ROBOEDIT: REFERENCE-GUIDED LOCAL VIDEO EDITING FOR ROBOTIC SCENES

Anonymous authors
Paper under double-blind review

ABSTRACT

Robotic manipulation videos are costly to collect and often tied to a specific embodiment. This embodiment-specific coupling limits reuse across robot platforms. We study reference-guided local video editing as a way to repurpose such demonstrations by replacing manipulators or objects while preserving the surrounding context. Since paired cross-embodiment videos are rarely available, we construct pseudo-paired supervision by extracting masks and reference images from the source video itself. However, this construction introduces a train-test mismatch: at training time, the reference is a crop of the source manipulator and is therefore tightly coupled to the foreground, whereas at inference, the reference is a novel image of a different manipulator. This mismatch can lead to reference leakage, boundary artifacts, and original-identity ghosting. We propose RoboEdit, a mask-aware consistency learning framework that localizes reference-induced changes to the editable region. RoboEdit compares reference-conditioned and null-reference forward passes and suppresses reference-induced changes in outside-mask token relations. We derive this objective from a flow-matching view of vector-field decoupling and implement it efficiently with token-level reservoir sampling. To evaluate this setting, we curate a robotic local editing benchmark and introduce Robot Manipulation Consistency (RMC), a task-specific metric for measuring local stability, trajectory preservation, and edited-region semantic consistency. Experiments show that RoboEdit improves reference fidelity, scene preservation, and manipulation consistency over existing video editing baselines; downstream evaluations further indicate that the edited videos improve simulation representation transfer and provide preliminary gains on a real OpenArm platform.

1 INTRODUCTION

Robotic manipulation videos provide a valuable source of supervision for embodied intelligence Brohan et al. (2023); Open X-Embodiment Collaboration et al. (2023); Black et al. (2024). These demonstrations encapsulate rich spatio-temporal dynamics, including precise robot kinematics, nuanced object interactions, and evolving scene contexts. However, scaling the collection of such high-quality videos is prohibitively expensive, and the resulting data is inherently coupled to a specific robot embodiment. This embodiment-specific coupling makes it difficult to reuse existing manipulation videos across robot platforms and deployment scenarios Open X-Embodiment Collaboration et al. (2023). The ability to replace the manipulator in a video while preserving the scene, the manipulated object, and the original motion dynamics would offer an appealing way to synthesize new embodied visual data from existing demonstrations. It would also allow existing demonstrations to be repurposed across robot embodiments, complementing recent efforts that leverage video generative modeling and large-scale visual pretraining for manipulation Wu et al. (2023); Jing et al. (2023).

Recent video editing methods have made impressive progress in text-guided generation, stylization, and general-purpose video-to-video editing Wu et al. (2023); Molad et al. (2023); Qi et al. (2023); Liu et al. (2024); Geyer et al. (2024); Kara et al. (2024); Ku et al. (2024); Shin et al. (2025); Jiang et al. (2025); Wei et al. (2025). However, robotic local replacement presents a distinct setting. First, the desired edit is spatially localized: the goal is to replace a specified target, such as the manipulator or a manipulated object, while keeping the rest of the video faithful to the source. Second, linguistic descriptions are inherently insufficient to capture the intricate geometric and textural details of robotic manipulators; hence, target appearance must be governed by precise image references. This paradigm

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

also aligns with recent video editing frameworks that move beyond purely text-driven control and support image- or subject-guided editing Ku et al. (2024); Jiang et al. (2025); Wei et al. (2025). Third, the non-edited regions are not merely passive background; they often include nearby objects, support surfaces, and contact cues that are essential to the task itself. Therefore, successful replacement requires not only accurate appearance transfer, but also spatially confined editing and preservation of action-relevant scene content. In practice, however, directly applying off-the-shelf editing methods often leads to two dominant failure modes: reference transfer remains superficial, changing color or texture while leaving the target geometry largely unchanged; and local edits corrupt action-relevant regions, producing artifacts near contact boundaries or unintentionally altering nearby objects, support surfaces, and background structures.

To address these challenges, we propose RoboEdit, a mask-aware consistency learning framework tailored for reference-guided local video editing in robotic scenes. At its core, RoboEdit enforces a structural bottleneck that constrains where the reference condition can take effect: it should drive the replacement inside the editable region while leaving the surrounding scene unchanged. To encourage this behavior, RoboEdit introduces a token-level correlation regularizer. During training, we compare two forward passes that share the same editing conditions and differ only in their reference images. This formulation measures how variations in the reference image perturb token relations between the editable region and the unedited global context. For efficiency, instead of forming dense correlations for all masked tokens, we sample a small set of query tokens from the editable region and compute their correlations with visual tokens across the video, yielding a compact estimate of reference-induced relational changes. Since relations to non-editable tokens should remain stable when the reference changes, the regularizer penalizes discrepancies only in outside-mask correlations and treats the aligned pass as a stop-gradient anchor. Finally, we train RoboEdit with this token-level regularizer together with localized reconstruction and masked temporal consistency objectives, encouraging edits that are locally faithful, temporally stable, and spatially confined. Ideally, optimizing such a framework demands strictly paired source-target videos that share identical tasks, scenes, and motion trajectories, differing solely in the targeted embodiments. Given the scarcity of such paired data in the robotics community, we circumvent this bottleneck by constructing pseudo-paired supervision directly from existing single-embodiment demonstrations. For evaluation, we build a curated benchmark covering manipulator and object replacement, and we introduce Robot Manipulation Consistency (RMC) to measure local stability, trajectory preservation, and edited-region semantic consistency.

Our contributions are summarized as follows:

1. We formulate cross-embodiment robotic video repurposing as reference-guided local editing, where a single target reference specifies a manipulator or object replacement while the source scene, motion, and interaction context are preserved.
2. We propose RoboEdit, a mask-aware consistency learning framework that combines token-level correlation regularization with localized reconstruction and temporal consistency to suppress reference leakage outside the editable region.
3. We construct a scalable pseudo-paired data generation pipeline with mask and reference quality filtering, enabling training from existing single-embodiment manipulation videos without paired cross-robot recordings.
4. We evaluate RoboEdit with visual metrics, Robot Manipulation Consistency (RMC), a blinded human preference study, simulation downstream control transfer, and a preliminary real-robot OpenArm evaluation.

2 Related Work

Reference-guided video editing and video customization. Diffusion-based video editing has shown strong capabilities in text-guided editing and general video-to-video transformation Wu et al. (2023); Molad et al. (2023); Qi et al. (2023); Liu et al. (2024); Geyer et al. (2024); Kara et al. (2024). However, text-based instructions alone often cannot specify fine-grained target appearance, especially instance-specific shape, texture, or geometry. To improve appearance control beyond text, recent methods introduce image references into video generation and editing. Early personalization approaches learn subject concepts through textual embeddings or model fine-tuning Gal et al. (2023); Ruiz et al. (2023), while more recent approaches use image adapters, reference encoders, or ControlNet-style conditions to inject target appearance more efficiently Ye et al. (2023); Wei et al. (2024); Chefer et al. (2024); Wu et al. (2025); Jiang et al. (2024); Wei et al. (2024). Building on this

trend, HunyuanCustom and OmniVCus further combine reference images with multimodal control signals such as masks, depth, camera trajectories, and text instructions, improving controllability and subject consistency in customized video generation Hu et al. (2025); Cai et al. (2025).

Another line of work extends video foundation models into unified generation and editing systems, including VACE, other Wan-style editing variants, and UniVideo Jiang et al. (2025); Wan Team et al. (2025); Wei et al. (2025). Although these systems support tasks such as reference-guided generation, video-to-video editing, and masked editing, they are not specifically designed for contact-sensitive local replacement in robotic manipulation videos. RoboEdit targets this setting by constraining reference-induced changes to the editable region while preserving nearby objects, contact regions, support surfaces, and motion.

Robot swapping and embodiment transfer. A related line of work transfers visual demonstrations across embodiments for robot learning, including cross-painting robot arms in shared environments (Chen et al., 2024), translating human interaction plans into robot task scenes (Bharadhwaj et al., 2024), human-to-robot data augmentation (Li et al., 2025), and cross-embodiment imitation through intermediate tools (Seo et al., 2025). These methods reduce embodiment-specific visual gaps, but their assumptions differ from our single-reference local editing setting. H2R (Li et al., 2025) uses simulator assets, retargeting, rendering, and compositing for a predefined target robot, while RoboSwap (Bai et al., 2025) learns a separate translator for each source–target robot pair. RoboEdit instead specifies the target appearance with a single reference image and directly edits the masked region, enabling new target embodiments and object replacements without target-domain data, pair-specific translators, or calibrated simulation assets.

3 Data

3.1 Scalable Data Generation Pipeline

A primary bottleneck in reference-guided robotic video editing is the scarcity of paired cross-embodiment videos: two clips rarely share the same task, scene, object, and motion trajectory while differing only in the manipulator. We therefore construct pseudo-paired supervision from existing single-embodiment demonstrations. Each training sample is represented as a tuple (V, M, R, p), where V is a source video, M is a framewise edit mask, R is a reference image specifying the target appearance, and p is a clip-level prompt.

Source videos. We aggregate clips from multiple open-source robotic manipulation datasets, primarily from Open X-Embodiment (OXE) Open X-Embodiment Collaboration et al. (2023), including BC-Z Jang et al. (2022), BridgeData V2 Walke et al. (2023), Berkeley AUTOLAB UR5 Chen et al. (2023), Berkeley FANUC Manipulation Zhu et al. (2023), Columbia CAIR Lab PushT Real Chi et al. (2023), Stanford Hydra Belkhale et al. (2023), and TOTO Zhou et al. (2023). We convert RLDS-style trajectories into coherent video clips and filter out low-quality or overly short sequences.

Masks, references, and prompts. For each clip, we use SAM3 Carion et al. (2025) to segment the target manipulator or object, producing framewise masks M and candidate reference crops. Since the target may be occluded or only partially visible in some frames, we rank crops by mask-area ratio and use Qwen3-VL Bai et al. (2025) to select the most complete candidate, which is composited onto a white background as the reference image R. The text prompt p is generated from task metadata when available; otherwise, Qwen3-VL captions the video directly. All prompts are normalized into an action-centric format describing the manipulator, object, and action when visually identifiable.

Pseudo-paired supervision. During training, R is extracted from the source video itself, yielding self-replacement supervision. This setup is particularly vulnerable to shortcut learning in robotic manipulation scenes Hu et al. (2025); Cai et al. (2025); Xie et al. (2024). Unlike human-centric editing, where moderate variation in body shape or pose does not usually determine identity from silhouette alone, robot manipulators often exhibit limited embodiment diversity and highly distinctive geometric contours. As a result, the edit mask itself can leak strong embodiment cues, allowing the model to infer the target manipulator directly from the masked shape and degenerate toward copy-paste behavior. To mitigate this issue, we dilate the edit mask and apply geometric perturbations, including translation and rotation, to the extracted reference foreground. These perturbations weaken the direct correspondence between the source-region shape and the reference, making pseudo-paired training a closer proxy to reference-guided editing at inference time. Figure 1 summarizes the pipeline.

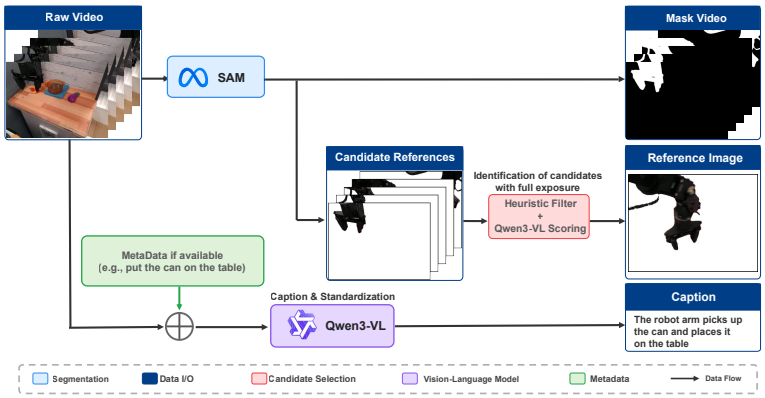

Figure 1: Overview of the data generation pipeline. Starting from raw robotic manipulation videos, we use SAM3 to obtain mask videos and candidate reference images. Candidates are first ranked by heuristic filters, then filtered by Qwen3-VL to identify the most complete reference view. For captioning, we provide Qwen3-VL with the raw video together with the task metadata, when available, to produce a standardized caption.

