TRIVE: TOKEN RESPONSIBILITY MODELING FOR STREAMING EGOCENTRIC VIDEO EDITING

Anonymous authors
Paper under double-blind review

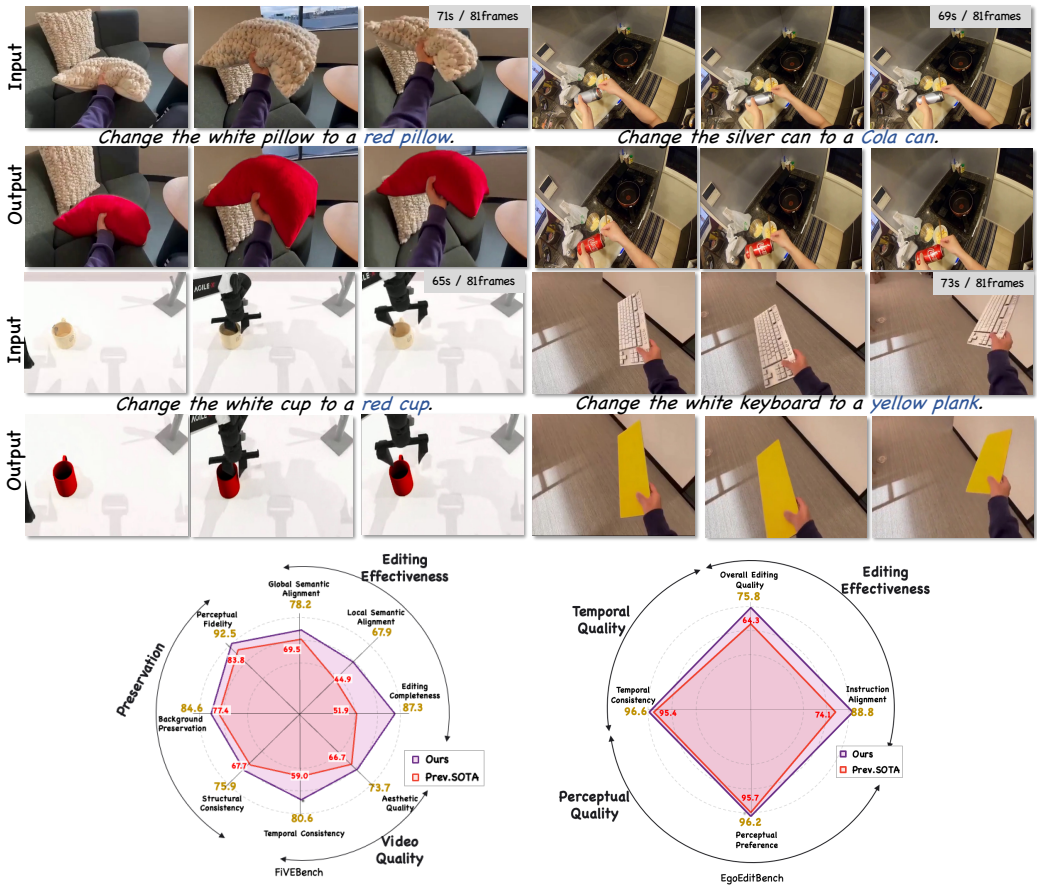

Figure 1: **Showcase of TRIVE.** Given a source interaction video, source and target prompts, and a hand representation, TRIVE performs streaming appearance editing while preserving the recorded interaction. Soft object, interface, hand, and background responsibilities regulate source retention, target-appearance access, and cross-block appearance consistency.

ABSTRACT

Egocentric video editing seeks to modify object appearance while preserving the recorded hand–object interaction, offering a practical route to visual data augmentation for augmented reality and embodied policy learning. This setting is challenging because latent video tokens near interaction boundaries often mix editable object content with hand structure that should remain unchanged. We introduce TRIVE, a framework for streaming egocentric video editing that models each token with soft object, interface, hand, and background responsibilities. These responsibilities form a shared representation from which we derive three operation-specific permissions: source retention, target-appearance access, and reference eligibility. We instantiate these permissions through role-aware attention control and a role-guided denoising update, including directional filtering that suppresses source corrections opposing the observed source-to-target response.

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.

To reduce appearance drift across causal blocks, we further construct a source-addressed appearance anchor from reliable object tokens and enable broader role-conditioned retrieval, including mixed interface tokens. TRIVE operates on a frozen streaming backbone and applies its responsibility-based controls during inference. Across egocentric and robot-manipulation benchmarks, TRIVE improves editing fidelity, source preservation, and temporal consistency over the evaluated baselines, while complementary hand- and interaction-focused diagnostics further support its preservation capability. Real-robot experiments further suggest that its edited demonstrations can improve policy performance under held-out visual conditions.

1 INTRODUCTION

Egocentric video is a central form of visual data for augmented reality (AR) and embodied AI, capturing the world from the perspective of a person or agent (Grauman et al., 2022; Li et al., 2026b). Compared with third-person footage, it directly records how actions unfold within the actor's workspace. During manipulation, hands often occupy a large foreground area, contact or occlude objects, and move together with a changing viewpoint. These observations connect visual appearance with action and physical interaction (Damen et al., 2018). Learning to understand activities, support interactive AR, and generalize manipulation skills therefore requires diverse egocentric data that faithfully captures these relationships. Acquiring such diversity at scale remains costly. Manual collection requires recording equipment, repeated demonstrations across objects and environments, and subsequent curation and annotation. Covering new object instances, users, and scene layouts adds further recording effort. Simulation and learned video generation offer synthetic alternatives (Li et al., 2024; Wang et al., 2025), but require scene construction or model training and must synthesize plausible hand–object dynamics along with appearance. Generated sequences also need validation of grasping, contact, and motion consistency. Video editing provides a complementary route: it reuses a recorded interaction and changes its visual content (Li et al., 2026b). Preserving the source motion, contact timing, and scene structure allows new appearance variations to be created without repeating the physical action, making editing attractive for data augmentation.

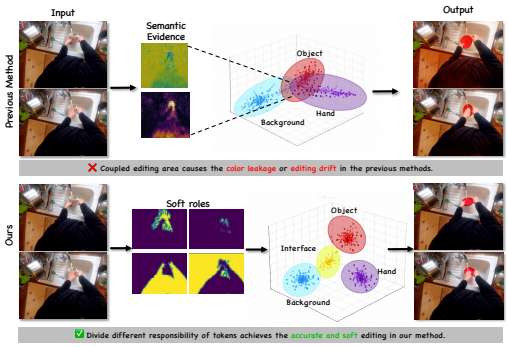

Figure 2: **Motivation.** Generic editing signals can entangle editable object content with the interacting hand in egocentric videos, leading to inaccurate edits and appearance leakage. TRIVE instead models soft interaction responsibilities to separate appearance adaptation from interaction preservation.

Recent advances in streaming video editing make this route more practical by processing video incrementally with few-step sampling. LiveEdit distills a causal editor and reuses computation in unedited regions to achieve low-latency editing (Wang et al., 2026b). StreamEdit adapts pretrained streaming generators for training-free editing (Jiao et al., 2026b). EgoEdit specifically targets first-person interactions through curated editing pairs and task-specific training (Li et al., 2026b). These methods demonstrate progress in both efficiency and visual quality. However, achieving reliable interaction preservation without constructing a specialized training set remains an open challenge.

General-purpose editors commonly infer editing regions from semantic or appearance-change signals. For example, StreamEdit uses object-word cross-attention maps for localization and source–target flow-velocity differences to modulate source preservation (Jiao et al., 2026b). Such signals do not explicitly distinguish object appearance from the hand that interacts with it. Under frequent contact and occlusion, latent compression further combines hand and object content within the same tokens. Editing responses can then extend into protected regions, producing unstable hand structure or target-color leakage, as illustrated in Figure 2. Strengthening source constraints can instead restore the original object appearance and weaken the edit. More accurate localization alone cannot resolve this conflict: even a correctly localized interface must retain interaction geometry while ac-

cepting new appearance. Editing therefore requires distinct preservation and adaptation controls for the same token.

We address this problem with an interaction-aware framework based on *token responsibility*. Semantic and generative responses are interpreted together with hand evidence and temporal source correspondence to represent soft object, interaction-interface, hand, and background roles. These overlapping roles determine how each token participates in editing: requiring source structure need not prevent a token from receiving target appearance. Role-guided attention regulates target-appearance access, while a role-guided denoising update retains source corrections according to interaction needs and filters directions that oppose the edit. This allows mixed interface tokens to preserve stronger source constraints than object-core tokens while still receiving the requested appearance change.

Causal generation also allows local errors to propagate through the evolving history, destabilizing appearance after occlusion. We therefore use reliable object-dominant tokens to establish a *source-addressed appearance anchor*, a fixed target-minus-source feature response that guides later blocks. Stricter reference eligibility lets mixed tokens receive appearance guidance without necessarily defining the reference themselves. Together, these controls support interaction-preserving appearance editing on a frozen streaming backbone, without task-specific training, parameter optimization, or inversion. We evaluate the framework on egocentric and robot-manipulation videos and examine its use for offline policy-training augmentation, where edited observations retain the original action labels.

Our contributions are:

- We introduce a novel framework for streaming egocentric video editing based on interaction-aware token responsibility. Its soft representation captures overlapping object, interface, hand, and background roles, distinguishing the requirements of source preservation, appearance adaptation, and persistent reference construction within the same token.
- We develop complementary role-guided attention and denoising controls for mixed hand–object tokens. Target-appearance access is regulated jointly with directional filtering and role-dependent allocation of source reconstruction residuals, allowing stronger interaction constraints at contact regions while retaining the capacity for appearance changes.
- We propose a source-addressed appearance anchor that stores a fixed target-minus-source feature response from reliable object tokens. Selective reference construction, broader role-conditioned retrieval, and correction against the current response provide persistent appearance guidance across causal blocks alongside the evolving generation history.

2 Related Work

Video editing. Several editors preserve source structure by reusing representations from pretrained video models. FateZero transfers spatiotemporal attention, TokenFlow propagates diffusion-feature correspondences, and VideoGrain modulates attention at multiple spatial and temporal scales (Qi et al., 2023; Geyer et al., 2024; Yang et al., 2025b). With rectified-flow backbones, UniEdit-Flow improves inversion and feature injection, while FlowDirector steers the generation trajectory without inversion (Jiao et al., 2026a; Li et al., 2025a). These methods provide strong generic editing priors, but they typically process a fixed clip and treat editability as a spatial decision. In hand–object interaction, however, one latent token may simultaneously contain editable object appearance and hand structure that must be preserved. Our work instead models overlapping token responsibilities and converts them into operation-specific controls.

