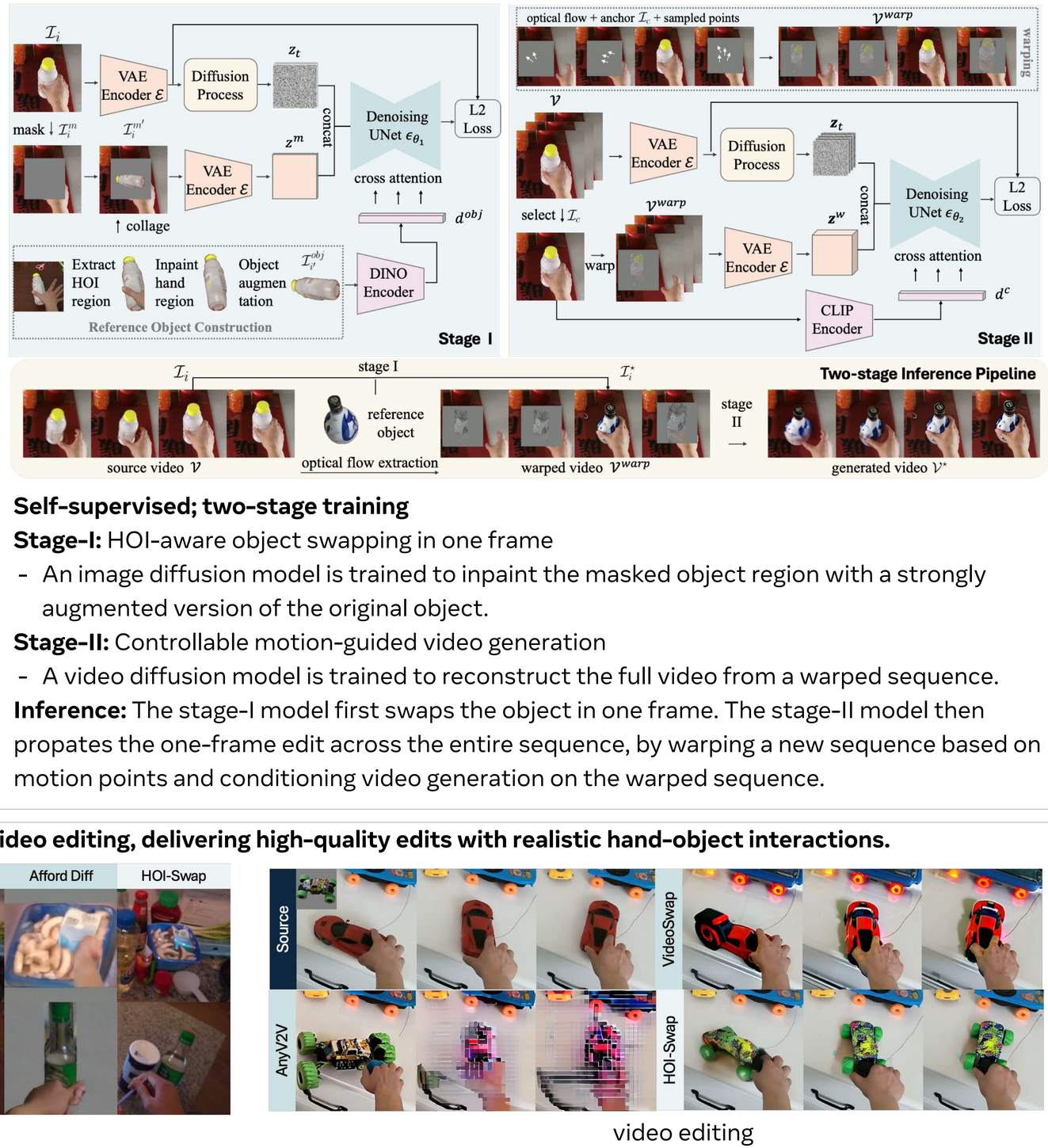


Challenges for the in-contact object swapping problem:

(a) HOI awareness → adjust the grasp patterns to accommodate HOIs
(b) spatial alignment with source → automatically reorient objects
(c) temporal alignment with source → controllable motion guidance



Framework



Self-supervised; two-stage training

Stage-I: HOI-aware object swapping in one frame
- An image diffusion model is trained to inpaint the masked object region with a strongly augmented version of the original object.

Stage-II: Controllable motion-guided video generation
- A video diffusion model is trained to reconstruct the full video from a warped sequence.

Inference: The stage-I model first swaps the object in one frame. The stage-II model then propates the one-frame edit across the entire sequence, by warping a new sequence based on motion points and conditioning video generation on the warped sequence.

HOI-Swap surpasses SOTA editing approaches for both image and video editing, delivering high-quality edits with realistic hand-object interactions.

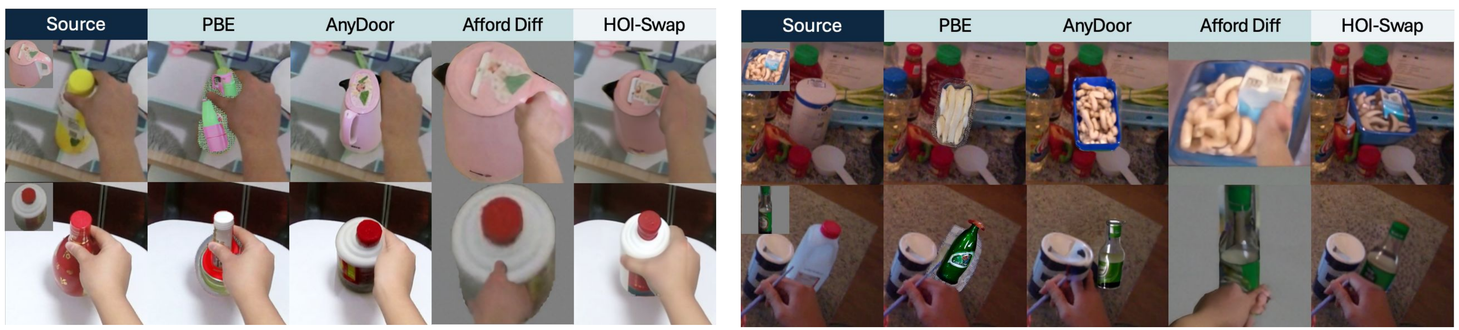


image editing

video editing