4 METHOD

We study reference-guided local video editing for robotic manipulation scenes. Our framework operates in a pseudo-paired training setting, where the reference image specifying the target manipulator is extracted directly from the source video. This avoids the need for paired cross-embodiment videos, but also entangles the dynamic generation trajectories of the foreground and background. We first formalize this distribution shift within the continuous-time Flow Matching framework (Sec. 4.1). We then formalize an invariance objective and give sufficient conditions under which an outside-mask structural anchor enforces it (Sec. 4.2). We bridge this analysis to an efficient discrete implementation using a token-level reservoir sampling strategy within the MMDiT architecture (Sec. 4.3), and conclude with supporting reconstruction and temporal objectives (Sec. 4.4).

4.1 FLOW MATCHING AND BIDIRECTIONAL ENTANGLEMENT

Let $Z \in \mathbb{R}^{N \times C}$ denote the video latent tokens, partitioned by an expanded spatial mask M into the foreground interaction region Z_{in} (containing the robot arm and the interactive object) and the unmasked environment Z_{out}. In practice, we apply a morphological expansion to the arm's mask to ensure the interactive object is fully absorbed into the generation region, allowing the model to adapt the physical interaction pattern to the novel reference arm.

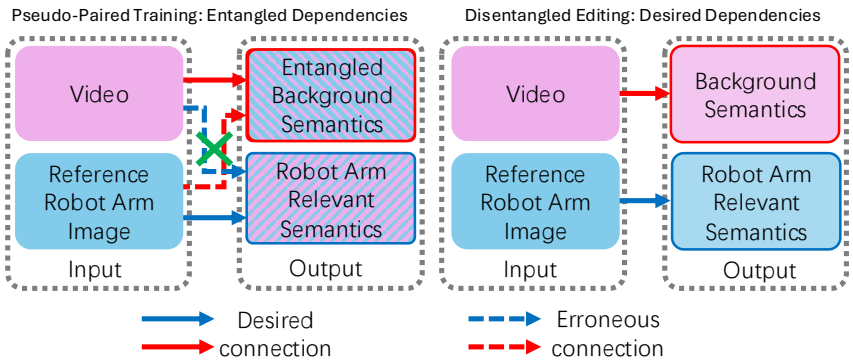

Figure 2: **Bidirectional Trajectory Entanglement.** In standard reference-guided editing (left), the source video and reference image leak information across modalities, causing the background to be corrupted by the new reference (red error connection) and the novel object to ghost original video semantics (blue error connection). Our method aims to sever these gradients to achieve disentangled local editing (right).

Our generative process is governed by a conditional vector field $v_\theta(z_t, t, V, R, M, p)$, which transports noise to data conditioned on the source video V, reference image R, mask M, and clip-level

prompt p. We write v_θ^{in} and v_θ^{out} for the components of v_θ acting on inside- and outside-mask tokens, respectively. The pseudo-paired training setup makes the reference image R and the original foreground interaction region V_{in} strongly correlated. At inference, however, the model is required to edit based on a novel, out-of-distribution reference R'.

As illustrated in Figure 2, this train–test mismatch leads to **Bidirectional Trajectory Entanglement**:

1. **Background Corruption:** R' heavily influences $\frac{dz_{out}}{dt}$, causing the new arm's features to bleed into the environment (red erroneous connection).
2. **Object Ghosting:** The original video's semantics interfere with $\frac{dz_{in}}{dt}$, preventing the novel R' from exclusively controlling the interaction identity (blue erroneous connection).

We therefore seek a vector field in which the environment is independent of R' and the manipulated interaction region is exclusively driven by R'.

4.2 Anchor-Mediated Reference Invariance

We address the entanglement above by training the model to make an outside-mask structural quantity invariant to reference change. Throughout this section we apply explicit input masking, so the network observes $V^{out} = V \odot M_{out}$ rather than the full source video. Consequently, holding z_t and M fixed, v_θ^{in} has no direct computational dependence on V_{in}, and $\nabla_{V_{in}} v_\theta^{in} = 0$ holds by construction.

To prevent reference leakage into the background, we extract two anchors per training step: a *reference-conditioned* anchor $A_\theta(z_t, t, c_R)$ from a forward pass with the true paired reference, and a *null-reference* anchor $A_\theta(z_t, t, c_\emptyset)$ from a forward pass with R replaced by $\emptyset$, where $c_R = (V^{out}, R, M, p)$ and $c_\emptyset = (V^{out}, \emptyset, M, p)$. We instantiate the anchor as $A_\theta(z_t, t, c) = W_{out} \odot \Phi_\theta(z_t, t, c)$, where Φ_θ is a token-correlation representation extracted from the network (concretely, a cross-Gram matrix; see Sec. 4.3). The reference-conditioned anchor is detached (stop-gradient) and serves as the alignment target for the null-reference one.

Lemma 1 (Anchor-Mediated Reference Invariance)**.** *Consider the reference-conditioned and null-reference outside-mask anchors $A_\theta(z_t, t, c_R)$ and $A_\theta(z_t, t, c_\emptyset)$. Under a reference-free null branch, an outside-mask anchor bottleneck, differentiability, and exact anchor alignment in the training limit, the outside-mask anchor and vector field become locally invariant to the reference: $\nabla_R A_\theta(z_t, t, c_R) = 0$ and $\nabla_R v_\theta^{\text{out}}(z_t, t, c_R) = 0$ almost everywhere.*

The proof and assumptions are given in Appendix B. This result should be read as an idealized justification for the regularizer rather than an architectural guarantee: MMDiT attention still permits bypass paths, so our objective encourages, but does not prove, exact factorization. Empirically, the reduced outside-mask correlation discrepancy in Sec. 5.5 and Appendix L indicates that the regularizer suppresses the dominant reference-dependent leakage paths.

4.3 Efficient Implementation via DiT Reservoir Sampling

Lemma 1 is agnostic to the specific choice of Φ_θ. Inspired by prior work on Gram matrices Simeoni et al. (2025); Yan et al. (2026), we instantiate Φ_θ within the MMDiT backbone as a token-level cross-Gram statistic, $G_t = Z_{\text{in}}(t) Z_{\text{out}}(t)^\top$, extracted from a chosen transformer layer. This statistic serves as a tractable low-order surrogate for cross-region feature interactions, capturing boundary relations between the editable foreground and the unmasked environment. Although it does not guarantee complete control of v_θ^{out}, Sec. 5.5 and Appendix L empirically show that reducing this Gram discrepancy coincides with reduced outside-mask leakage and fewer boundary artifacts.

Evaluating the dense query-to-all matrix across all spatial-temporal tokens is computationally prohibitive. Inspired by classical reservoir sampling for streaming uniform subsampling Vitter (1985), we maintain a fixed-size reservoir S_K of K query tokens drawn from the masked region. Streaming the sequence of masked tokens, the first K tokens deterministically populate the reservoir; thereafter, the token at stream index $i > K$ is accepted with probability K/i, replacing a uniformly chosen reservoir element when accepted. Crucially, the same sampled query indices are used for both the reference-conditioned and null-reference passes, so the comparison is over identical token positions. The sampled correlation matrix should be viewed as a random row-submatrix of the full query-to-token correlation matrix. What is strictly unbiased is the scalar row-averaged loss computed over these sampled rows.

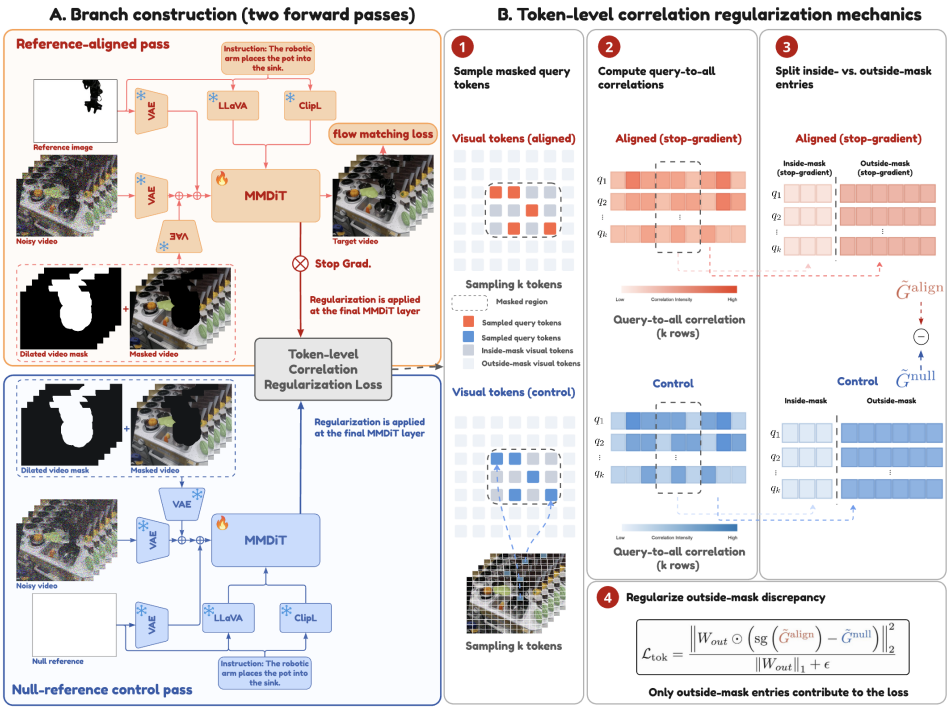

Figure 3: Overview of the proposed token-level correlation regularization. During training, we perform two forward passes: a reference-aligned pass and a null-reference control pass. The aligned pass is optimized with the standard flow-matching loss, whereas the regularization is computed from token features extracted from the final MMDiT layer in both passes. Panel B illustrates the regularization mechanism. We first sample masked query tokens from the edit region and compute query-to-all correlations with all visual video tokens, yielding one correlation matrix per pass. After splitting each matrix into inside-mask and outside-mask entries, we compute the regularization loss using only the outside-mask entries, with the aligned pass serving as a stop-gradient anchor.

Proposition 1 (Unbiased Sampled Estimator of the Token-Correlation Loss)**.** *Let I denote the inside-mask query tokens ($|I| = N_{\text{in}}$) and O the outside-mask tokens. For each inside token $i \in I$, define the row loss*

$$\ell_i = \frac{1}{|O| + \epsilon} \sum_{j \in O} \left(G_{ij}^{\text{align}} - G_{ij}^{\text{null}} \right)^2,$$

and the exact outside-mask token-correlation loss $L_{\text{exact}} = \frac{1}{N_{\text{in}}} \sum_{i \in I} \ell_i$. Let $S_K \subset I$ be a reservoir sample of K inside-mask tokens drawn uniformly without replacement and set $\widehat{L}_K = \frac{1}{K} \sum_{i \in S_K} \ell_i$. Then, conditioned on the token features, $\mathbb{E}_{S_K}[\widehat{L}_K] = L_{\text{exact}}$ and

$$\text{Var}_{S_K}(\widehat{L}_K) = \left(1 - \frac{K}{N_{\text{in}}}\right) \frac{\sigma_\ell^2}{K} = \mathcal{O}(1/K),$$

where $\sigma_\ell^2 = \frac{1}{N_{\text{in}} - 1} \sum_{i \in I} (\ell_i - L_{\text{exact}})^2$.