Streaming generation and editing. Causal video generators synthesize frames or short blocks under bounded history, enabling low-latency and potentially open-ended generation. Diffusion Forcing connects diffusion with autoregressive prediction, while CausVid, Self Forcing, and StreamDiT distill or train bidirectional video models for causal rollout (Chen et al., 2024; Yin et al., 2025; Huang et al., 2025; Kodaira et al., 2026). Long-horizon systems further manage finite context through sinks, recaching, or compressed state (Liu et al., 2025; Yang et al., 2025a; Yi et al., 2025). StreamEdit repurposes a few-step streaming generator for training-free editing through paired source and target generation (Jiao et al., 2026b). We build on this causal setting, but address a distinct failure mode:

preservation and appearance adaptation conflict at mixed interaction tokens, while evolving causal history provides no fixed reference for the edited appearance.

Egocentric interaction synthesis. Egocentric generation models synthesize first-person observations from scene, pose, or action conditions (Li et al., 2024; Hassan et al., 2025; Wang et al., 2025; Bai et al., 2025). Editing an observed interaction imposes a stricter constraint because the source hand motion, contact timing, object dynamics, and camera trajectory should remain unchanged. HOI-Swap studies object replacement through a modified keyframe and motion-conditioned propagation (Xue et al., 2024). EgoEdit introduces paired supervision for bidirectional and real-time egocentric editors with explicit hand preservation (Li et al., 2026b). In contrast, we study training-free causal editing of an already recorded interaction. Rather than generating a new action, our method determines how source reconstruction, target appearance, and future reference construction should treat each mixed token. An extended discussion of these areas is provided in Appendix D.

3 METHODOLOGY

3.1 OVERVIEW

Given a source video X_s, source and target prompts c_s, c_t, their manipulated-object expressions w_s, w_t, and a hand representation $\mathcal{H}$ comprising spatial support and temporally aligned motion cues, our goal is to realize the target appearance while preserving the recorded hand–object interaction, camera motion, and non-target scene content.

Our key idea is to explicitly represent each token's responsibility, and then use this representation to guide how editing operations act on that token. We infer a shared *interaction responsibility* field over object, interaction-interface, hand, and background roles. This field acts as a common representation from which we derive three distinct permissions: source retention, target-appearance access, and reference construction. Within a block, role-aware attention control and a role-guided denoising update apply the first two permissions without forcing them to coincide spatially. Across blocks, a source-addressed appearance anchor uses the stricter reference permission to preserve a fixed appearance response. Figure 3 summarizes this formulation: *the same mixed token can require different operational permissions*.

3.2 INTERACTION RESPONSIBILITY REPRESENTATION

We describe a token along two axes: its association with the manipulated object and its occupancy by the hand or gripper. Let $p_i, h_i \in [0, 1]$ denote these quantities. Actor occupancy accounts for the pixel frames compressed into a causal latent, so a token can carry substantial evidence of both the actor and the object. We express this overlap through four soft responsibilities:

$$\begin{aligned} q_i^{obj} &= p_i(1-h_i), \quad q_i^{int} = p_i h_i, \\ q_i^{hand} &= (1-p_i)h_i, \quad q_i^{bg} = (1-p_i)(1-h_i). \end{aligned} \tag{1}$$

The entries of $\mathbf{q}_i$ sum to one and form a factorized responsibility field induced by the two evidence axes; they are not independently learned class posteriors. Interface responsibility records overlapping evidence within the latent representation rather than assigning a token exclusively to the hand or object. Normalized role entropy $u_i = H(\mathbf{q}_i)/\log 4$ moderates interventions when responsibility is dispersed; it measures role mixture rather than calibrated prediction error.

Object support is initialized from source-expression attention and temporal source correspondence, while the hand representation identifies interaction overlap and constrains the support. Together these signals define a connected, source-grounded extent and preserve support when visibility weakens. The first paired prediction then supplies branch-response evidence to refine confidence *within this extent*; off-support responses cannot independently authorize new edit regions. Appendix B.1 details the evidence construction and refinement.

3.3 FROM RESPONSIBILITIES TO OPERATIONAL PERMISSIONS

Paired source and target predictions expose two controls at each denoising step: reconstruction acts on the velocity update, whereas appearance access acts on the features predicting that update. The

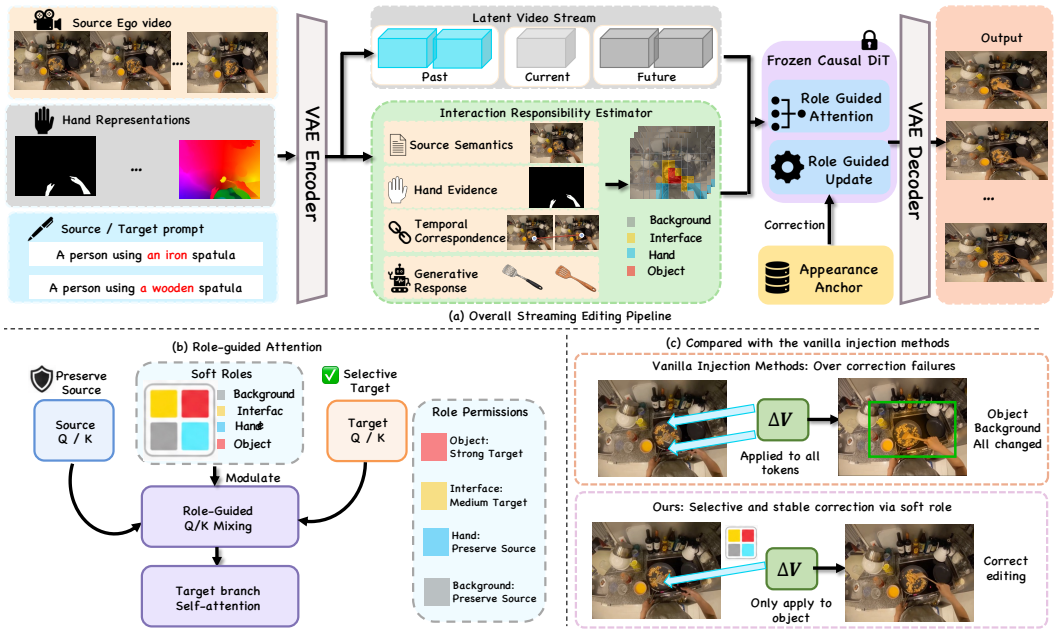

Figure 3: **Overview of TRIVE.** (a) Soft interaction responsibilities are inferred from source, hand, temporal, and generative evidence to guide causal editing on a frozen streaming backbone. (b) Role-aware attention controls target-appearance access, while the role-guided update independently allocates source reconstruction, enabling mixed interaction tokens to preserve structure without blocking appearance change. (c) A source-addressed appearance anchor is written once from reliable object tokens and provides bounded response correction to later chunks, maintaining the edited appearance alongside the evolving causal history.

shared responsibility state therefore gives interface tokens stronger source constraint than object-core tokens without blocking target appearance (Figure 4).

Role-guided denoising update. The source residual $\boldsymbol{\rho}_i = \mathbf{v}_i^{src,*} - \mathbf{v}_i^{src}$ corrects the predicted source velocity but can also restore the appearance being replaced. We remove its component opposing the branch response $\mathbf{d}_i = \mathbf{v}_i^{trg} - \mathbf{v}_i^{src}$ to obtain $\boldsymbol{\rho}_i^{safe}$, preserving non-opposing source correction without assuming appearance–geometry disentanglement.

The same roles determine the coefficient s_i of this filtered correction and the appearance-access score a_i:

$$\begin{aligned} s_i &= \alpha_o q_i^{obj} + \alpha_c q_i^{int}, & 0 \le \alpha_o < \alpha_c \le 1, \\ a_i &= (q_i^{obj} + \lambda_{\text{mix}} q_i^{int})(1 - u_i), & 0 < \lambda_{\text{mix}} < 1. \end{aligned} \tag{2}$$

Thus the interface receives stronger filtered source correction but discounted appearance access relative to the object role. Hand/background handling and the entropy fallback are specified in Appendix B.2; "safe" denotes non-opposition to the observed response, not semantic correctness.

Role-aware attention control. We restrict a_i to connected object support and augment it with correspondence-supported object access to obtain g_i^{acc}. Source correspondence can therefore sustain access when the role score alone is weak. This weight increases the target contribution to current-token query/key blending while source correction remains separately allocated. Refined roles apply after the first paired prediction, whereas historical keys retain the backbone schedule, making retention and adaptation compatible permissions at the same location.

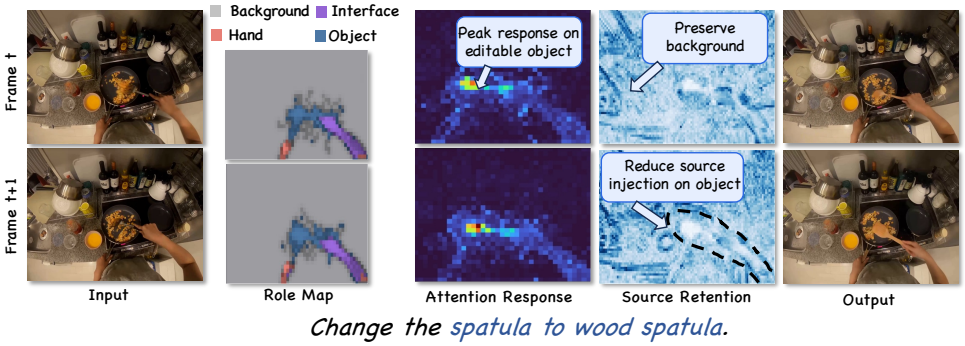

Figure 4: **Visualization of role-aware editing control.** Soft token roles induce distinct attention and source-retention patterns: target-appearance access concentrates on the editable object, while source information is retained more strongly in hand and background regions. Mixed interface tokens retain stronger source constraints than object-core tokens while remaining accessible to the target appearance.

3.4 Source-Addressed Appearance Anchor

Across blocks, occlusion and imperfect edits can contaminate an evolving causal history. Moreover, a mixed token may need access to the edited appearance even when its own response is unsuitable for defining that appearance in the future. We therefore retain a fixed reference constructed from selected source-supported responses alongside the dense causal history.