The proof is in Appendix C, which further shows that $\nabla_\theta \widehat{L}_K$ is unbiased for $\nabla_\theta L_{\text{exact}}$, so the dense-to-compact substitution is faithful under SGD training. The row-loss formulation is what makes this clean: a naive submatrix Frobenius-norm estimator with subsampled normalizer would be a ratio of sample means, and generally biased. Let $\tilde{G}_t^{\text{align}}, \tilde{G}_t^{\text{null}} \in \mathbb{R}^{K \times |O|}$ denote the reservoir-sampled cross-Gram matrices from the reference-conditioned and null-reference passes, sharing the same sampled query indices, and let W_{out} be the outside-mask weighting introduced in Sec. 4.2. Approximating the continuous-time integration via stochastic sampling $t \sim \mathcal{U}(0, 1)$, the empirical token-level regularizer takes the form

$$\mathcal{L}_{\text{tok}} = \mathbb{E}_t \left[\frac{\| W_{out} \odot (\text{sg}(\tilde{G}_t^{\text{align}}) - \tilde{G}_t^{\text{null}}) \|_F^2}{\| W_{out} \|_1 + \epsilon} \right] \tag{1}$$

which directly penalizes reference-dependent variation of outside-mask token relations relative to the stop-gradient aligned pass.

4.4 Overall Training Objective

We train the backbone with a primary flow-matching loss augmented by our structural regularizer and two auxiliary constraints (details in Appendix A).

Masked Flow Matching. To align the generated manipulator with the identity of R, we adopt a mask-weighted conditional flow-matching loss supervised by $u_t = z_1 - z_0$: $\mathcal{L}_{\text{fm}} = \mathbb{E}_{t,z_0,z_1}\left[\|(1+\alpha M)\odot(u_t - v_\theta(z_t,t,V,R,M,p))\|_2^2\right]$. where α upweights the edited region.

Token-Level Correlation (Gram Loss). Our token-level regularizer $\mathcal{L}_{\text{tok}}$ (Sec. 4.3) targets the structural entanglement induced by reference conditioning by preserving boundary interactions between the edited region and its surroundings.

Auxiliary Objectives. A masked L_1 reconstruction loss $\mathcal{L}_{\text{rec}}$ improves local precision, and a temporal consistency loss $\mathcal{L}_{\text{temp}}$ suppresses frame-to-frame jitter.

Full Objective: $\mathcal{L} = \mathcal{L}_{\text{fm}} + \lambda_{\text{tok}}\mathcal{L}_{\text{tok}} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{temp}}\mathcal{L}_{\text{temp}}$, where the λ weights balance local accuracy and temporal stability against the structural regularizer.

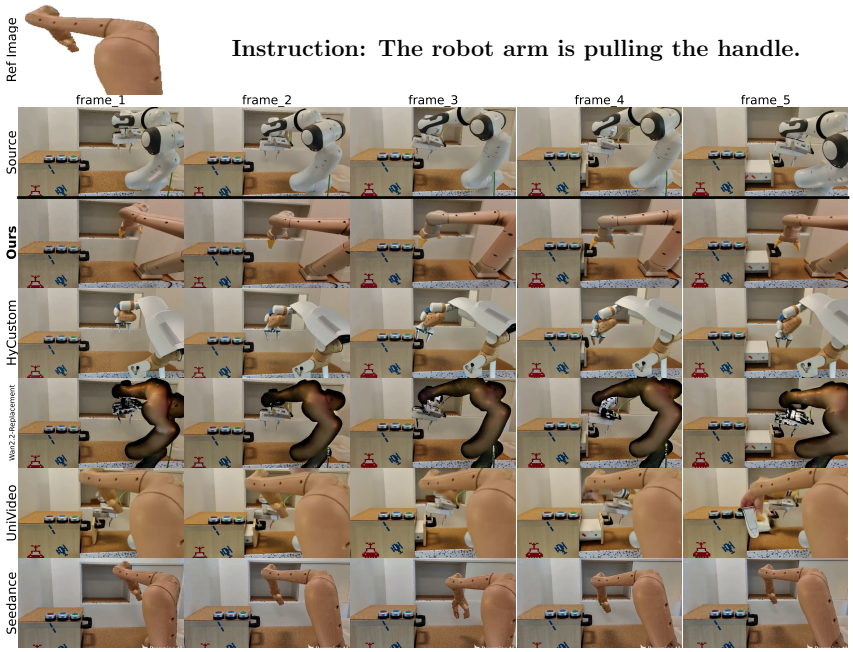

Figure 4: RoboEdit replaces manipulators while preserving surrounding content and motion dynamics, outperforming several existing video editing methods in this setting.

5 Experiments

5.1 Experimental Setup

In our experiments, we adopt HunyuanCustom as the base video generation model and fine-tune it with LoRA for reference-guided local video editing. Training is conducted on a curated dataset of approximately 57K video-caption pairs. Unless otherwise specified, all experiments are conducted on 16 NVIDIA H100 GPUs (2 nodes) with FSDP for distributed training. Due to the high memory cost of processing 64-frame clips at a resolution of 320×576, we employ sequence parallelism with a group size of 8, such that each training sample is partitioned across 8 GPUs along the spatio-temporal token dimension. This results in an effective global batch size of 2. We optimize the model using AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay 0.01 for 5 epochs. The learning rate is initialized at 1×10^{-4} and decayed to 1×10^{-6} with a cosine schedule. The training objective consists of the original flow-matching loss, two auxiliary masked losses, and the proposed token-level correlation regularization term. For the token-level term, we extract token representations from the final transformer layer, project them into a 128-dimensional subspace using a fixed random projection,

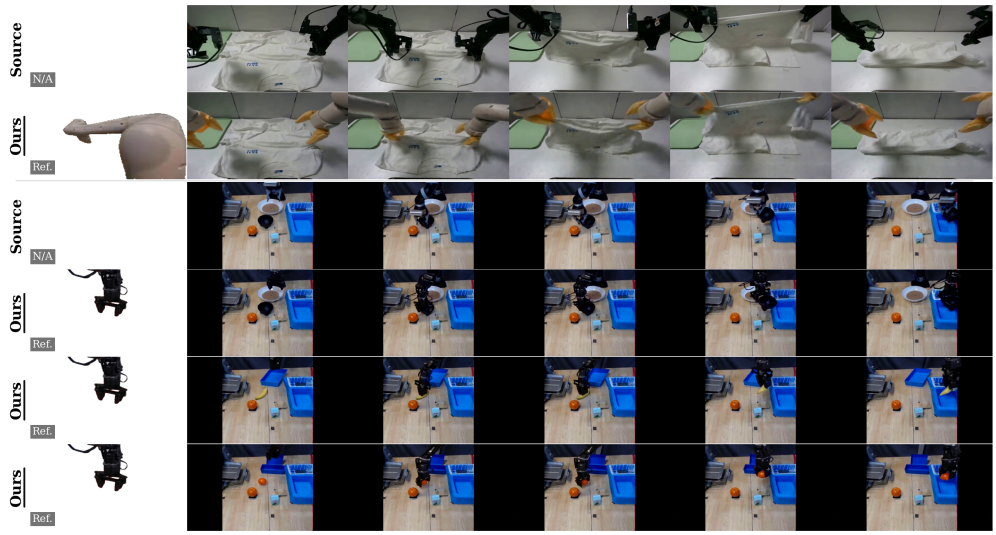

Figure 5: Qualitative results of RoboEdit on robotic manipulation videos. RoboEdit handles both manipulator replacement and object replacement (like replacing the black bowl with a banana or an orange) while preserving surrounding scenes. Results are best viewed zoomed in.

and select 4 query tokens within the masked region. We then compute query-to-all correlation rows and penalize discrepancies on tokens outside the edit mask, which helps preserve the structure of unedited regions while allowing flexible generation inside the edited area. Detailed settings of the loss design are provided in Appendix E.

5.2 Qualitative Results

Table 1: Quantitative comparison between RoboEdit and HyCustom.

Model	CLIP ↑	DINO-Sim ↑	Temp-Consis ↑	RMC ↑	Scene Pres. ↑	Ref. Fid. ↑
RoboEdit	**0.3050**	**0.2914**	**0.9163**	**0.3426**	**7.28**	**7.01**
HyCustom	0.3017	0.2248	0.9100	0.3122	6.45	5.87

Table 2: Objective ablation of RoboEdit.

Variant	CLIP ↑	DINO-Sim ↑	Temp-Consis ↑	RMC ↑	Scene Pres. ↑	Ref. Fid. ↑
Full	**0.3050**	**0.2914**	0.9163	**0.3426**	**7.28**	**7.01**
w/o $\mathcal{L}_{\text{tok}}$	0.3042	0.2318	**0.9184**	0.3200	6.56	6.19
w/o $\mathcal{L}_{\text{rec}}$	0.2912	0.2223	0.9152	0.3136	6.20	6.30
w/o $\mathcal{L}_{\text{temp}}$	0.3037	0.2802	0.9114	0.3311	7.01	6.54

We qualitatively evaluate RoboEdit on diverse local editing tasks, including manipulator and object replacement, as shown in Figure 4 and Figure 5. Additional results are provided in Appendix I. Across several examples, RoboEdit produces high-quality edits with strong local fidelity, clean background preservation, and stable temporal consistency. We further compare RoboEdit with several representative video editing baselines; detailed qualitative results are provided in Appendix J.

5.3 Quantitative Results

We introduce, to the best of our knowledge, the first benchmark dedicated to local video editing in robotic scenes. Our test dataset generation follows the training data pipeline but uses out-of-distribution videos to evaluate generalization. Each masked video is paired with two to three suitable robot-arm references and manually verified, yielding 335 high-quality video–reference pairs. Following prior work, we evaluate model performance on this benchmark using (i) reference-image similarity (DINO-Sim), (ii) text–video alignment (CLIP-B-T), and (iii) temporal consistency (Temp-Consis). In addition, we propose Robot Manipulation Consistency (RMC), a new metric that specifically measures manipulation faithfulness in robot videos by jointly assessing temporal mask stability, trajectory preservation with respect to the source video, and semantic consistency in the

edited robot region. Detailed implementation is provided in Appendix F. Since the above metrics do not directly capture reference leakage into non-edited regions or local scene corruption, we further adopt a multimodal judge to assess scene preservation and reference fidelity. Details are provided in Appendix G. Table 1 summarizes the comparison results between RoboEdit and HunyuanCustom. UniVideo and Wan2.2-Replace are excluded due to their high edit failure rates. In addition, a blinded four-way human preference study on complete videos gives RoboEdit 125/150 votes (83.3%), with RoboEdit preferred most often on 9 of 10 video sets ($\chi^2(3) = 273.3$, $p < 0.001$); the full protocol is provided in Appendix H.

5.4 Downstream Control Evaluation

Video-editing metrics alone cannot determine whether the edited videos are useful for robot learning. We therefore evaluate downstream control under a fixed representation-transfer protocol following H2R (Li et al., 2025). For each video source, we start from an ImageNet-pretrained MAE encoder (He et al., 2022), continue pretraining it on the corresponding robot-video frames, freeze the encoder, and train the same behavior-cloning head on the same target demonstrations. Thus, differences in task success reflect the transferability of the learned visual representation rather than changes in the policy or action data. As shown in Table 3, RoboEdit achieves the best average success across nine simulated Robomimic and MimicGen tasks and three source embodiments, reaching 41.1% compared with 29.3% for unedited videos, 30.1% for HyCustom, 33.2% for Mirage-like, and 32.6% for Ro-Aug. We also report a preliminary real-robot evaluation on a dual-arm OpenArm platform, where AIST/ALOHA videos (Motoda et al., 2025) are edited into the OpenArm appearance and used for the same frozen-encoder BC pipeline. RoboEdit reaches 65.0% average success, compared with 53.3% for HyCustom and 40.0% for unedited AIST videos. Full task-level results and real-robot protocol details are provided in Appendix M and Appendix N.

Table 3: Downstream control evaluation.

Method (Simulation transfer to Panda)	Sawyer	IIWA	UR5e	Avg.
Unedited	21.0%	36.2%	30.8%	29.3%
HyCustom	29.6%	30.8%	30.0%	30.1%
Mirage-like	31.7%	35.7%	32.3%	33.2%
Ro-Aug	29.3%	35.0%	33.4%	32.6%
RoboEdit	**40.1%**	**43.0%**	**40.3%**	**41.1%**
MAE pretraining data (Real-robot OpenArm)	Lemon	Banana	Pear	Avg.
Unedited AIST	13/20	11/20	0/20	40.0%
HyCustom AIST→OpenArm	15/20	16/20	1/20	53.3%
RoboEdit AIST→OpenArm	**17/20**	**19/20**	**3/20**	**65.0%**

5.5 Ablation Study

We ablate the training objectives in Table 2. Besides the original ablation of the token-level correlation regularizer, we remove $\mathcal{L}_{\text{rec}}$ and $\mathcal{L}_{\text{temp}}$ while keeping all other settings fixed. Removing $\mathcal{L}_{\text{rec}}$ causes the largest drops in CLIP, DINO, RMC, and scene preservation, indicating its role in accurate local replacement. Removing $\mathcal{L}_{\text{temp}}$ mainly decreases temporal consistency. Removing $\mathcal{L}_{\text{tok}}$ reduces DINO similarity, RMC, scene preservation, and reference fidelity, suggesting weaker confinement of reference-induced changes. Qualitative comparisons and the layer-placement study are provided in Appendix K and Appendix L.

6 Conclusion

In this work, we address the challenge of cross-embodiment robotic video repurposing by framing it as a reference-guided local editing problem. Our goal is to seamlessly replace a manipulator or an interacted object based on a target reference, while strictly preserving the original motion and surrounding scene context. To this end, we present RoboEdit, a mask-aware consistency learning framework that combines token-level correlation regularization with localized reconstruction and temporal consistency objectives to localize the effect of the reference condition to the intended edit region. Experimental results demonstrate that RoboEdit improves reference fidelity, scene preservation, and manipulation consistency over representative video editing baselines. Downstream evaluations further show that RoboEdit-edited videos improve simulation representation transfer across multiple source embodiments and provide preliminary real-robot gains on OpenArm. These results suggest that reference-guided local editing offers a practical path toward synthesizing reusable embodied visual data from existing robotic demonstrations.

AI USE STATEMENT

In this work, we used generative AI tools for dataset construction, evaluation support, and manuscript preparation. Specifically, Qwen3-VL was used to select complete reference crops, reject low-quality mask/reference candidates, and generate or standardize clip-level captions when metadata was unavailable, as described in Section 3 and Appendix D. GPT-5 was used as a multimodal judge to score scene preservation and reference fidelity under the fixed prompt reported in Appendix G. These uses supported dataset cleaning/reformatting and qualitative evaluation. Additionally, generative AI tools were used for language polishing and LaTeX editing assistance. We did not use generative AI tools to autonomously generate research hypotheses, formulate mathematical claims, prove theoretical claims, or draw experimental conclusions. All AI-assisted outputs used in the data pipeline, evaluation, and manuscript preparation were reviewed by the authors. We take responsibility for the final content of this work, including text, claims, artifacts, and results produced with the aid of generative AI.

ETHICS STATEMENT

RoboEdit can create realistic edited robotic videos and therefore may be misused to exaggerate robotic capabilities or create misleading demonstrations. We encourage clear disclosure whenever RoboEdit-generated videos are used in datasets, benchmarks, or demonstrations, and we do not recommend replacing real-world validation with edited videos. Our experiments use publicly available robot datasets and self-collected robot trials. The human preference study involved blinded judgments of edited videos and collected task responses rather than sensitive personal information.

REPRODUCIBILITY STATEMENT

We provide implementation details for the training objectives, model configuration, optimization settings, metric definitions, multimodal-judge protocol, human study, and downstream control evaluation in the appendix. The data construction pipeline, including source datasets, mask/reference filtering, caption generation, and pseudo-paired supervision, is described in Section 3 and Appendix D.

REFERENCES

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil J. Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Andy Zeng, and Pete Florence. RT-1: Robotics Transformer for Real-World Control at Scale. In *Proceedings of Robotics: Science and Systems*, 2023.

Open X-Embodiment Collaboration, Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, and others. Open X-Embodiment: Robotic Learning Datasets and RT-X Models. *arXiv preprint arXiv:2310.08864*, 2023.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, and others. π_0: A Vision-Language-Action Flow Model for General Robot Control. *arXiv preprint arXiv:2410.24164*, 2024.

Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing Large-Scale Video Generative Pre-Training for Visual Robot Manipulation. *arXiv preprint arXiv:2312.13139*, 2023.

Ya Jing, Xuelin Zhu, Xingbin Liu, Qie Sima, Taozheng Yang, Yunhai Feng, and Tao Kong. Exploring Visual Pre-training for Robot Manipulation: Datasets, Models, and Methods. *arXiv preprint arXiv:2308.03620*, 2023.

Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Eyal Molad, Eliahu Horwitz, Dani Valevski, Alex Rav Acha, Yossi Matias, Yael Pritch, Yaniv Leviathan, and Yedid Hoshen. Dreamix: Video Diffusion Models are General Video Editors. *arXiv preprint arXiv:2302.01329*, 2023.

Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. FateZero: Fusing Attentions for Zero-shot Text-based Video Editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-P2P: Video Editing with Cross-attention Control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. TokenFlow: Consistent Diffusion Features for Consistent Video Editing. In *International Conference on Learning Representations*, 2024.

Ozgur Kara, Bariscan Kurtkaya, Hidir Yesiltepe, James M. Rehg, and Pinar Yanardag. RAVE: Randomized Noise Shuffling for Fast and Consistent Video Editing with Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *International Conference on Learning Representations*, 2023.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models. *arXiv preprint arXiv:2308.06721*, 2023.

Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Zhang, Jingren Zhou, and Hongming Shan. DreamVideo: Composing Your Dream Videos with Customized Subject and Motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Hila Chefer, Shiran Zada, Roni Paiss, Ariel Ephrat, Omer Tov, Michael Rubinstein, Lior Wolf, Tali Dekel, Tomer Michaeli, and Inbar Mosseri. Still-Moving: Customized Video Generation without Customized Video Data. *ACM Transactions on Graphics*, 2024.

Tao Wu, Yong Zhang, Xintao Wang, Xianpan Zhou, Guangcong Zheng, Zhongang Qi, Ying Shan, and Xi Li. CustomCrafter: Customized Video Generation with Preserving Motion and Concept Composition Abilities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

Yuming Jiang, Tianxing Wu, Shuai Yang, Chenyang Si, Dahua Lin, Yu Qiao, Chen Change Loy, and Ziwei Liu. VideoBooth: Diffusion-Based Video Generation with Image Prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Yujie Wei, Shiwei Zhang, Hangjie Yuan, Xiang Wang, Haonan Qiu, Rui Zhao, Yutong Feng, Feng Liu, Zhizhong Huang, Jiaxin Ye, Yingya Zhang, and Hongming Shan. DreamVideo-2: Zero-Shot Subject-Driven Video Customization with Precise Motion Control. *arXiv preprint arXiv:2410.13830*, 2024.

Teng Hu, Zhentao Yu, Zhengguang Zhou, Sen Liang, Yuan Zhou, Qin Lin, and Qinglin Lu. HunyuanCustom: A Multimodal-Driven Architecture for Customized Video Generation. *arXiv preprint arXiv:2505.04512*, 2025.

Yuanhao Cai, He Zhang, Xi Chen, Jinbo Xing, Yiwei Hu, Yuqian Zhou, Kai Zhang, Zhifei Zhang, Soo Ye Kim, Tianyu Wang, Yulun Zhang, Xiaokang Yang, Zhe Lin, and Alan Yuille. OmniVCus: Feedforward Subject-Driven Video Customization with Multimodal Control Conditions. *arXiv preprint arXiv:2506.23361*, 2025.

Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. OmniControl: Control Any Joint at Any Time for Human Motion Generation. In *International Conference on Learning Representations*, 2024.

Max Ku, Cong Wei, Weiming Ren, Harry Yang, and Wenhu Chen. AnyV2V: A Tuning-Free Framework for Any Video-to-Video Editing Tasks. *arXiv preprint arXiv:2403.14468*, 2024.

Inkyu Shin, Qihang Yu, Xiaohui Shen, In So Kweon, Kuk-Jin Yoon, and Liang-Chieh Chen Enhancing Temporal Consistency in Video Editing by Reconstructing Videos with 3D Gaussian Splatting. *Transactions on Machine Learning Research*, 2025.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. VACE: All-in-One Video Creation and Editing. *arXiv preprint arXiv:2503.07598*, 2025.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, and others. Wan: Open and Advanced Large-Scale Video Generative Models *arXiv preprint arXiv:2503.20314*, 2025.

Cong Wei, Quande Liu, Zixuan Ye, Qiulin Wang, Xintao Wang, Pengfei Wan, Kun Gai, and Wenhu Chen. UniVideo: Unified Understanding, Generation, and Editing for Videos. *arXiv preprint arXiv:2510.08377*, 2025.

Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. BC-Z: Zero-Shot Task Generalization with Robotic Imitation Learning In *Proceedings of The 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pages 991–1002, 2022.

Homer Walke, Kevin Black, Abraham Lee, Moo Jin Kim, Max Du, Chongyi Zheng, Tony Zhao, Philippe Hansen-Estruch, Quan Vuong, Andre He, Vivek Myers, Kuan Fang, Chelsea Finn, and Sergey Levine. BridgeData V2: A Dataset for Robot Learning at Scale. In *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pages 1723–1736, 2023.

Gaoyue Zhou, Victoria Dean, Mohan Kumar Srirama, Aravind Rajeswaran, Jyothish Pari, Kyle Hatch, Aryan Jain, Tianhe Yu, Pieter Abbeel, Lerrel Pinto, Chelsea Finn, and Abhinav Gupta. Train Offline, Test Online: A Real Robot Learning Benchmark. *arXiv preprint arXiv:2306.00942*, 2023.

Lawrence Y. Chen, Samson Adebola, and Ken Goldberg. Berkeley UR5 Demonstration Dataset. 2023.

Xinghao Zhu, Ran Tian, Chenfeng Xu, Mingxiao Huo, Wei Zhan, Masayoshi Tomizuka, and Mingyu Ding. Fanuc Manipulation: A Dataset for Learning-based Manipulation with FANUC Mate 200iD Robot 2023.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. In *Proceedings of Robotics: Science and Systems*, 2023.

Suneel Belkhale, Yuchen Cui, and Dorsa Sadigh. HYDRA: Hybrid Robot Actions for Imitation Learning. *arXiv preprint arXiv:2306.17237*, 2023.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman R"adle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment Anything with Concepts. *arXiv preprint arXiv:2511.16719*, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*, 2025.

Lawrence Yunliang Chen, Kush Hari, Karthik Dharmarajan, Chenfeng Xu, Quan Vuong, and Ken Goldberg. Mirage: Cross-Embodiment Zero-Shot Policy Transfer with Cross-Painting. In *Proceedings of Robotics: Science and Systems*, 2024.

Guangrun Li, Yaoxu Lyu, Zhuoyang Liu, Chengkai Hou, Jieyu Zhang, and Shanghang Zhang. H2R: A Human-to-Robot Data Augmentation for Robot Pre-training from Videos. *arXiv preprint arXiv:2505.11920*, 2025.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

Homanga Bharadhwaj, Abhinav Gupta, Vikash Kumar, and Shubham Tulsiani. Towards Generalizable Zero-Shot Manipulation via Translating Human Interaction Plans. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 6904–6911, 2024.