Separating a reference's address from its appearance. At attention layer ℓ, source keys address the appearance response encoded by clean target–source value differences:

$$\mathcal{M}^{\ell} = \left\{ \left(\mathbf{K}_j^{\ell,src,*}, \mathbf{V}_j^{\ell,trg,*} - \mathbf{V}_j^{\ell,src,*} \right) : j \in \mathcal{W} \right\}, \tag{3}$$

where $*$ denotes clean-state features and $\mathcal{W}$ the write set. The anchor is initialized after the first eligible completed block and then fixed, avoiding replacement by later generated responses.

Writing requires correspondence-supported object ownership, role-supported object extent, and limited hand evidence; persistent hand occupancy is excluded. Reading instead uses g_i^{acc} together with valid source matches, allowing mixed tokens to receive the appearance reference without qualifying as writers. This selective-write, broader-read policy separates the evidence needed to define an appearance from the evidence needed to receive it; Appendix B.3 gives the complete rules.

Correcting the response already present. A clean-source query independently retrieves reference and current responses, $\mathbf{D}_i^{ref}$ and $\mathbf{D}_i^{cur,\tau}$, by source address. Their bounded difference corrects target attention:

$$\mathbf{O}'_i = \mathbf{O}_i^{native} + \lambda_{\text{anc}} g_i^{acc} c_i^{addr} \, \text{Clip}_1 \left(\mathbf{D}_i^{ref} - \mathbf{D}_i^{cur,\tau} \right). \tag{4}$$

Here c_i^{addr} requires valid source matches in both searches, λ_{anc} sets the correction strength, and unit-relative clipping bounds its magnitude relative to the retrieved responses. The correction vanishes when the two responses agree. Because the reference is clean while the current response depends on noise time, their discrepancy serves as bounded feature feedback rather than a calibrated appearance error.

The same correction guides authorized clean-target values committed to causal history, carrying its influence into subsequent generation. An anchor is created only after its source block completes and becomes available from the next block onward; Appendix B.4 gives the schedule.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Protocol and implementation. We evaluate complete egocentric and robot-manipulation videos, asking whether target appearance can change while hand or gripper motion, contact geometry, and non-target content remain intact. Our set contains 200 videos: 50 from EgoEditBench (Li et al., 2026b), 50 collected by us, 50 from RoboTwin (Mu et al., 2025), and 50 from RoboMIND (Wu et al., 2025a). Each 81-frame source is paired with two target instructions, yielding 400 video–instruction pairs that cover changes in color, material, and object appearance. Every source is processed over seven causal blocks. We use a frozen Self-Forcing backbone without task-specific training, inversion, or per-video optimization at 832×480 with 15 sampling steps and shared source–target noise. Dataset construction and complete configurations are in Appendices C.1 and C.2.

Baselines. Our baselines span feature propagation, attention and flow control, and streaming generation: TokenFlow (Geyer et al., 2024), Pyramid-Edit and Wan-Edit (Li et al., 2025b), UniEdit-Flow (Jiao et al., 2026a), StreamDiffusionV2 (Feng et al., 2025), and StreamEdit (Jiao et al., 2026b). LucyEdit (Decart AI Team, 2025) and LiveEdit (Wang et al., 2026b) are trained cross-paradigm references; EgoEdit (Li et al., 2026b) is not reproduced because no public checkpoint is available. All methods receive identical videos and target instructions, with only release-required auxiliary inputs.

4.2 COMPARISON WITH BASELINES

Quantitative comparison. Following StreamEdit (Jiao et al., 2026b) and FiVE-Bench (Li et al., 2025b), Table 1 evaluates structure, background preservation, text alignment, image quality, temporal consistency, editing accuracy, and runtime; metric definitions and aggregation are given in Appendix C.3. Ours achieves the strongest results across all quality metrics while retaining streaming inference speed. Additional evaluation on the appearance-editing subset of EgoEditBench is provided in Appendix C.5.

Table 1: **Quantitative comparison following StreamEdit and FiVE-Bench.** Structure Distance and background metrics evaluate source preservation; CLIP_S, CLIP_S^{edit}, and FiVE-Acc evaluate editing effectiveness; NIQE and temporal consistency measure visual quality and motion fidelity. Time is reported in seconds per frame. The best and second-best results are marked in **red** and **blue**, respectively.

Method	Structure	Background Preservation				Text Alignment		IQA	Temp.	FiVE	Time (s)
	Dist.$\times 10^3 \downarrow$	PSNR$\uparrow$	LPIPS$\times 10^3 \downarrow$	MSE$\times 10^4 \downarrow$	SSIM$\times 10^2 \uparrow$	$\text{CLIP}_S \uparrow$	$\text{CLIP}_S^{edit} \uparrow$	NIQE$\downarrow$	Cons.$\times 10^2 \uparrow$	Acc.$\uparrow$	Per Frame$\downarrow$
TokenFlow	19.14	20.75	**38.31**	432.23	82.28	26.88	22.74	13.57	84.07	65.48	137.06
Pyramid-Edit	9.09	24.55	117.71	575.03	72.28	27.47	23.40	12.33	90.65	75.60	75.00
Wan-Edit	11.22	24.68	117.28	573.46	71.79	27.52	23.48	12.16	89.96	75.60	3.89
UniEdit-Flow	**5.63**	19.07	99.67	652.38	70.50	27.55	**24.72**	10.91	**95.30**	**84.52**	44.37
StreamDiffusionV2	16.85	27.15	61.80	236.40	86.95	27.20	23.15	13.40	94.70	66.80	**0.12**
StreamEdit	6.47	**28.47**	49.84	218.59	87.52	26.95	21.73	10.33	91.80	51.94	0.89
LucyEdit	8.72	26.80	70.40	250.15	86.20	27.58	24.10	10.70	94.20	80.40	18.60
LiveEdit	7.15	27.90	55.20	**194.60**	**88.10**	**27.61**	24.40	**10.12**	95.05	81.60	**0.08**
Ours	**4.82**	**30.12**	**34.72**	**185.32**	**89.63**	**27.82**	**25.18**	**9.84**	**96.12**	**87.35**	0.93

Qualitative comparison. Figure 5 compares representative interaction edits. Generic editors frequently under-edit the manipulated object, propagate the target appearance into the hand, or alter the source interaction when the two overlap. In contrast, our method assigns different source-retention and appearance-access permissions to mixed tokens, allowing the requested appearance change while retaining the recorded hand–object relation. Additional videos and comparisons are provided in Appendix A.

Real-Robot Execution Evaluation. We evaluate offline video augmentation on four AgileX Cobot Magic skills: Lift, Pick, Capture, and Put (Fig. 6). Edited observations retain each source demonstration's actions and proprioceptive states and are added only to the policy-training set. We compare the original demonstrations with matched StreamEdit (Jiao et al., 2026b) and Ours augmentations under identical policy architecture, optimization budget, demonstration count, and test initializations. Held-out execution changes object or workspace appearance; the editor is absent at deployment. Table 2 reports 20 trials per task.



Figure 5: **Qualitative comparison.** The results show appearance edits under strong hand–object interaction and occlusion. TRIVE more consistently realizes the requested object appearance while reducing changes to the interacting hand and non-target content.

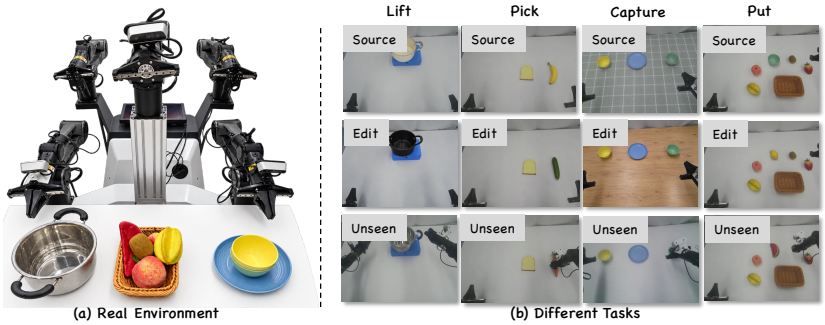

Figure 6: **Real-robot setup and evaluation protocol.** (a) The AgileX Cobot Magic platform and evaluation objects. (b) Four task families. Each column shows a source demonstration (top), its offline edited training variant (middle), and execution under a held-out visual condition (bottom).

Table 2: **Real-robot execution success rate.** We report success rate (%) over 20 trials per task under held-out visual conditions. Edited videos are used only for offline policy training; no editor is used at deployment.

Policy Training Data	Lift	Pick	Capture	Put	Avg.
Original Demonstrations	55.0	45.0	35.0	40.0	43.8
Original + StreamEdit	**65.0**	**50.0**	**45.0**	**50.0**	**52.5**
Original + Ours	**80.0**	**70.0**	**60.0**	**70.0**	**70.0**

StreamEdit augmentation yields a higher average success rate than training on the original demonstrations alone, suggesting that additional visual variation can be beneficial under the evaluated conditions. TRIVE augmentation yields higher success rates than StreamEdit across the four evaluated tasks. Together with the editing-preservation diagnostics, these results are consistent with TRIVE providing useful label-preserving visual augmentation.

4.3 Ablation Study

Effectiveness of Interaction Responsibility. Without interaction responsibility, the editor retains the aligned hand evidence but does not decompose tokens into object, interface, hand, and background roles. As shown in Table 3, this produces the largest overall degradation: strengthening the edit contaminates the hand–object interaction, whereas strengthening preservation leaves source

appearance in manipulated objects. We separately isolate the representation itself through input-matched hard-versus-soft responsibility comparisons in Appendix C.7.

Effectiveness of Role-Aware Attention Control. Removing RAC restores the native source–target attention coupling while retaining responsibility inference, RGU, and the appearance anchor. The target branch consequently loses token-wise appearance access: object-core and interface tokens receive the same blending policy despite requiring different source constraints. This primarily weakens CLIP_S^{edit} and FiVE-Acc and yields incomplete or spatially uneven appearance changes around the interaction boundary.

Table 3: **Quantitative ablation.** RAC denotes role-aware attention control and RGU the role-guided denoising update.