Mingyo Seo, H. Andy Park, Shenli Yuan, Yuke Zhu, and Luis Sentis. LEGATO: Cross-Embodiment Imitation Using a Grasping Tool. *IEEE Robotics and Automation Letters*, 10(3):2854–2861, 2025.

Yang Bai, Liudi Yang, George Eskandar, Fengyi Shen, Dong Chen, Mohammad Altillawi, Ziyuan Liu, and Gitta Kutyniok. RoboSwap: A GAN-driven Video Diffusion Framework for Unsupervised Robot Arm Swapping. *arXiv preprint arXiv:2506.08632*, 2025.

Oriane Siméoni and Huy V. Vo and Maximilian Seitzer and Federico Baldassarre and Maxime Oquab and Cijo Jose and Vasil Khalidov and Marc Szafraniec and Seungeun Yi and Michaël Ramamonjisoa and Francisco Massa and Daniel Haziza and Luca Wehrstedt and Jianyuan Wang and Timothée Darcet and Théo Moutakanni and Leonel Sentana and Claire Roberts and Andrea Vedaldi and Jamie Tolan and John Brandt and Camille Couprie and Julien Mairal and Hervé Jégou and Patrick Labatut and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Dingkun Yan, Xinrui Wang, Ru Wang, Zhuoru Li, Jinze Yu, Yusuke Iwasawa, Yutaka Matsuo, and Jiaxian Guo. Towards High-resolution and Disentangled Reference-based Sketch Colorization. *arXiv preprint arXiv:2603.05971*, 2026.

Tomohiro Motoda, Masaki Murooka, Ryoichi Nakajo, Muhammad A. Muttaqien, Koshi Makihara, Hanbit Oh, Keisuke Shirai, Floris Erich, Ryo Hanai, and Yukiyasu Domae. AIST Bimanual Manipulation Dataset: A Dataset for Learning Bimanual Robot Control, Collected via Leader-Follower Teleoperation. `https://aistairc.github.io/aist_bimanip_site/`, 2025.

Jeffrey S. Vitter. Random Sampling with a Reservoir. *ACM Transactions on Mathematical Software*, 11(1):37–57, 1985.

APPENDIX

A AUXILIARY TRAINING OBJECTIVES

In addition to our primary token-level correlation regularizer ($\mathcal{L}_{\text{tok}}$) and the masked flow matching objective ($\mathcal{L}_{\text{fm}}$), we employ two auxiliary spatial-temporal losses during training. These terms operate on the predicted clean latents to improve local replacement accuracy and cross-frame stability.

Let x_0 denote the target clean video latents and $\hat{x}_0$ denote the model's prediction of the clean latents, which can be derived from the predicted vector field v_θ and the current noisy state z_t.

Masked Reconstruction Loss ($\mathcal{L}_{\text{rec}}$). To improve the accuracy of the local replacement, we apply an L_1 reconstruction loss exclusively inside the edited region. Let W denote the spatial mask weight mapped to the latent grid. We define the reconstruction loss as:

$$\mathcal{L}_{\text{rec}} = \frac{\|W \odot (\hat{x}_0 - x_0)\|_1}{\|W\|_1} \tag{2}$$

This term improves local reconstruction fidelity inside the edited region. In the context of robotic manipulation, it helps mitigate under-editing by explicitly encouraging the edited latent to closely match the target reference identity within the masked boundary.

Masked Temporal Consistency Loss ($\mathcal{L}_{\text{temp}}$). To reduce frame-to-frame instability during the generation process, we further constrain the temporal differences between adjacent reconstructed latents. Let W_t denote the temporal mask weight between neighboring frames t and $t+1$. The temporal consistency loss is formulated as:

$$\mathcal{L}_{\text{temp}} = \frac{\left\|W_t \odot \left((\hat{x}_0^{t+1} - \hat{x}_0^{t}) - (x_0^{t+1} - x_0^{t})\right)\right\|_1}{\|W_t\|_1} \tag{3}$$

This penalty enforces that the temporal gradients (i.e., the frame-to-frame evolution) of the generated region match the source dynamics. It is particularly helpful for suppressing high-frequency flicker, stabilizing contact boundaries over time, and preventing temporal drift as the manipulator moves through the scene.

While these two auxiliary terms are essential for local precision and temporal smoothness, they do not inherently resolve the structural entanglement introduced by reference conditioning. That primary role is fulfilled by our proposed $\mathcal{L}_{\text{tok}}$ regularizer described in the main text.

B PROOF OF LEMMA 1

Proof. This proof directly addresses the four assumptions (A1–A4) to establish the pointwise zero-gradient claims. Proof of Lemma 1:By Assumption (A3), the ideal continuous objective achieves exact distillation, meaning $\mathcal{L}_{\text{anch}} = 0$. Because the integrand is a squared Frobenius norm (which is strictly non-negative) and expected over a continuous domain, it must hold that the integrand is zero almost everywhere. Therefore, we have:

$$A_\theta(z_t, t, c_R) = A_\theta(z_t, t, c_\emptyset) \quad \text{a.e.}$$

By Assumption (A1), the null-reference branch $A_\theta(z_t, t, c_\emptyset)$ is strictly independent of the reference R. Consequently, $A_\theta(z_t, t, c_R)$ is also independent of R almost everywhere. Given the differentiability condition in Assumption (A4), taking the Jacobian with respect to R directly yields the first claim:

$$\nabla_R A_\theta(z_t, t, c_R) = 0 \quad \text{a.e.}$$

Next, we evaluate the gradient of the outside-mask vector field. By Assumption (A2), the dependence of v_θ^{out} on R is bottlenecked entirely through the anchor A_θ. Applying the multivariate chain rule yields:

$$\nabla_R v_\theta^{out}(z_t, t, c_R) = \left(\frac{\partial A_\theta}{\partial R}\right)^T \nabla_{A_\theta} f_\theta(z_t, t, V^{out}, M, p, A_\theta)$$

Substituting $\nabla_R A_\theta = 0$ into this expression immediately eliminates the right-hand side, establishing the second claim:

$$\nabla_R v_\theta^{out}(z_t, t, c_R) = 0 \quad \text{a.e.}$$

This completes the proof. ■ □

C Proof of Proposition 1 and Gradient Corollary

On the row-loss formulation. Eq. (1) is written as a Frobenius norm over a reservoir-sampled submatrix, but it is equivalent to the sample mean of per-row scalar losses ℓ_i. This equivalence is what enables the analysis below. A natural alternative would be to subsample both the rows of the cross-Gram difference *and* the normalizer, yielding an estimator of the form $\|W_{\text{out}} \odot (\tilde{G}^{\text{align}} - \tilde{G}^{\text{null}})\|_F^2 / \|\tilde{W}_{\text{out}}\|_1$ with $\tilde{W}_{\text{out}}$ restricted to the sampled rows. Such an estimator is a ratio of two sample means and is generally biased for the population ratio L_{exact}. Holding the per-row normalizer $|O| + \epsilon$ fixed inside ℓ_i and averaging the resulting scalars converts the problem into a standard finite-population sample-mean estimation, for which the unbiasedness, variance, and gradient results below all follow cleanly.

Proof. We prove (i) unbiasedness and (ii) the finite-population variance.

Unbiasedness. Our reservoir algorithm uniformly samples K indices without replacement from the set of inside-mask tokens I. Let $\mathbb{I}_i = \mathbb{I}(i \in S_K)$ be the indicator variable denoting whether token $i \in I$ is included in the sample S_K. Since the sampling is uniform, the probability of any specific token i being included is:

$$P(i \in S_K) = \frac{K}{N_{\text{in}}}.$$

The empirical loss can be rewritten as a sum over all tokens in I using the indicator function:

$$\widehat{L}_K = \frac{1}{K} \sum_{i \in I} \ell_i \cdot \mathbb{I}_i.$$

Taking the expectation over the sampling process S_K and using $\mathbb{E}_{S_K}[\mathbb{I}_i] = K/N_{\text{in}}$:

$$\mathbb{E}_{S_K}[\widehat{L}_K] = \frac{1}{K} \sum_{i \in I} \ell_i \cdot \frac{K}{N_{\text{in}}} = \frac{1}{N_{\text{in}}} \sum_{i \in I} \ell_i = L_{\text{exact}}.$$

This guarantees that the sampled loss is an unbiased estimator of L_{exact}.

Variance. Because S_K is drawn without replacement, $\widehat{L}_K$ is the sample mean of a simple random sample drawn from a finite population of size N_{in}. By the standard formula for the variance of a sample mean under finite-population sampling,

$$\text{Var}_{S_K}(\widehat{L}_K) = \left(1 - \frac{K}{N_{\text{in}}}\right) \frac{\sigma_\ell^2}{K}, \qquad \sigma_\ell^2 = \frac{1}{N_{\text{in}} - 1} \sum_{i \in I} (\ell_i - L_{\text{exact}})^2.$$

Conditioned on the token features, the row losses ℓ_i are fixed, making σ_ℓ^2 a constant with respect to the sampling process. The finite-population correction factor $1 - K/N_{\text{in}} \in [0, 1)$ provides a strict variance reduction relative to with-replacement sampling. We therefore obtain the bound

$$\text{Var}_{S_K}(\widehat{L}_K) \leq \frac{\sigma_\ell^2}{K},$$

so that $\text{Var}_{S_K}(\widehat{L}_K) = \mathcal{O}(1/K)$. □

Corollary 1 (Unbiased gradient for SGD training)**.** *Under the conditions of Proposition 1, and assuming each row loss ℓ_i is differentiable in θ,*

$$\mathbb{E}_{S_K}\left[\nabla_\theta \widehat{L}_K\right] = \nabla_\theta L_{\text{exact}}, \qquad \text{Var}_{S_K}\left(\nabla_\theta \widehat{L}_K\right) = \mathcal{O}(1/K).$$

Optimizing $\widehat{L}_K$ via stochastic gradient descent therefore shares its stationary points with the dense objective L_{exact}, with stochastic-gradient variance directly controlled by the reservoir size K.

Proof. For each $i \in I$, the row loss $\ell_i(\theta)$ depends on θ through the cross-Gram entries $G_{ij}^{\text{align}}, G_{ij}^{\text{null}}$ and is differentiable by assumption. Writing $\widehat{L}_K = \frac{1}{K} \sum_{i \in I} \ell_i(\theta)\, \mathbb{I}_i$, the indicators $\mathbb{I}_i$ depend only on the sampling process S_K and are independent of θ. Gradient and expectation therefore commute:

$$\mathbb{E}_{S_K}\left[\nabla_\theta \widehat{L}_K\right] = \frac{1}{K} \sum_{i \in I} \nabla_\theta \ell_i(\theta) \cdot \mathbb{E}_{S_K}[\mathbb{I}_i] = \frac{1}{K} \sum_{i \in I} \nabla_\theta \ell_i(\theta) \cdot \frac{K}{N_{\text{in}}} = \nabla_\theta L_{\text{exact}}.$$

Applying the finite-population variance bound coordinate-wise to the parameter-indexed family $\{\nabla_\theta \ell_i\}_{i \in I}$ yields $\mathrm{Var}_{S_K}(\nabla_\theta \widehat{L}_K) = \mathcal{O}(1/K)$ in each coordinate, and hence in any fixed norm. The training-convergence consequence is then the standard SGD result for unbiased stochastic gradients with bounded variance. □

D Mask and Reference Quality Control

For each video, SAM3 generates frame-wise masks for the manipulator or object. We remove clips where the target is not detected or where valid foreground coverage is consistently too small. We further evaluate masked crops with Qwen3-VL to reject samples with severely incomplete foregrounds or excessive background contamination. For retained samples, we dilate the masks to tolerate moderate boundary errors and to cover the gripper–object interaction region. Reference crops are selected from multiple candidate frames by mask-area ratio and VLM assessment, composited onto a white background, and perturbed by translation and rotation during training to reduce pose-level shortcut learning. At inference time, difficult cases can be refined with positive and negative points in an interactive SAM3-based mask tool.