Metric	w/o Resp.	w/o RAC	w/o RGU	w/o Anchor	Full
PSNR↑	23.42	27.76	25.31	**29.38**	**30.12**
LPIPS$\times 10^3$ ↓	111.60	61.35	87.90	**46.85**	**34.72**
SSIM$\times 10^2$ ↑	74.85	86.48	80.36	**88.91**	**89.63**
CLIP_S ↑	26.82	27.08	27.39	**27.65**	**27.82**
CLIP_S^{edit} ↑	22.48	23.71	24.16	**24.91**	**25.18**
FiVE-Acc↑	63.70	75.80	78.40	**84.80**	**87.35**

Effectiveness of the Role-Guided Update. Without RGU, the model uses the native denoising update even though roles remain available to attention and reference retrieval. Source reconstruction residuals are no longer filtered against the current source–target branch response or allocated according to token responsibility. The resulting edits retain target semantics but more readily perturb the hand, contact geometry, and non-target content, leading to larger degradation in PSNR, LPIPS, and SSIM.



Figure 7: **Qualitative ablation.** Representative sequences expose the effects of each component on appearance transfer, hand–object structure preservation, and long-horizon consistency.

Effectiveness of the Appearance Anchor. Removing the appearance anchor leaves both current-block controls intact but eliminates the fixed target-response reference used by later blocks. Early chunks therefore remain close to the full model, while appearance gradually drifts with the evolving causal history and becomes less reliable after occlusion and reappearance. This explains why the anchor-free variant remains the strongest ablation in frame-averaged metrics while exhibiting clear long-horizon inconsistency in the qualitative sequence.

5 Conclusion

We introduced TRIVE, a novel framework for streaming egocentric video editing that addresses the conflict between modifying object appearance and preserving hand–object interaction within mixed latent tokens. TRIVE represents object, interface, hand, and background responsibilities as soft roles and maps this shared representation to three operation-specific permissions: source retention, target-appearance access, and reference construction. Role-aware attention and role-guided denoising controls regulate current-block editing, while a source-addressed appearance anchor reduces appearance drift across causal blocks without recursively redefining the reference from evolving generated history. Experiments on egocentric and robot-manipulation videos show improved editing, preservation, and temporal-consistency metrics under the evaluated protocols, and component analyses support the complementary contributions of the proposed mechanisms. Real-robot results further suggest that augmenting policy training with our edited videos while reusing source actions improves performance under held-out visual conditions. Overall, TRIVE shows that separating operational permissions at mixed interaction tokens provides an effective formulation for interaction-aware egocentric video editing.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing, brainstorming, and limited code assistance. We did not use generative AI tools to generate experimental results, fabricate data, or make final scientific claims, and other disclosure categories are not applicable to this work. All AI-assisted text, code, and suggestions were manually reviewed, revised, and verified by the authors. We take full responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies training-free editing of egocentric and robot-manipulation videos. Such editing may support visual data augmentation and embodied policy learning, but realistic synthetic interaction videos could also be used to fabricate actions or misrepresent human or robotic capabilities. Generated content should therefore be clearly identified, and policies trained with edited data should not be deployed in safety-critical settings without human oversight and real-world validation. Participation in the perceptual user study is voluntary, and no sensitive personal information is collected. We will follow the licenses and terms of use of all source datasets and document the intended use, limitations, and potential misuse risks of released code, annotations, and generated data.

REPRODUCIBILITY STATEMENT

TRIVE is training-free and does not update the weights of its frozen streaming backbone. Section 3 specifies the token-responsibility representation, role-guided denoising update, role-aware attention control, and source-addressed appearance anchor. The supplementary material provides the role-construction procedure, causal execution order, control parameters, evaluation-set construction, prompt and mask preprocessing, baseline protocol, metric definitions, user-study design, and real-robot policy protocol. We will release the implementation, configuration files, data-processing and evaluation scripts, and evaluation manifests needed to reproduce the reported results, subject to the licenses of the source models and datasets.

REFERENCES

Yutong Bai, Danny Tran, Amir Bar, Yann LeCun, Trevor Darrell, and Jitendra Malik. Whole-body conditioned egocentric video prediction. In *Advances in Neural Information Processing Systems*, volume 38, pp. 164375–164418, 2025.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. In *Advances in Neural Information Processing Systems*, volume 37, pp. 24081–24125, 2024.

Yiyang Chen, Xuanhua He, Xiujun Ma, and Yue Ma. Contextflow: Training-free video object editing via adaptive context enrichment. *arXiv preprint arXiv:2509.17818*, 2025.

Jiaxin Cheng, Tianjun Xiao, and Tong He. Consistent video-to-video transfer using synthetic dataset. *arXiv preprint arXiv:2311.00213*, 2023.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The EPIC-KITCHENS dataset. In *Proceedings of the European Conference on Computer Vision*, 2018.

Decart AI Team. Lucy edit: Open-weight text-guided video editing. *Technical report*, 2025.

Tianrui Feng, Zhi Li, Shuo Yang, Haocheng Xi, Muyang Li, Xiuyu Li, Lvmin Zhang, Keting Yang, Kelly Peng, Song Han, Maneesh Agrawala, Kurt Keutzer, Akio Kodaira, and Chenfeng Xu. StreamDiffusionV2: A streaming system for dynamic and interactive video generation. *arXiv preprint arXiv:2511.07399*, 2025.

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.

Kristen Grauman et al. Ego4D: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18973–18990, 2022.

Mariam Hassan, Sebastian Stapf, Ahmad Rahimi, Pedro M. B. Rezende, Yasaman Haghighi, David Brüggemann, Isinsu Katircioglu, Lin Zhang, Xiaoran Chen, Suman Saha, Marco Cannici, Elie Aljalbout, Botao Ye, Xi Wang, Aram Davtyan, Mathieu Salzmann, Davide Scaramuzza, Marc Pollefeys, Paolo Favaro, and Alexandre Alahi. Gem: A generalizable ego-vision multimodal world model for fine-grained ego-motion, object dynamics, and scene composition control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22404–22415, 2025.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In *Advances in Neural Information Processing Systems*, 2025.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17191–17200, 2025.

Guanlong Jiao, Biqing Huang, Kuan-Chieh Wang, and Renjie Liao. Uniedit-flow: Unleashing inversion and editing in the era of flow models. In *International Conference on Learning Representations*, 2026a.

Guanlong Jiao, Chenyangguang Zhang, Jia Jun Cheng Xian, Zewei Zhang, and Renjie Liao. Streamedit: Training-free video editing via few-step streaming video generation. *arXiv preprint arXiv:2605.21466*, 2026b.

Xuan Ju, Tianyu Wang, Yuqian Zhou, He Zhang, Qing Liu, Nanxuan Zhao, Zhifei Zhang, Yijun Li, Yuanhao Cai, Shaoteng Liu, Daniil Pakhomov, Zhe Lin, Soo Ye Kim, and Qiang Xu. Editverse: Unifying image and video editing and generation with in-context learning. *arXiv preprint arXiv:2509.20360*, 2025.

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*, pp. 6507–6516, 2024.

Akio Kodaira, Tingbo Hou, Ji Hou, Markos Georgopoulos, Felix Juefei-Xu, Masayoshi Tomizuka, and Yue Zhao. Streamdit: Real-time streaming text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Max Ku, Cong Wei, Weiming Ren, Harry Yang, and Wenhu Chen. Anyv2v: A tuning-free framework for any video-to-video editing tasks. *Transactions on Machine Learning Research*, 2024.

Gen Li, Kaifeng Zhao, Siwei Zhang, Xiaozhong Lyu, Mihai Dusmanu, Yan Zhang, Marc Pollefeys, and Siyu Tang. Egogen: An egocentric synthetic data generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14497–14509, 2024.

Guangzhao Li, Yanming Yang, Chenxi Song, and Chi Zhang. Flowdirector: Training-free flow steering for precise text-to-video editing. *arXiv preprint arXiv:2506.05046*, 2025a.

Minghan Li, Chenxi Xie, Yichen Wu, Lei Zhang, and Mengyu Wang. FiVE: A fine-grained video editing benchmark for evaluating emerging diffusion and rectified flow models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025b.

Ruibin Li, Tao Yang, Fangzhou Ai, Tianhe Wu, Shilei Wen, Bingyue Peng, and Lei Zhang. Long-horizon streaming video generation via hybrid attention with decoupled distillation. *arXiv preprint arXiv:2604.10103*, 2026a.

Runjia Li, Moayed Haji-Ali, Ashkan Mirzaei, Chaoyang Wang, Arpit Sahni, Ivan Skorokhodov, Aliaksandr Siarohin, Tomas Jakab, Junlin Han, Sergey Tulyakov, Philip Torr, and Willi Menapace. Egoedit: Dataset, real-time streaming model, and benchmark for egocentric video editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026b.

Zekun Li, Xiaoyan Cong, Hongyu Li, Zhiyang Dou, Chuan Guo, Abhay Mittal, Sizhe An, and Srinath Sridhar. Ms. Forcing: Efficient streaming video generation with multi-scale patchification and attention. *arXiv preprint arXiv:2607.20940*, 2026c.

Zhiyuan Li, Chi-Man Pun, Peng-Tao Jiang, Bo Li, and Xiaodong Cun. Editalive! unified character video editing for live streaming. *arXiv preprint arXiv:2608.27123*, 2026d.

Yiqi Lin, Guoqiang Liang, Ziyun Zeng, Zechen Bai, Yanzhe Chen, and Mike Zheng Shou. Kiwi-edit: Versatile video editing via instruction and reference guidance. *arXiv preprint arXiv:2603.02175*, 2026.

Jinxiu Liu, Xuanming Liu, Kangfu Mei, Yandong Wen, Ming-Hsuan Yang, and Weiyang Liu. Streaming autoregressive video generation via diagonal distillation. In *International Conference on Learning Representations*, 2026a.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. *arXiv preprint arXiv:2509.25161*, 2025.

Qi Liu, Gang Yue, Mingyu Yin, Lisai Zhang, Yidi Wu, Yaole Wang, Yaohui Wang, Chang Yao, Jingyuan Chen, and Lin Ma. Tide: Task-isolated diffusion for unified video editing and generation. *arXiv preprint arXiv:2606.08260*, 2026b.

Yao Mu, Tianxing Chen, Zanxin Chen, Shijia Peng, Zhiqian Lan, Zeyu Gao, Zhixuan Liang, Qiaojun Yu, Yude Zou, Mingyu Ding, and Ping Luo. Robotwin: Dual-arm robot benchmark with generative digital twins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

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.

Yuanpeng Tu, Hao Luo, Xi Chen, Xiang Bai, Fan Wang, and Hengshuang Zhao. Playerone: Egocentric world simulator. In *Advances in Neural Information Processing Systems*, volume 38, pp. 145235–145261, 2025.

Weicheng Wang, Zhicheng Zhang, Zhongqi Zhang, Juncheng Zhou, Yongjie Zhu, Wenyu Qin, Meng Wang, Pengfei Wan, and Jufeng Yang. Live: Leveraging image manipulation priors for instruction-based video editing. *arXiv preprint arXiv:2604.17021*, 2026a.

Xiaofeng Wang, Kang Zhao, Feng Liu, Jiayu Wang, Guosheng Zhao, Xiaoyi Bao, Zheng Zhu, and Yingya Zhang. Egovid-5m: A large-scale video-action dataset for egocentric videos generation. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Xinyu Wang, Chongbo Zhao, Fangneng Zhan, and Yue Ma. Liveedit: Towards real-time diffusion-based streaming video editing. *arXiv preprint arXiv:2606.26740*, 2026b.

Kun Wu, Chengkai Hou, Jiaming Liu, Zhengping Che, Xiaozhu Ju, Zhuqin Yang, Meng Li, Yinuo Zhao, Zhiyuan Xu, Guang Yang, et al. Robomind: Benchmark on multi-embodiment intelligence normative data for robot manipulation. In *Robotics: Science and Systems*, 2025a.

Yuhui Wu, Liyi Chen, Ruibin Li, Shihao Wang, Chenxi Xie, and Lei Zhang. Insvie-1m: Effective instruction-based video editing with elaborate dataset construction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16692–16701, 2025b.

Yicheng Xiao, Wenxun Dai, Xinran Qin, Lin Song, Maoquan Zhang, Hang Xu, Yukang Chen, Yitong Li, Guohui Zhang, Yuan Zhang, et al. Joyai-video-edit: Real-time open-ended video editing with autoregressive diffusion. *arXiv preprint arXiv:2608.03974*, 2026.

Zihui Xue, Mi Luo, Changan Chen, and Kristen Grauman. Hoi-swap: Swapping objects in videos with hand-object interaction awareness. In *Advances in Neural Information Processing Systems*, volume 37, pp. 77132–77164, 2024.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, Song Han, and Yukang Chen. Longlive: Real-time interactive long video generation. *arXiv preprint arXiv:2509.22622*, 2025a.

Xiangpeng Yang, Linchao Zhu, Hehe Fan, and Yi Yang. Videograin: Modulating space-time attention for multi-grained video editing. In *International Conference on Learning Representations*, 2025b.

Zixuan Ye, Xuanhua He, Quande Liu, Qiulin Wang, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, Qifeng Chen, and Wenhan Luo. Unic: Unified in-context video editing. *arXiv preprint arXiv:2506.04216*, 2025.

Jung Yi, Wooseok Jang, Paul Hyunbin Cho, Jisu Nam, Heeji Yoon, and Seungryong Kim. Deep forcing: Training-free long video generation with deep sink and participative compression. *arXiv preprint arXiv:2512.05081*, 2025.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22963–22974, 2025.

Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. *arXiv preprint arXiv:2504.12626*, 2025.

Yixuan Zhu, Haolin Wang, Shilin Ma, Wenliang Zhao, Yansong Tang, Lei Chen, and Jie Zhou. Fade: Frequency-aware diffusion model factorization for video editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

A Additional Qualitative Results

Figures 8–10 provide additional qualitative results across diverse egocentric and robot-manipulation videos. Figure 11 shows a representative limitation under severe initial occlusion.

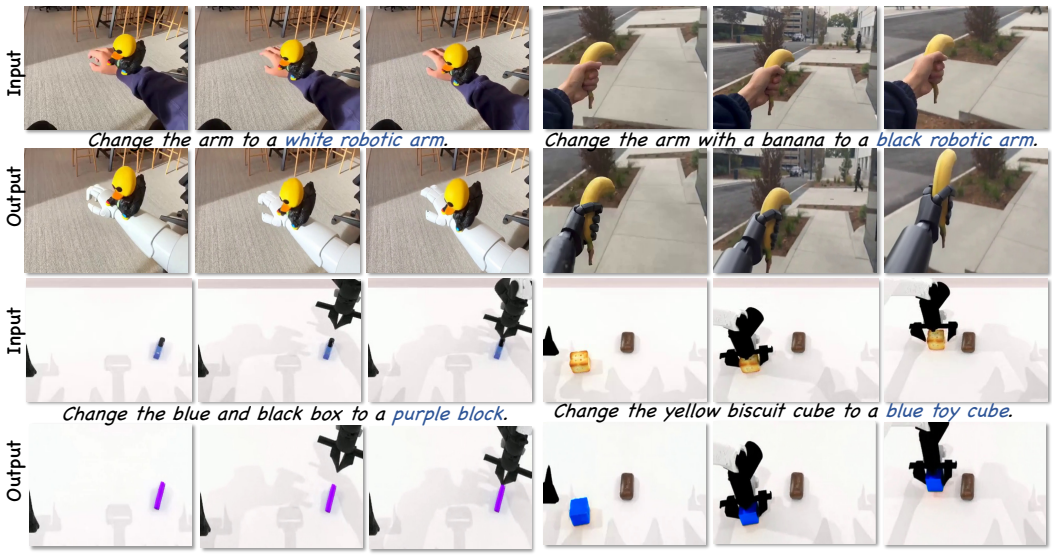

Figure 8: **Additional qualitative results.** Our method changes the appearance of manipulated objects across diverse environments while retaining the source interaction and temporal progression.



Figure 9: **Additional qualitative results.** Results on egocentric human and robot-manipulation videos with varied target objects and appearance instructions.

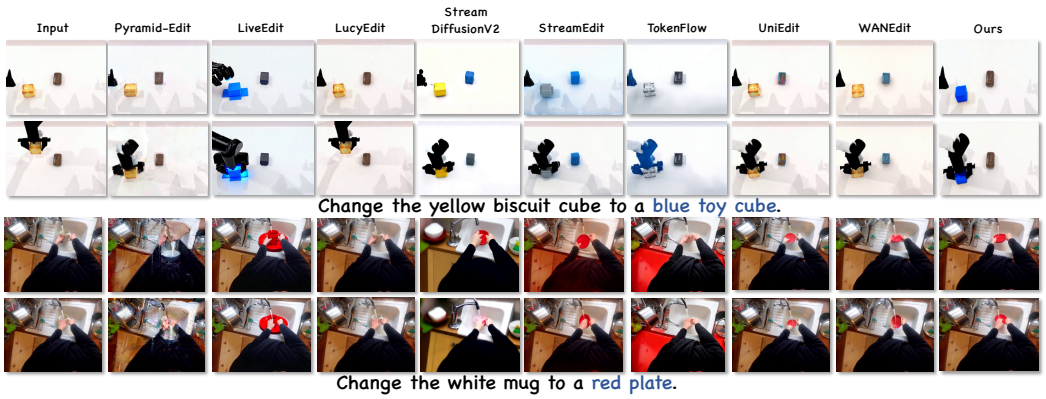

Figure 10: **Additional qualitative comparison.** We compare manipulation-video editing results under object-appearance changes.

A.1 Failure Case

TRIVE relies on source-grounded object evidence to initialize token responsibilities before propagating them through the video. When the manipulated object is already severely occluded by the hand in the first frame, the visible evidence may be insufficient to localize its object tokens reliably. Because subsequent propagation and anchor construction remain constrained by this initial support, missing object regions may not recover even after they become visible, resulting in an incomplete target-appearance edit. A multi-frame initialization strategy that delays role assignment until sufficient object evidence is available could mitigate this limitation.



Figure 11: **Failure under severe initial occlusion.** The hand covers most of the manipulated bottle in the first frame, preventing reliable initialization of the corresponding object tokens. Consequently, the requested black appearance is cannot be established.

B Supplementary Method Details

This appendix expands the evidence construction, control rules, and reference operations in Sec. 3. We use b for a causal block, t for a latent video frame, i for a spatial location, and τ for noise time. Temporal and layer indices are omitted where unambiguous.

B.1 Constructing and Refining Interaction Roles

Hand representation and current observation. The input hand representation $\mathcal{H}$ comprises a per-frame hand-mask sequence together with its temporally aligned motion/flow cues. For robot videos, gripper evidence is represented in the same form. The mask component is aligned with the causal VAE: the first latent represents the first pixel frame, and subsequent latents represent temporal frame groups. Within each group, the temporal union identifies where the hand has been present and supplies the proximity cue $d_{t,i} \in [0,1]$; temporal averaging defines fractional hand occupancy $h_{t,i} \in [0,1]$. Persistent support records locations occupied throughout the group for hard-exclusion decisions, while the aligned motion cues retain the temporal evolution of that support. Subsequent controls consume these derived statistics rather than treating a single binary mask as the complete hand representation.

A clean-source forward pass provides object-word attention and source query features. Per-frame spatial quantile normalization produces semantic evidence $e^{sem}_{t,i} \in [0,1]$. High semantic responses near the hand seed the manipulated instance. Separation between near-hand and distant attention responses determines semantic reliability and adapts the seed and candidate thresholds. Local propagation through semantic candidates, followed by hand-connected growth, produces the current observation p^{obs}. Visibility is assessed from interaction support relative to its causal history.

Temporal transport and fusion. We normalize and average captured source query features across selected layers, and match each current token to the four most similar tokens in the preceding latent frame. Transported support and its fusion with the observation are

$$\begin{aligned} \widetilde{p}_{t,i} &= c_{t,i} \sum_{j \in \mathcal{N}_4(i)} w_{ij} p^{prev}_{t-1,j}, \\ p^{fuse}_{t,i} &= V_t \max\{p^{obs}_{t,i},\ \omega_t \widetilde{p}_{t,i}(0.2 + 0.8 e^{sem}_{t,i})\}. \end{aligned} \tag{5}$$

Here w_{ij} is a softmax over selected cosine similarities, with temperature adapted to the current frame's match distribution. Token confidence $c_{t,i}$ combines absolute match similarity with its relative strength among current-frame matches. The weight ω_t combines mean confidence on active transported support with semantic reliability. The visibility gate V_t admits observed support or recovery supported by sufficiently strong temporal evidence.

Transport follows latent-frame order. At a block boundary, the preceding support and features come from the previous block's stored final frame; within a block, they are the latest preceding-frame estimates in the source pass. The temporal term is zero when preceding state is unavailable. After fusion, we retain high-scoring support under an adaptive coverage budget, restore a minimum confidence on seed-connected regions, and select hand-anchored connected support S_t. This produces the source-grounded prior p^{src} and fixes the spatial extent available to response refinement.