E Detailed Experimental Settings

We adopt HunyuanCustom as the base video generation model and fine-tune it with LoRA. For parameter-efficient adaptation, we insert LoRA modules with rank $r = 8$ and scaling factor $\alpha = 16$ into all attention and feed-forward layers of both the dual-stream and single-stream transformer blocks. During training, in addition to the original flow-matching loss, we use three auxiliary objectives: a masked x_0 reconstruction loss, a masked temporal consistency loss, and a token-level correlation regularization loss. The masked x_0 reconstruction loss, with weight $\lambda_{x_0} = 0.1$, penalizes the ℓ_1 distance between the predicted and ground-truth clean latents inside a segmentation-derived mask, encouraging the model to focus on the edited region. The masked temporal consistency loss, with weight $\lambda_{\text{temp}} = 0.05$, enforces agreement between temporal differences of predicted and target latents, helping maintain smooth and coherent motion within the mask. The token-level correlation regularization loss, with weight $\lambda_{\text{tok}} = 5 \times 10^{-5}$, aligns the query-to-all token correlation structure between a reference-conditioned forward pass and a null-reference control pass. We further apply classifier-free guidance dropout with probability $p_{\text{drop}} = 0.1$.

F Robot Manipulation Consistency

To evaluate whether edited robot videos preserve plausible and consistent manipulation behavior, we propose **Robot Manipulation Consistency (RMC)**. Specifically, for a given edited video $V^e = \{I_t^e\}_{t=1}^T$, a source video $V^o = \{I_t^o\}_{t=1}^T$, and their robot masks $\{M_t^e\}_{t=1}^T$, we compute

$$\mathrm{RMC} = \mathrm{ref}_{\text{sim}} \cdot \mathrm{RMC}_{\text{inner}}, \qquad \mathrm{RMC}_{\text{inner}} = \alpha\, \mathrm{IoU}_{\text{temp}} + \beta\, \mathrm{Traj}_{\text{score}} + \gamma\, \mathrm{Feat}_{\text{sim}}, \tag{4}$$

where $\alpha + \beta + \gamma = 1$.

Temporal mask consistency. We measure the short-term geometric stability of edited masks using adjacent-frame IoU:

$$\mathrm{IoU}_{\text{temp}} = \frac{1}{|S|} \sum_{t \in S} \mathrm{IoU}(M_{t-1}^e, M_t^e) = \frac{1}{|S|} \sum_{t \in S} \frac{\left|M_{t-1}^e \cap M_t^e\right|}{\left|M_{t-1}^e \cup M_t^e\right|}, \tag{5}$$

where S includes only frame indices for which both adjacent masks are non-empty. The score ranges from 0 to 1, with lower values indicating temporal jitter or inconsistent mask shapes and positions.

Trajectory preservation to source video. Let $\mathbf{c}_t^e$ and $\mathbf{c}_t^s$ be mask centroids in edited and source videos. We compute normalized trajectory error:

$$\mathrm{Traj}_{\text{err}} = \frac{1}{|U|} \sum_{t \in U} \frac{\|\mathbf{c}_t^e - \mathbf{c}_t^s\|_2}{\sqrt{W^2 + H^2}}, \tag{6}$$

where U includes frames with valid centroids in both videos, and (H, W) is the frame size. We then convert the error to a score:

$$\mathrm{Traj}_{\mathrm{score}} = \frac{1}{1 + \mathrm{Traj}_{\mathrm{err}}}. \tag{7}$$

Semantic stability. To quantify appearance consistency of the manipulated region over time, we extract DINOv2 features from the masked robot crop of each edited frame, denoted by $\mathbf{f}_t^e$. We then measure temporal coherence using adjacent-frame cosine similarity:

$$\mathrm{Feat}_{\mathrm{sim}} = \frac{1}{|Q|} \sum_{t \in Q} \frac{\cos(\mathbf{f}_{t-1}^e, \mathbf{f}_t^e) + 1}{2}, \tag{8}$$

where Q is the set of adjacent frame pairs with valid foreground masks. Higher values indicate more stable robot appearance and reduced temporal flicker.

Reference-image gating. Given a reference image I^r, we compute a reference-consistency factor from patch-level DINOv2 matching. For each edited frame t, let $\mathcal{E}_t$ denote foreground patch features and $\mathcal{R}$ denote reference-image patch features. We define a frame-level match score as

$$s_t = \frac{1}{|\mathcal{E}_t|} \sum_{\mathbf{e} \in \mathcal{E}_t} \max_{\mathbf{r} \in \mathcal{R}} \cos(\mathbf{e}, \mathbf{r}), \tag{9}$$

and aggregate across frames:

$$\mathrm{ref}_{\mathrm{sim}} = \frac{1}{|P|} \sum_{t \in P} s_t. \tag{10}$$

Implementation details. All masks described above are generated by SAM3. We use $(\alpha, \beta, \gamma) = (0.3, 0.5, 0.2)$ for Eq. equation 4.

G Multimodal Judge Evaluation

Standard video metrics provide useful information for reference similarity, text alignment, and frame-to-frame stability. However, they do not directly measure a key failure mode in local robotic video editing: corruption of the non-edited scene. In our setting, this includes reference leakage into the background, manipulated-object corruption, contact-boundary artifacts, or unintended changes to nearby objects and support surfaces. These failures are especially important for manipulation videos, since local artifacts around the gripper or the object can mislead the task context.

We apply a multimodal judge as a complementary evaluation protocol. For each video, we sample a fixed number of evenly spaced frames. For each sampled frame, the judge receives three images in order: the reference image specifying the target manipulator appearance, the source frame before editing, and the edited frame after editing. The judge is also given the editing instruction and the frame index. It then returns structured scores and diagnostic tags in JSON format. We average the frame-level scores to obtain video-level scores, and then average across the benchmark.

We report two judge-based scores. **Scene Preservation** measures whether the edited frame preserves the non-edited regions, including the manipulated object and background, and penalizes reference leakage outside the intended robot-arm region. On the other hand, **Reference Fidelity** evaluates whether the edited manipulator matches the reference image and is naturally integrated into the scene. Both scores are reported on a 1–10 scale, where higher is better. We use GPT-5 as the multimodal judge and provide the full prompt below.

Prompt for multimodal judge evaluation.

You are evaluating one frame from a robot video editing result.

You are given three images in order:

- Image 1: reference image – the target robot arm appearance that should be edited into the video.
- Image 2: input frame before editing.
- Image 3: edited frame after editing.

Editing instruction: `{instruction}`

The intended edit is local: the robot arm / manipulator in the input frame should be replaced to resemble the reference robot arm. Everything else – the manipulated object, contact context, nearby objects, table surface, and background – should remain faithful to the input frame unless the change is directly caused by normal robot motion or occlusion.

Evaluate two aspects:

[A] Reference Fidelity.

- Does the edited robot arm's shape, color, and texture match Image 1?
- Is the arm naturally integrated into the scene without seams or blending artifacts?

[B] Scene Preservation. Does the edit corrupt areas outside the intended robot-arm region? Consider:

1. Manipulated object corruption, disappearance, identity change, or physically implausible state/position change.
2. Contact-region artifacts around the robot gripper and manipulated object, such as broken boundaries, melting, duplicated parts, or unnatural blending.
3. Nearby object, table, support surface, or background changes not explained by normal robot motion.
4. Reference leakage: the reference robot's color, texture, shape, or shadow appearing on the object, table, or background outside the edit region.

Do not penalize:

- normal robot-arm movement or pose change;
- normal occlusion caused by the moving arm;
- small lighting or compression differences;
- changes inside the intended robot-arm edit region.

Score guides:

Scene Preservation:

- 10: no visible artifact outside the intended edit region.
- 7–9: only minor boundary artifacts; object and background are mostly preserved.
- 4–6: visible artifacts near contact regions or nearby objects, but the task context remains recognizable.
- 1–3: severe background corruption, wrong object state, or strong reference leakage.

Reference Fidelity:

- 10: the edited arm closely matches the reference in shape, color, and texture, and is seamlessly integrated.
- 7–9: mostly matches the reference, with minor color/texture deviation or slight seams.
- 4–6: partial match, with noticeable shape or color mismatch.
- 1–3: the edited arm looks unlike the reference or is severely distorted.

Return only a JSON object:

```
{
  "scene_preservation_score": <int 1-10>,
  "reference_fidelity_score": <int 1-10>,
  "brief_reason": "<one short sentence>"
}
```

H HUMAN STUDY PROTOCOL

We conduct a blinded four-way preference study on complete edited videos. Each participant views 10 examples, each containing four anonymized outputs labeled A–D, and selects the best overall edit. The interface does not reveal method identities. After excluding incomplete submissions, 15 participants provide 150 valid votes. RoboEdit receives 125/150 preferences (83.3%), is preferred most often on 9 of 10 video sets, and significantly differs from a uniform preference distribution ($\chi^2(3) = 273.3, p < 0.001$).

I ADDITIONAL QUALITATIVE RESULTS

We provide additional qualitative editing results in Figure 6, covering diverse manipulator- and object-replacement scenarios. For manipulator replacement, RoboEdit takes a reference image of the target manipulator and replaces the original arm while preserving the surrounding scene, manipulation trajectory, and temporal dynamics. For object replacement, RoboEdit uses the manipulator reference image together with a text prompt describing the target object, enabling localized object editing while maintaining consistent robot motion and scene context.These examples further illustrate that RoboEdit preserves scene context, local contact structure, and motion dynamics while following the target appearance specified by the reference. Animated versions of five edited examples, together with their source videos, are provided in Figure 7, where each panel loops in JavaScript-enabled PDF viewers (*e.g.*, Adobe Acrobat).

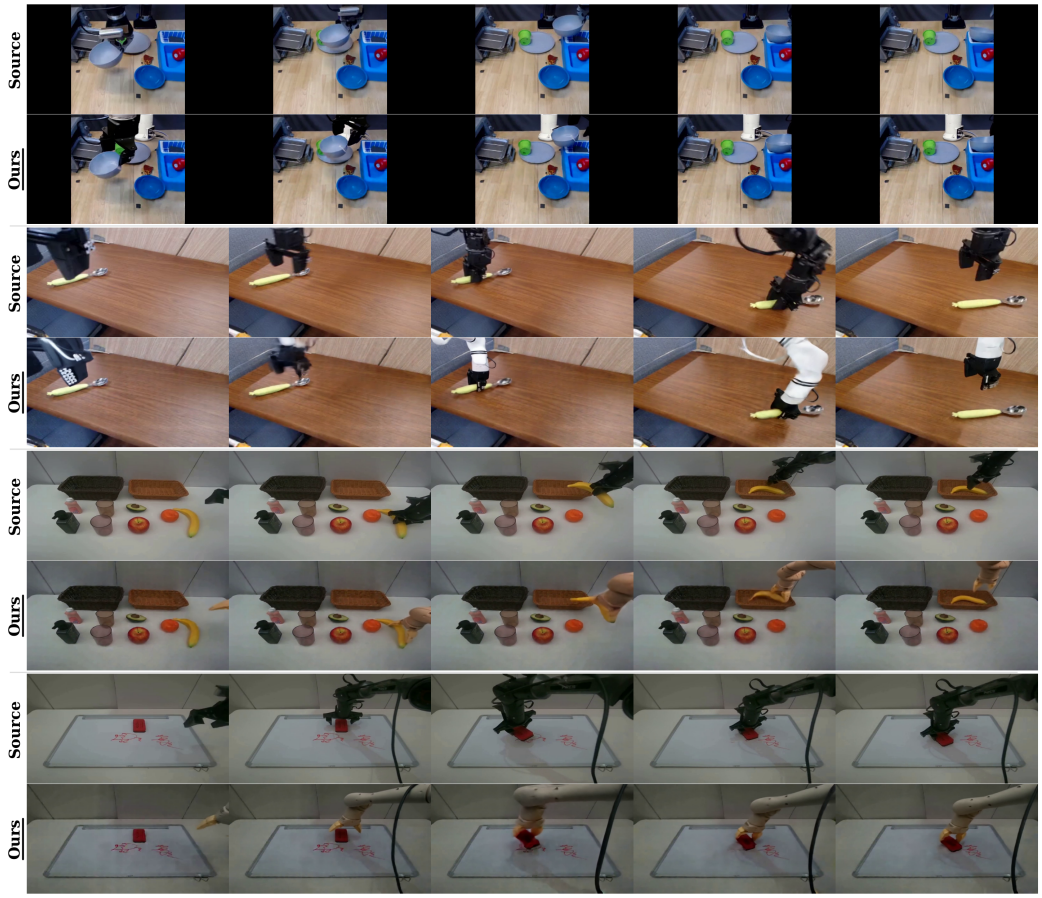

Figure 6: Additional qualitative results of RoboEdit on robotic local video editing. Results are best viewed zoomed in.