Source–target branch-response evidence. The first paired prediction of each block supplies

$$f_b = \mathrm{Pool}\Big[\mathrm{QNorm}\Big(\mathrm{Mean}_c \Big|\mathbf{v}^{trg}_{\tau,b} - \mathbf{v}^{src}_{\tau,b}\Big|\Big)\Big]. \tag{6}$$

f_b is a block tensor; $f_{t,i}$ denotes its entry at latent frame t and location i. The channel mean gives a scalar magnitude at each location. QNorm applies per-frame spatial quantile normalization and clips to $[0,1]$; average pooling aligns the result with the support grid. Velocity denotes the update direction along noise time. Its branch difference depends on the current source and target predictions, including their latent states and histories, and is not an estimate of physical motion.

For each frame, let μ^{seed}_t and μ^{ring}_t be the median responses over object seeds and their surrounding ring. With MAD denoting median absolute deviation, set $s_t = 1.4826(\mathrm{MAD}_{seed} + \mathrm{MAD}_{ring})$. Response reliability and attenuation are

$$\begin{aligned} \delta_t &= \frac{[\mu^{seed}_t - \mu^{ring}_t]_+}{s_t + \epsilon}, \qquad & \kappa_t &= \frac{\delta_t}{1 + \delta_t}, \\ L_{t,i} &= \exp\left(-\kappa_t \frac{[\mu^{seed}_t - f_{t,i}]_+}{\max(s_t, \epsilon)}\right), \end{aligned} \tag{7}$$

where $[z]_+ = \max(z, 0)$ and ϵ is the numerical stability floor. Weak seed-to-ring separation makes κ_t small, limiting the contribution of response evidence. With stronger separation, $L_{t,i}$ attenuates support whose response falls below the seed reference.

Let p^{cal} denote the coverage-limited product $p^{src}L$. Local candidates $\mathcal{C}_t$ require semantic support, proximity to prior support or seeds, hand proximity, visibility, and occupancy below a hand-dominated threshold. Their contribution is

$$\begin{aligned} e^{loc}_{t,i} &= \kappa_t L_{t,i}(0.25 + 0.75 e^{sem}_{t,i})\sqrt{d_{t,i}}(1 - h_{t,i})\mathbf{1}[i \in \mathcal{C}_t], \\ \widehat{p}_{t,i} &= \max\left\{p^{cal}_{t,i}, e^{loc}_{t,i}\right\}. \end{aligned} \tag{8}$$

We reapply the coverage limit, restore the confidence floor on seed-connected support, and restrict the result to the previously selected S_t. The final $p \in [0, 1]$ therefore adjusts confidence within the source-grounded extent.

Role normalization and ambiguity. Object support p and hand occupancy h are aligned to the latent control grid before applying Eq. (1). When another control uses a different grid, the four roles are jointly resampled and renormalized. Their normalized role-mixture entropy is

$$u_i = -\frac{1}{\log 4} \sum_{r \in \{obj, int, hand, bg\}} q^r_i \log \max(q^r_i, \epsilon). \tag{9}$$

The source prior defines the initial roles. Response refinement updates them after the first paired prediction, and the final-frame support is stored for the next block.

B.2 Current-Block Control Rules

Source residual and directional filtering. For the source interpolation $\mathbf{x}^{src}_\tau = (1 - \tau)\mathbf{x}^{src}_0 + \tau\boldsymbol{\epsilon}$, the reconstruction velocity is $\mathbf{v}^{src,*} = \boldsymbol{\epsilon} - \mathbf{x}^{src}_0$. Source and target predictions use shared sampled noise, but have their own latent states and causal histories. At each token we form $\boldsymbol{\rho}_i = \mathbf{v}^{src,*}_i - \mathbf{v}^{src}_i$ and $\mathbf{d}_i = \mathbf{v}^{trg}_i - \mathbf{v}^{src}_i$. The filtered residual is

$$\boldsymbol{\rho}^{safe}_i = \boldsymbol{\rho}_i - \frac{\min(\langle \boldsymbol{\rho}_i, \mathbf{d}_i \rangle, 0)}{\|\mathbf{d}_i\|^2_2 + \epsilon}\mathbf{d}_i. \tag{10}$$

Filtering is applied when $\|\mathbf{d}_i\|^2_2 > 10^{-6}$; otherwise the residual is unchanged. The subtraction acts along the latent channel dimension. Its reference is the current branch response, which need not be a pure appearance direction.

Complete velocity allocation. Let B_i be one minus the channel-averaged branch-response magnitude, min–max-normalized over the current block separately for each batch element. This is the backbone preservation gate, distinct from the background responsibility q^{bg}_i. With s_i from Eq. (2), the complete role update is

$$\begin{aligned} \widetilde{\mathbf{v}}^{role}_i &= \mathbf{v}^{trg}_i + s_i\boldsymbol{\rho}^{safe}_i + (q^{hand}_i + q^{bg}_i B_i)\boldsymbol{\rho}_i, \\ \zeta_i &= \lambda_{\text{mix}} q^{bg}_i B_i, \\ \mathbf{v}^{role}_i &= (1 - \zeta_i)\widetilde{\mathbf{v}}^{role}_i + \zeta_i \mathbf{v}^{src,*}_i. \end{aligned} \tag{11}$$

The background gate moderates both residual injection and the direct source reconstruction anchor. The entropy fallback uses the same background anchor:

$$\begin{aligned} \mathbf{v}^{fallback}_i &= (1 - \zeta_i)(\mathbf{v}^{trg}_i + B_i\boldsymbol{\rho}_i) + \zeta_i \mathbf{v}^{src,*}_i, \\ \mathbf{v}^{controlled}_i &= (1 - u_i)\mathbf{v}^{role}_i + u_i \mathbf{v}^{fallback}_i. \end{aligned} \tag{12}$$

Here u_i measures how dispersed the role weights are. A high value attenuates the departure from the backbone-derived preservation fallback; it does not establish that the underlying evidence is incorrect. The scalar s_i describes the filtered-residual coefficient before background anchoring and fallback, rather than the magnitude of the final vector correction.

Spatial query/key coupling. Let $S_i = \mathbf{1}[i \in S_t]$ on the attention grid. For the access score a_i in Eq. (2), define $g_i^{role} = S_i a_i \mathbf{1}[S_i a_i \geq \theta_{\text{role}}]$. The full configuration combines this with the source ownership permission g_i^{can} defined in Appendix B.3, giving $g_i^{acc} = \max\{g_i^{role}, g_i^{can}\}$. This combined gate supplies both spatial appearance access and denoising-time reference reads. At a subsequent prediction, the base blending rate β is spatialized as

$$\begin{aligned} \beta_i &= \beta + (1-\beta) g_i^{acc}, \\ \mathbf{Q}_i &= \beta_i \mathbf{Q}_i^{trg} + (1-\beta_i)\mathbf{Q}_i^{src}, \\ \mathbf{K}_i &= \beta_i \mathbf{K}_i^{trg} + (1-\beta_i)\mathbf{K}_i^{src}. \end{aligned} \tag{13}$$

The rate is configured for the next call using its noise level and the fixed quadratic backbone schedule $\beta = 1 - \tau^2$. The first paired prediction uses scalar base blending, since its response has not yet refined the roles. Spatial modulation changes current-token self-attention queries and keys; values and historical keys retain the backbone's respective paths. It does not modify cross-attention probabilities.

B.3 Appearance Reference Selection and Retrieval

Reference write and read permissions. The full configuration constructs the reference from source ownership, transported and verified through source correspondence. Let $o_i \in [0,1]$ be its continuous confidence, $\mathcal{V}_t$ its verified support, and $H_i \in \{0,1\}$ the persistent-hand exclusion indicator. These quantities are aligned with the attention grid. Role authorization requires sufficient object-associated responsibility relative to background:

$$A_i = \mathbf{1}\left[q_i^{obj} + q_i^{int} \geq \theta_{\text{role}},\ q_i^{obj} + q_i^{int} \geq q_i^{bg}\right]. \tag{14}$$

The write set includes verified owners and a one-token spatial neighborhood, restricted by role authorization and hand evidence:

$$\mathcal{W}_t = \left\{ i \in \mathcal{D}_1(\mathcal{V}_t) : A_i = 1,\ q_i^{hand} \leq \theta_{hand},\ H_i = 0 \right\}. \tag{15}$$

Here $\mathcal{D}_1$ includes the original support and its connected boundary ring. The hand test uses hand-only responsibility q_i^{hand}, while H_i excludes persistent occupancy. This permits role-supported interface details to enter the reference when the ownership and hand conditions are satisfied.

The canonical ownership permission combines the role access score a_i with the source ownership confidence:

$$\begin{aligned} \widetilde{g}_i^{can} &= A_i \max\{a_i, o_i A_i\}(1 - q_i^{hand})(1 - H_i), \\ g_i^{can} &= \widetilde{g}_i^{can} \mathbf{1}[\widetilde{g}_i^{can} \geq \theta_{\text{role}}]. \end{aligned} \tag{16}$$

Source correspondence can therefore sustain read access where the role-only score is attenuated. Read eligibility does not require membership in $\mathcal{W}_t$. The write/read asymmetry is a difference in admission criteria, rather than a guarantee that every frame has more readers than writers. During denoising, the combined gate $g_i^{acc} = \max\{g_i^{role}, g_i^{can}\}$ supplies both appearance access and reference reads. Clean-value commit uses g_i^{can} alone.

At the first completed block with nonempty $\mathcal{W} = \bigcup_t \{(t,i) : i \in \mathcal{W}_t\}$, paired clean passes supply the source keys and target-minus-source values in Eq. (3). We denote the stored value difference by $\Delta\mathbf{V}^* = \mathbf{V}^{trg,*} - \mathbf{V}^{src,*}$. Source addresses are taken before positional rotation. Initialization is deferred when the write set is empty; after initialization the bank is fixed.

Two searches with a shared source query. The reference and current responses are

$$\begin{aligned} \mathbf{D}_i^{ref} &= \mathcal{R}_i(\mathbf{Q}^{src,*}, \mathbf{K}_{mem}^{src,*}, \Delta\mathbf{V}_{mem}^{*}), \\ \mathbf{D}_i^{cur,\tau} &= \mathcal{R}_i(\mathbf{Q}^{src,*}, \mathbf{K}_{cur}^{src,*}, \mathbf{V}_{cur}^{trg,\tau} - \mathbf{V}_{cur}^{src,\tau}). \end{aligned} \tag{17}$$

The reference search uses stored write tokens; the current search uses positions with nonzero g^{acc}. Each search selects top-k keys using cosine similarity of source addresses flattened across heads, rejects matches below θ_{sim}, and aggregates response values using per-head scaled dot-product attention over the retained keys. The shared query is clean-source, while the current value difference comes from the paired predictions at the current noise level. The two searches select their matches independently.

Let m_i^{ref} and m_i^{cur} be their best similarities. When both searches admit a valid match, $c_i^{addr} = \mathrm{clip}_{[0,1]}[(\min(m_i^{ref}, m_i^{cur}) - \theta_{sim})/(1 - \theta_{sim})]$; otherwise $c_i^{addr} = 0$. Thus an unsupported query receives no correction.

Bounding and carrying the response correction. For $\mathbf{e}_i^\tau = \mathbf{D}_i^{ref} - \mathbf{D}_i^{cur,\tau}$, the clipping in Eq. (4) is

$$\mathrm{Clip}_1(\mathbf{e}_i^\tau) = \mathbf{e}_i^\tau \min\left\{1,\ \frac{\max\{\mathrm{RMS}(\mathbf{D}_i^{ref}), \mathrm{RMS}(\mathbf{D}_i^{cur,\tau})\}}{\max\{\mathrm{RMS}(\mathbf{e}_i^\tau), \epsilon\}}\right\}. \tag{18}$$

RMS is computed per token over heads and value channels. The full configuration applies the additive correction after native self-attention and before the output projection at layers 8, 12, 16, and 20. Unit-relative clipping bounds the intervention without implying convergence or a guaranteed appearance error reduction. Attention value differences are distinct from diffusion velocities.

Once an anchor exists, a completed block also undergoes a clean-value correction. After the clean target pass, the same dual retrieval and bounded response correction are applied to the newly written target values, using clean current value differences, g^{can}, and the shared coefficient λ_{mix}. Only admitted queries receive a change to their target values; keys remain on the backbone path. These values then serve as history for subsequent blocks. The backbone's dense causal cache is retained, while the reference bank itself is never overwritten.

Parameter configuration. TRIVE uses independently specified control parameters: $\alpha_o = 0.10$ and $\alpha_c = 0.35$ for object and interface source retention, the shared role threshold $\theta_{\mathrm{role}} = 0.20$, the anchor correction strength $\lambda_{\mathrm{anc}} = 0.20$, and retrieval parameters $\theta_{sim} = 0.35$ and $k = 8$. The interface-access discount, background anchoring, and clean-value commit share the fixed balanced coefficient $\lambda_{\mathrm{mix}} = 0.50$. The hand-majority exclusion uses $\theta_{hand} = 0.50$, and $\epsilon = 10^{-6}$ is used only for numerical stability. Unit gains are absorbed into the equations, while the quadratic blending schedule and controlled layers $\{8, 12, 16, 20\}$ are fixed architectural settings rather than tuned controls. We use this configuration for every video without per-example adjustment.

B.4 Causal Execution Order

For each causal block, the operations occur in the following order:

1. A clean-source probe supplies semantic and correspondence features. The hand representation and preceding support establish the source prior, initial roles, and source ownership permissions.
2. The first paired prediction uses the base scalar Q/K blending schedule. If a reference already exists, its correction uses the available source-grounded read permissions. The prediction's branch response then refines object support and the roles.
3. Refined roles control the velocity correction at this step and configure spatial Q/K blending for subsequent predictions. Reference read and write permissions are updated from the refined state. An existing reference supplies response feedback during later predictions.
4. After denoising, a clean target pass writes the current block's dense causal state. If a reference already exists, its matched correction is applied to authorized clean-target values before that state is used by subsequent blocks.
5. If the reference bank is empty and eligible write tokens exist, the paired clean features initialize it. This new reference is first read in subsequent blocks. Final-frame support and source features are retained for role propagation.

The reference is therefore formed from a completed edit, and each response refinement follows the prediction that supplies its evidence. Causality here concerns blockwise generation and reference use; the hand representation is a provided input.

C Supplementary Experimental Protocol

C.1 Evaluation Set and Prompt Construction

The primary evaluation set contains 200 manipulation videos. Its egocentric subset combines 50 videos from EgoEditBench (Li et al., 2026b) with 50 videos collected by us. Its embodied subset contains 50 robot-manipulation videos from RoboTwin (Mu et al., 2025) and 50 from RoboMIND (Wu et al., 2025a). Every source contains 81 frames and is paired with two compatible target instructions, yielding 400 video–instruction pairs. The instruction set contains 134 color, 133 material, and 133 object-appearance changes. We retain clips in which the manipulated object is visible before or during interaction and the hand or gripper exhibits sustained contact, overlap, or occlusion with it. Instructions alter the manipulated object without changing the recorded action, camera trajectory, or scene layout. Each sequence corresponds to 21 causal latent frames processed as seven blocks of three latent frames; clips containing object occlusion and reappearance are tagged for long-horizon analysis.

Qwen3-VL is used only to draft source descriptions and candidate target instructions. Human annotators verify object identities and attributes, remove ambiguous descriptions, and retain two edits that preserve the source action, geometry, and camera trajectory. Grounding DINO localizes the hand or gripper, and SAM 2 propagates the corresponding masks through the video. The resulting hand tracks are manually screened for quality, and inaccurate or failed tracks are excluded without manual mask refinement. Temporally aligned motion/flow cues are computed from the same source sequence.

C.2 Implementation and Baseline Settings

We use the public Self-Forcing checkpoint at 832×480 resolution on a single NVIDIA A800 GPU. Each causal block contains three latent frames, and standard evaluation uses 15 stochastic sampling steps. Source and target branches share the sampled noise at each step. We apply responsibility-guided controls to attention layers 8, 12, 16, and 20 and use a fixed appearance anchor after its first eligible completed block. The same parameters are used for every video; no model weight is updated and no per-video optimization or inversion is performed.

For each baseline, we use its official checkpoint, implementation, and released inference configuration when available. Every method receives the same source clip, target instruction, and frame range; random seeds are matched whenever the released interface permits. Inputs follow each method's official resolution and are restored to a common evaluation resolution before metric computation. Training-free methods are evaluated without target images or manually edited first frames. Trained editors are reported in a separate group because their supervision and backbones differ from the primary training-free comparison. Runtime is measured on the same NVIDIA A800 GPU, and we record network evaluations, wall-clock time, peak memory, and hardware for every run.

C.3 Metric Definitions

Structure distance. Following FiVE-Bench, structure distance compares the self-similarity of DINO spatial features between the source and edited videos. Lower values indicate that the edit better preserves the source layout and structure.

Background preservation. PSNR, LPIPS, MSE, and SSIM are calculated in editing-irrelevant regions using the benchmark masks. Higher PSNR and SSIM and lower LPIPS and MSE indicate stronger preservation. We report LPIPS multiplied by 10^3, MSE by 10^4, and SSIM by 10^2, following the presentation in StreamEdit.

Text alignment. CLIP_S averages frame-level similarity between the edited video and the target prompt. CLIP_S^{edit} evaluates the same similarity inside the annotated edit region, reducing the contribution of unchanged background content.

Image quality and temporal consistency. NIQE measures no-reference image distortion, with lower values indicating better visual quality. Temporal consistency uses the motion-fidelity protocol

adopted by FiVE-Bench, comparing tracked trajectories in the source and edited videos; higher values indicate stronger preservation of source motion.

Editing accuracy. FiVE-Acc aggregates VLM-based questions that test whether the source concept disappears and the requested target concept appears. We use the official questions, answer parser, and aggregation protocol for every method.

Efficiency. We report wall-clock seconds per output frame, including all model forwards required by the editing method but excluding one-time model loading and metric evaluation. Hardware, precision, and peak memory are recorded together with the runtime.

C.4 Human Study

We conduct a user study with 25 participants on 50 cases sampled from the primary evaluation set. For each case, participants are shown the source video, target instruction, and anonymized outputs in a randomized method order. They rate each output from 1 (best) to 5 (worst) along four criteria: preservation of the source interaction structure, preservation of editing-irrelevant background content, alignment with the target instruction, and overall video quality. We use a balanced assignment in which each participant evaluates 10 cases and each case receives five independent judgments. Table 4 reports the mean and standard deviation of case-level scores.

Table 4: **Human evaluation on 50 cases.** Scores range from 1 (best) to 5 (worst); lower is better. Each case receives five independent judgments. We report mean ± standard deviation over case-level scores.

Method	Structure↓	Background↓	Text Align.↓	Overall↓
TokenFlow	3.48 ± 0.72	2.63 ± 0.68	3.12 ± 0.75	3.18 ± 0.71
Pyramid-Edit	2.74 ± 0.64	3.21 ± 0.74	2.72 ± 0.69	2.88 ± 0.67
Wan-Edit	2.91 ± 0.66	3.18 ± 0.72	2.64 ± 0.67	2.84 ± 0.65
UniEdit-Flow	$\mathbf{1.91 \pm 0.53}$	3.36 ± 0.76	$\mathbf{1.82 \pm 0.49}$	2.28 ± 0.57
StreamDiffusionV2	3.26 ± 0.70	2.36 ± 0.61	3.19 ± 0.73	2.91 ± 0.68
StreamEdit	2.12 ± 0.56	2.05 ± 0.55	3.27 ± 0.76	2.49 ± 0.60
LucyEdit	2.58 ± 0.62	2.51 ± 0.64	2.08 ± 0.54	2.27 ± 0.56
LiveEdit	2.26 ± 0.58	$\mathbf{1.86 \pm 0.50}$	1.94 ± 0.51	$\mathbf{1.90 \pm 0.49}$
Ours	$\mathbf{1.39 \pm 0.40}$	$\mathbf{1.34 \pm 0.38}$	$\mathbf{1.43 \pm 0.42}$	$\mathbf{1.41 \pm 0.39}$

C.5 Evaluation on EgoEditBench

We additionally evaluate on the appearance-editing subset of EgoEditBench (Li et al., 2026b). This evaluation is distinct from the mixed-source evaluation set used in Table 1: it retains the benchmark videos and instructions associated with appearance changes and applies the same subset to every method. We exclude instructions that alter the source action, camera trajectory, or scene structure, and determine the supported task list before running any method. The complete subset manifest will be released.