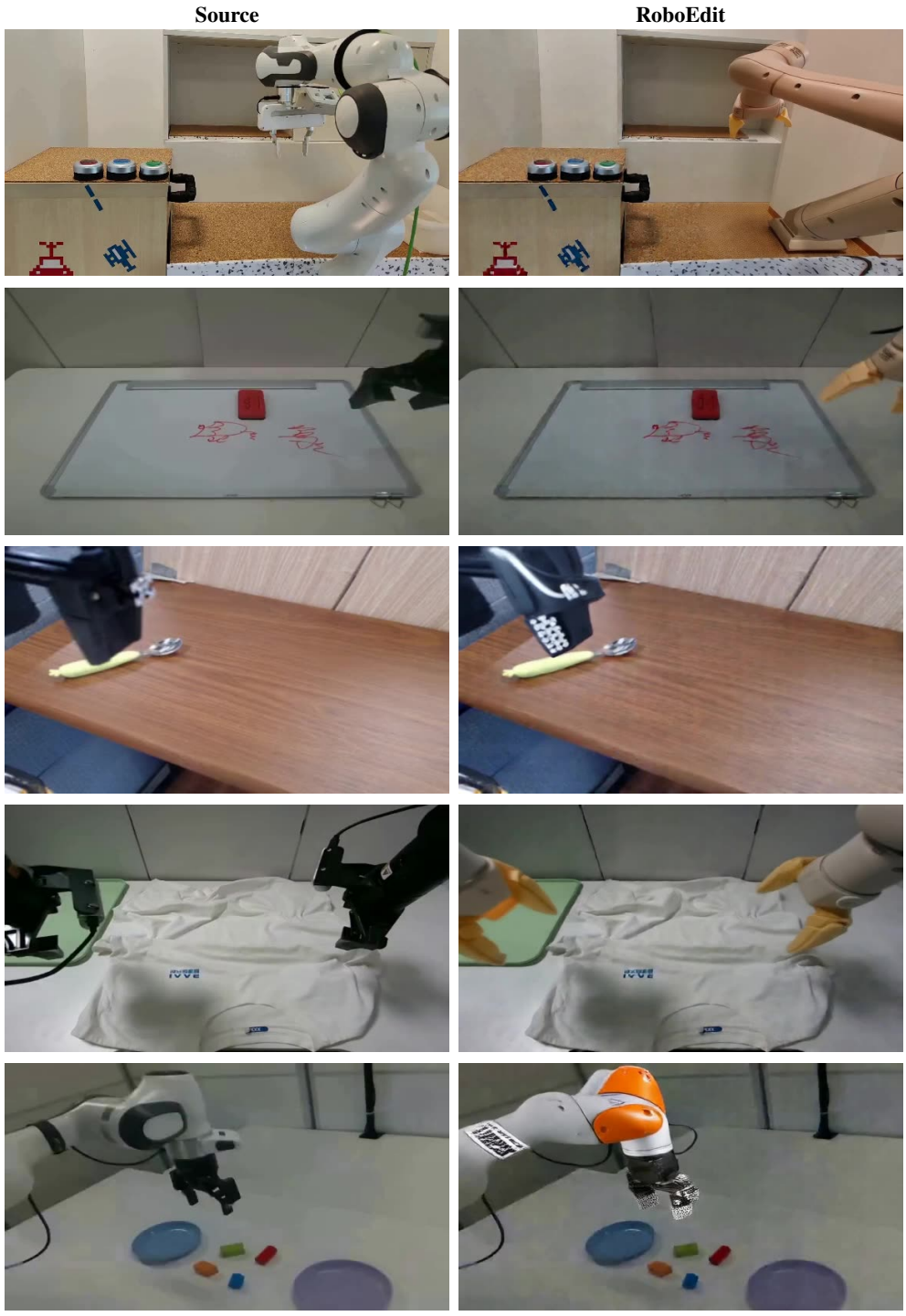

Figure 7: Animated RoboEdit results on five robotic manipulation examples. Left: original source videos. Right: RoboEdit-edited videos, where the manipulator is replaced according to a target reference. In JavaScript-enabled PDF viewers (*e.g.*, Adobe Acrobat) each panel plays as a loop; other viewers display the first frame.

J Qualitative Comparison with Video Editing Baselines

We provide a qualitative comparison between **RoboEdit** and several representative video editing methods, including **HunyuanCustom**, **Wan2.2-Replace**, **UniVideo**, and **Seedance**, in Fig. 8.

HunyuanCustom and UniVideo show some ability to transfer basic texture, color, and coarse structure, but they often fail to follow the original manipulation trajectory. The manipulated object also changes abruptly or drifts across frames. Seedance 2.0 achieves better transfer of the manipulator appearance, especially in color and overall structure. However, it still fails to preserve the original motion trajectory, and the manipulated object is often changed from the first frame. Although the video remains temporally stable, the object state is consistently incorrect. Finally, Wan2.2-Replace fails to produce meaningful editing results in this setting.

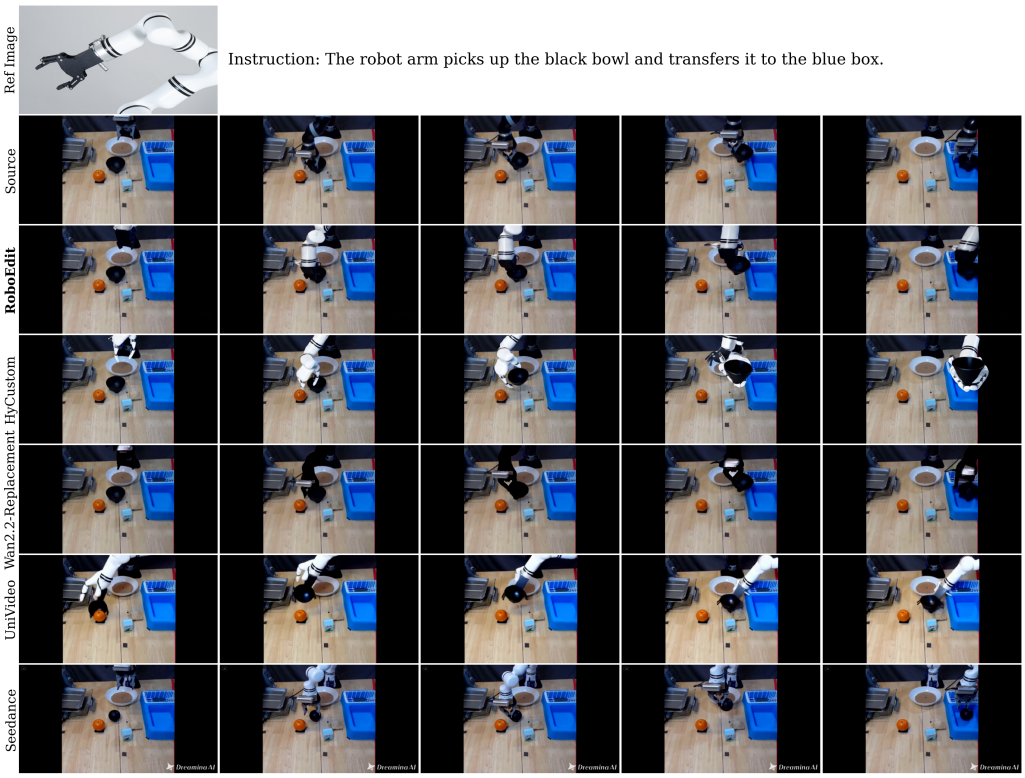

Figure 8: Qualitative comparison with representative video editing baselines on robotic manipulation scenes. RoboEdit better preserves manipulation dynamics, surrounding scene content, and edited-region fidelity under local reference-guided editing.

K Qualitative Video Comparison for the Token-Level Regularizer

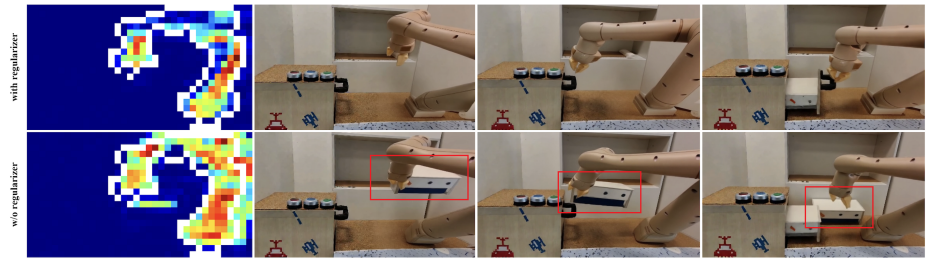

Figure 9: Visualization of $\Delta\tilde{G}$ under reference-conditioned and null-reference settings. The white border marks the local editing region, while the outside area corresponds to the surrounding scene. Without the regularizer, the discrepancy leaks beyond the mask and produces unintended cabinet-like artifacts near the manipulator, indicating weaker spatial confinement.

Figure 9 shows the effect of the token-level correlation regularizer through the correlation-difference heatmap $\Delta\tilde{G}$. Without the regularizer, the discrepancy between the reference-conditioned and null-reference runs extends beyond the mask, but it is mostly concentrated near the mask boundary and adjacent manipulator-motion regions rather than over the entire frame. This pattern suggests that the main failure mode is localized perturbation around the editable region, rather than broad corruption of the overall scene structure.

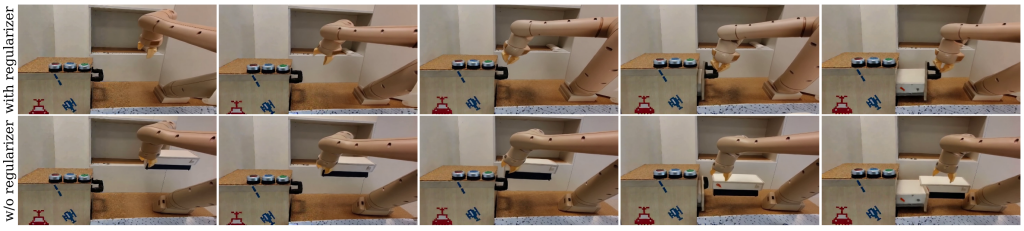

Figure 10: Effectiveness of the token-level correlation regularizer.

To complement the analysis, we further provide qualitative comparisons on the edited videos. Figure 10 shows that such boundary-local leakage often corresponds to visible artifacts in surrounding content that interacts closely with the edited manipulator. In particular, when the manipulator is near neighboring objects or contact regions, the version without the regularizer tends to introduce abrupt local changes in the nearby context. By contrast, with the regularizer, the edited video remains visually cleaner around the boundary, with better preservation of surrounding scene content.

L Layer-wise Analysis of Token-Level Correlation Regularization

We further study the effect of where the token-level correlation regularizer is imposed. Recent multimodal diffusion transformers often contain modality-specific streams followed by shared single-stream blocks. The former process each modality within its own stream, while the latter mix different modalities (like visual, textual, and speech tokens) in a shared space. Since reference-induced leakage is most directly reflected after such cross-modal mixing, we analyze the regularizer within the single-stream blocks and the final layer.

To do so, we use a layer-wise diagnostic probe to measure how much the reference condition perturbs the relation between the edited region and the surrounding scene. For each sampled video, we run two forward passes under identical settings, differing only in the given reference image. At each layer, we record the corresponding token-correlation discrepancy as a diagnostic signal. Figure 11 reports the average token-level correlation discrepancy over 100 sampled videos.

Table 4: Correlation-regularizer layer ablation.

Model	CLIP ↑	DINO-Sim ↑	Temp-Consis ↑	RMC ↑	Scene Pres. ↑	Ref. Fid. ↑
RoboEdit@Final layer	0.3050	**0.2914**	0.9163	**0.3426**	**7.28**	**7.01**
RoboEdit@Middle layer	**0.3066**	0.2471	0.9127	0.3335	6.68	6.27
RoboEdit@Shallow layer	0.3059	0.2457	0.9108	0.3307	6.67	6.53
w/o Reg.	0.3042	0.2318	**0.9184**	0.3200	6.56	6.19

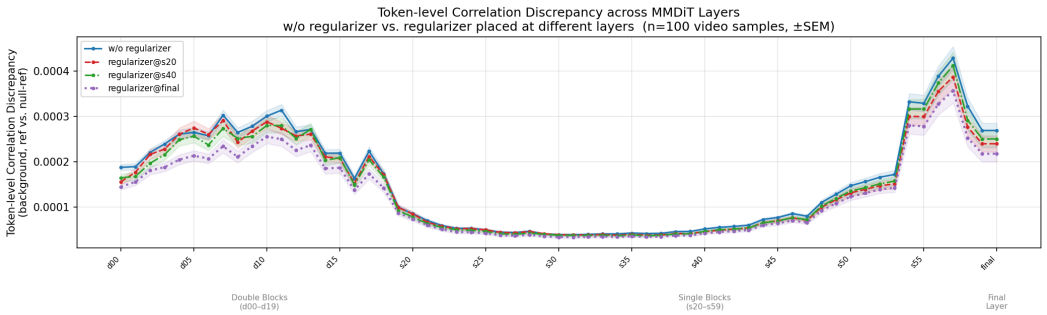

Figure 11: Layer-wise token-correlation discrepancy under different regularizer placements. The final-layer placement most effectively suppresses the discrepancy near the output side.

The results show a clear pattern. When the regularizer is applied to a shallow single-stream layer, the discrepancy is reduced around that layer, but it increases again in later layers. This suggests that subsequent multimodal mixing can reintroduce reference-induced changes into the surrounding scene. The middle-layer variant reduces this rebound, but still leaves a noticeable discrepancy near the output side. In contrast, applying the regularizer at the final layer gives the lowest discrepancy near the prediction layer, where the output representation is already formed and no later blocks can amplify the reference signal again. This trend is consistent with Table 4, where the final-layer variant achieves the best scene preservation and reference fidelity. We therefore apply the token-level correlation regularizer at the final layer by default.

M Downstream Simulation Evaluation

Video-editing metrics alone cannot show whether the generated videos are useful for robot control, so we include a downstream representation-transfer evaluation following H2R (Li et al., 2025). The target embodiment is always Franka Panda in MuJoCo. For each pretraining condition, we start from the same ImageNet-pretrained MAE encoder (He et al., 2022), continue MAE pretraining on the corresponding robot-video frames, freeze the encoder, and train the same behavior-cloning action head on the same Panda demonstrations. We then evaluate the resulting policy with 100 simulation rollouts per task. Since the policy architecture, action-labeled demonstrations, and rollout protocol are fixed, the success-rate differences mainly reflect the visual representation learned from each video source.

We first report the full Sawyer-to-Panda setting in Table 5. The original three Robomimic tasks, Lift, MoveCan, and Square, are extended with six MimicGen tasks: Stack, ThreePiece, Coffee, MugCleanup, StackThree, and SquareD0. The compared pretraining sources include unedited Sawyer videos, native Panda videos, HyCustom-edited Sawyer-to-Panda videos, Mirage-like edited videos, Ro-Aug generated videos, RoboEdit-edited Sawyer-to-Panda videos, and mixtures with native Panda videos. We then test whether the improvement is specific to Sawyer by repeating the same protocol with Sawyer, IIWA, and UR5e as source embodiments while keeping Panda as the target embodiment; these benchmark-level averages are reported in Table 6.

Table 5: Full downstream simulation results for Sawyer-to-Panda representation transfer. Values are success rates in percent.

Task	Out-of-domain	In-domain	HyCustom	Mirage-like	Ro-Aug	RoboEdit	RoboEdit + in-domain
Lift	62%	92%	80%	85%	88%	**95%**	93%
MoveCan	32%	49%	59%	39%	54%	63%	**72%**
Square	4%	20%	5%	3%	2%	15%	**25%**
Stack	45%	65%	55%	64%	40%	66%	**77%**
ThreePiece	3%	11%	2%	**27%**	20%	23%	21%
Coffee	2%	10%	3%	2%	4%	8%	**13%**
MugCleanup	15%	18%	14%	16%	9%	**29%**	27%
StackThree	4%	7%	6%	5%	14%	7%	**16%**
SquareD0	22%	50%	42%	44%	33%	55%	**58%**
Average	21.0%	35.8%	29.6%	31.7%	29.3%	40.1%	**44.7%**

Table 6: Downstream simulation averages across source embodiments. Values are success rates in percent.

Benchmark	Method	Sawyer	IIWA	UR5e
Robomimic	Unedited	32.7%	55.0%	45.3%
	HyCustom	48.0%	47.0%	39.0%
	Mirage-like	42.3%	48.3%	47.7%
	Ro-Aug	48.0%	49.0%	47.7%
	RoboEdit	**57.7%**	**60.0%**	**50.0%**
MimicGen	Unedited	15.2%	26.8%	23.5%
	HyCustom	20.3%	22.7%	25.5%
	Mirage-like	26.3%	29.3%	24.7%
	Ro-Aug	20.0%	28.0%	26.3%
	RoboEdit	**31.3%**	**34.5%**	**35.5%**
Overall	Unedited	21.0%	36.2%	30.8%
	HyCustom	29.6%	30.8%	30.0%
	Mirage-like	31.7%	35.7%	32.3%
	Ro-Aug	29.3%	35.0%	33.4%
	RoboEdit	**40.1%**	**43.0%**	**40.3%**

N Preliminary Real-Robot Evaluation

We report an additional real-robot evaluation on a dual-arm OpenArm platform. Since our OpenArm is bimanual, we use the AIST-Bimanual Manipulation Dataset Motoda et al. (2025), collected with an ALOHA dual-arm system, as the source data and edit the ALOHA embodiment into our OpenArm appearance. We compare three MAE pretraining settings: unedited AIST videos, HyCustom-edited AIST-to-OpenArm videos, and RoboEdit-edited AIST-to-OpenArm videos.

For all settings, we freeze the resulting MAE encoder and train the same BC policy, with an action-chunk horizon of 16, on the same 200 self-collected OpenArm pick-and-place demonstrations: 100 with a lemon and 100 with a banana. We evaluate each policy over 20 real-robot trials per object using a lemon, a banana, and an unseen pear, which does not appear in the BC demonstrations. We vary the initial object position across trials, including near and far placements, to introduce different levels of difficulty.

Table 7: Preliminary real-robot success rates on OpenArm.

MAE pretraining data	Lemon	Banana	Pear (unseen)	Avg.
Unedited AIST	13/20	11/20	0/20	40.0%
HyCustom AIST-to-OpenArm	15/20	16/20	1/20	53.3%
RoboEdit AIST-to-OpenArm	**17/20**	**19/20**	**3/20**	**65.0%**

We omit Mirage-like from this experiment because its geometry-based pipeline requires a target-specific simulator model and calibrated camera–robot geometry, which are not available for our dual-arm OpenArm platform. HyCustom and RoboEdit, in contrast, can be applied directly to the same source videos without additional platform-specific simulation infrastructure. While limited in scale, the consistent gains across the three objects provide preliminary evidence that embodiment-aligned video editing can benefit visual representation transfer for downstream control.

O Additional Comparison to Embodiment-Transfer Baselines

H2R Li et al. (2025) retargets human hand motion to a predefined robot, renders the robot with simulator assets, and composites it into the source video. This gives explicit control over geometry and kinematics, but requires target-specific assets and does not jointly generate the robot with the surrounding scene. RoboSwap Bai et al. (2025) is closer to robot-arm replacement, but learns a separate translator for each source–target pair and requires target-domain data and retraining for new targets. Mirage-like and Ro-Aug baselines are useful downstream controls, but they rely on fixed embodiment-transfer assumptions. RoboEdit instead performs local reference-guided video editing from a single target image, making it applicable to arbitrary references, object replacement, and new embodiments without pair-specific translator training.

P Scope and Future Directions

RoboEdit is designed for reference-guided local visual editing of robotic manipulation videos. It can generate visually diverse variants of existing demonstrations by replacing a manipulator or manipulated object while preserving the surrounding context. However, this formulation also introduces several limitations. First, RoboEdit relies on edit masks to separate the editable region from the preserved scene. Inaccurate masks can blur this separation, leading to incomplete replacement, boundary artifacts, or unintended changes to nearby objects and contact regions. This limitation may be partially mitigated by interactive mask refinement or by adapting the segmentation model to robotic manipulation videos. Second, RoboEdit primarily constrains reference-induced changes outside the edit mask, while the masked region remains generative. This can lead to occasional in-mask hallucinations, such as small spurious structures near the replaced manipulator, as shown in Figure 12. Such artifacts are usually localized and do not substantially corrupt the background or motion dynamics, but they indicate that stronger in-mask geometric constraints could further improve local fidelity.

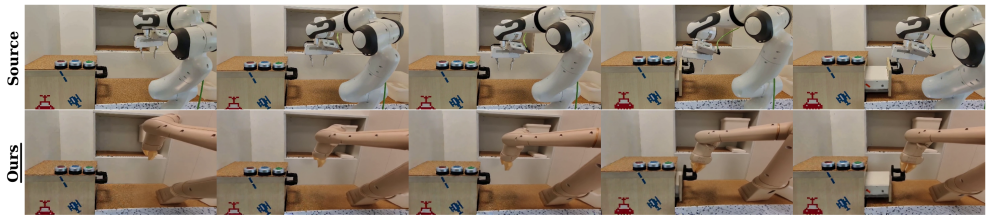

Figure 12: Example of in-mask hallucination. A small box-like structure appears behind the replaced manipulator within the edit mask, while the surrounding scene remains largely preserved.

Third, RoboEdit depends on the availability of a clear reference image, since occluded, incomplete, or ambiguous references can weaken appearance transfer and reduce editing fidelity. Finally, RoboEdit focuses on video-level visual augmentation rather than full cross-embodiment policy adaptation. The edited videos preserve the original motion dynamics, but different robot embodiments may have different kinematics, action spaces, or proprioceptive states. Thus, directly using the edited videos for policy learning may require additional alignment between the visual trajectory and the target embodiment's control space. This could be addressed by combining RoboEdit with trajectory retargeting, action relabeling, or learned embodiment-conditioned policy models, which we leave for future work.

Q BROADER IMPACT

RoboEdit aims to make robotic video data augmentation more scalable by locally replacing manipulators or objects in existing demonstrations. However, the synthetic robotic videos should not replace real-world validation. Edited videos may contain subtle artifacts or physically inconsistent details, especially when masks, references, or embodiment differences are imperfect. If used for robot training, such data require additional checks to ensure that task-relevant structure is preserved. More broadly, reference-guided video editing could be misused to create misleading robot demonstrations or overstate robotic capabilities. We encourage transparent disclosure when generated videos are used in datasets, benchmarks, or demos, and careful validation before deployment in safety-critical robotic applications.