Following the official EgoEditBench metric setting, we report VLM editing quality (VLM), PickScore (PS), video-level text alignment (TA), and CLIP-based temporal consistency (TC). VLM uses the benchmark evaluation prompts to assess whether the requested edit is correctly realized while preserving relevant source content. PS measures overall perceptual preference, TA measures correspondence to the target instruction, and TC measures consistency across frames. All methods are evaluated from their generated videos using the same metric implementations and aggregation rules.

C.6 Ablation Configurations

All ablations use the same checkpoint, prompts, masks, seeds, and sampling schedule as the full model. *w/o interaction responsibility* retains the aligned hand evidence but removes responsibility-

Table 5: **Evaluation on the appearance-editing subset of EgoEditBench.** We follow the official EgoEditBench metrics: VLM editing quality, PickScore (PS), video-level text alignment (TA), and CLIP-based temporal consistency (TC). All metrics are higher-is-better. The best and second-best results are marked in **red** and **blue**, respectively.

Method	VLM↑	PS↑	TA↑	TC↑
TokenFlow	4.72	18.82	15.54	94.78
Pyramid-Edit	5.18	18.91	15.89	95.06
Wan-Edit	5.46	18.96	16.08	95.24
UniEdit-Flow	5.93	19.03	16.42	95.63
StreamDiffusionV2	2.74	18.58	12.61	94.17
StreamEdit	6.08	19.06	16.31	95.71
LucyEdit	5.21	18.88	15.37	94.69
LiveEdit	**6.47**	**19.18**	**16.73**	**96.08**
Ours	**7.16**	**19.14**	**17.21**	**96.34**

conditioned attention, velocity routing, and reference access. *w/o RAC* restores the native source–target attention coupling while retaining role inference, RGU, and the appearance anchor. *w/o RGU* restores the native StreamEdit velocity update while retaining role inference, RAC, and the anchor. *w/o Appearance Anchor* disables reference construction and retrieval while retaining native causal history and both current-block controls.

C.7 Input-Matched Hand Evidence and Responsibility Diagnostic

TRIVE uses an explicit hand representation in addition to the source video and text conditions. To disentangle the benefit of this auxiliary evidence from the benefit of responsibility-based control, we construct input-matched variants on the full set of 100 egocentric evaluation videos. Starting from the base streaming editor, we first provide the same automatic hand masks used by TRIVE through direct mask-based hand preservation. We then add the same temporally aligned occupancy and motion/flow cues, while still using them without interaction-responsibility decomposition. Finally, TRIVE consumes the identical hand and temporal evidence to infer soft object, interface, hand, and background responsibilities and maps them to operation-specific controls. All variants use identical source videos, target instructions, sampled noise, and denoising settings. The hand tracks are generated automatically using Grounding DINO and SAM 2. They are screened only for input quality, with inaccurate or failed tracks excluded before editing; no mask pixels are manually refined, and the screening procedure is independent of the outputs of any editing method. We evaluate preservation using complementary hand- and interface-focused metrics. Hand-LPIPS measures perceptual change inside independently obtained hand regions, PCK@0.1 evaluates hand-keypoint consistency, and Hand Flow-EPE measures motion discrepancy inside the source hand region. Interface-LPIPS measures perceptual preservation near the hand–object boundary. CLIP_S^{edit} is reported jointly to verify that stronger preservation does not suppress the requested object edit. All evaluator masks, keypoints, and flows are produced independently from the hand representation consumed by the editor.

Table 6: **Input-matched diagnostic of hand evidence and responsibility representation.** All variants are evaluated on the same 100 egocentric videos. The first three variants progressively introduce the auxiliary hand evidence used by TRIVE. Hard Responsibility uses the same downstream controls as TRIVE but assigns each token to its maximum-responsibility role, whereas TRIVE retains soft responsibilities. Lower is better for Hand-LPIPS, Flow-EPE, and Interface-LPIPS; higher is better for PCK@0.1 and CLIP_S^{edit}.

Variant	Hand	Temporal	Role	Hand-LPIPS$\times 10^3$ ↓	PCK@0.1↑	Flow-EPE↓	Interface-LPIPS$\times 10^3$ ↓	CLIP_S^{edit} ↑
Base editor	No	No	–	141.6	75.8	4.79	149.8	24.07
+ Hand mask	Yes	No	–	119.4	82.1	4.18	121.6	23.76
+ Temporal hand representation	Yes	Yes	–	99.1	87.6	2.91	101.8	24.12
+ Hard responsibilities	Yes	Yes	Hard	82.7	90.5	2.46	84.1	24.68
+ Soft responsibilities (TRIVE)	Yes	Yes	Soft	**70.8**	**92.9**	**2.16**	**69.7**	**25.16**

Explicit hand evidence substantially improves hand and interface preservation over the base editor, but direct mask-based protection slightly reduces editing effectiveness, exposing a preservation–editing trade-off. Temporally aligned hand evidence further improves motion consistency. Introducing hard object/interface/hand/background responsibilities provides additional gains, showing that interaction-aware role decomposition is useful beyond hand localization alone.

Importantly, replacing the hard assignment with soft responsibilities further improves both interface preservation and editing effectiveness while keeping all auxiliary evidence and downstream controls unchanged. This comparison supports the use of soft responsibilities for mixed interaction tokens: rather than forcing a token into a single discrete role, TRIVE allows source retention and target-appearance access to coexist with different strengths at the same spatial location.

C.8 Downstream Policy Protocol

Hardware and control. All real-robot experiments use the AgileX Cobot Magic platform, which follows a Mobile-ALOHA-style leader–follower configuration with four 6-DoF arms. Visual observations are captured by three Orbbec Dabai RGB-D cameras: one overhead camera and two wrist-mounted cameras. We use their RGB streams together with robot proprioception as policy observations. We retain the platform's 16-dimensional action interface,

$$\mathbf{a}_t = \left[\mathbf{q}^L_{t,1:6}, g^L_t, \mathbf{q}^R_{t,1:6}, g^R_t, v_t, \omega_t\right] \in \mathbb{R}^{16}, \tag{19}$$

where $\mathbf{q}^L, \mathbf{q}^R$ are the six joint-position commands of the left and right execution arms, g^L, g^R are their gripper commands, and v_t, ω_t are the linear and angular base-velocity commands. The Tracer mobile base remains fixed throughout our tabletop experiments, so we set $v_t = \omega_t = 0$ during both data collection and policy execution. We set the low-level control rate to 80 Hz in all experiments.

Demonstrations and offline augmentation. We collect 20 action-labeled source demonstrations for each of four manipulation skills: Lift, Pick, Capture, and Put, yielding 80 source trajectories in total. Each demonstration contains the visual observations, robot proprioceptive states, and corresponding action sequence. For offline data augmentation, every edited observation sequence remains paired with the actions and proprioceptive states of its source demonstration.

For each source trajectory, StreamEdit and TRIVE each generate two edited visual variants, corresponding to appearance changes that preserve the original task and action sequence. Consequently, both editing-based augmentation conditions contain 20 original and 40 edited observation sequences per task, for a total of 60 training sequences per task. StreamEdit and TRIVE therefore use identical source trajectories and augmentation budgets. We compare policies trained on the original demonstrations, the same demonstrations augmented by StreamEdit, and the same demonstrations augmented by our method. All settings use the same source trajectories, policy architecture, initialization, optimizer, and number of gradient updates.

Evaluation protocol. Each method is tested for 20 trials per task under held-out visual conditions. The task objective, camera configuration, action space, and initial-state protocol are shared across training variants. Depending on the task, the visual shift changes the manipulated object instance or appearance, the workspace appearance, or both. Every physical test condition is absent from both the original source demonstrations and the edited training variations. The editor is used only to prepare training data and is never invoked during real-robot execution.

D Supplementary Related Work

D.1 Streaming Video Generation

Conventional video diffusion models typically denoise a fixed-length clip with bidirectional spatiotemporal attention, requiring future frames to be available and increasing computation with video duration. Streaming generation instead produces frames or short blocks causally under bounded context. Diffusion Forcing assigns different noise levels to sequence tokens and connects full-sequence diffusion with autoregressive prediction (Chen et al., 2024). FramePack compresses variable-length history into a fixed context budget and introduces mechanisms to suppress long-term drift (Zhang

et al., 2025). CausVid converts bidirectional diffusion models into causal generators through asymmetric distribution-matching distillation (Yin et al., 2025); Self Forcing trains on self-generated rollouts to reduce exposure bias (Huang et al., 2025); and StreamDiT combines buffered flow matching with chunk-wise multi-step distillation (Kodaira et al., 2026).

Maintaining quality over open-ended horizons remains difficult because prediction errors accumulate while bounded context discards appearance and motion cues. Rolling Forcing jointly denoises a sliding window and retains global anchors through attention sinks (Liu et al., 2025). LongLive combines short-window attention, frame sinks, and KV recaching for long-form generation and prompt changes (Yang et al., 2025a). Deep Forcing constructs sinks without training and compresses KV cache according to token importance (Yi et al., 2025); Hybrid Forcing augments local attention with recurrent state (Li et al., 2026a); and Diagonal Distillation and Ms. Forcing improve few-step efficiency through temporally aware distillation and noise-dependent spatial granularity (Liu et al., 2026a; Li et al., 2026c).

Most of these systems focus on text- or image-conditioned synthesis rather than editing an observed video. StreamEdit reformulates training-free editing as source-conditioned few-step streaming generation (Jiao et al., 2026b). Egocentric editing adds rapid camera motion, repeated hand–object occlusion, and local entanglement between editable appearance and interaction structure. These properties motivate controls designed specifically for paired causal editing rather than generic long-video generation.

D.2 Video Editing

Text-guided video editing modifies selected semantics while preserving source identity, layout, and motion. Early diffusion methods transferred image-editing priors to video through explicit cross-frame constraints. FateZero reuses spatiotemporal attention maps (Qi et al., 2023), TokenFlow propagates correspondences in diffusion features (Geyer et al., 2024), RAVE jointly edits shuffled frame grids (Kara et al., 2024), and AnyV2V combines an edited first frame with image-to-video generation (Ku et al., 2024). These approaches require little or no task-specific training, but inversion, repeated denoising, and feature injection can limit efficiency and robustness on long videos.

Pretrained video diffusion and rectified-flow models provide stronger native motion priors. VideoGrain controls edits at multiple granularities through attention modulation (Yang et al., 2025b). UniEdit-Flow improves inversion and attention injection for flow models (Jiao et al., 2026a), whereas FlowDirector avoids inversion by steering an ODE trajectory in data space (Li et al., 2025a). FADE factorizes video-model components and applies frequency-aware modulation (Zhu et al., 2025), while ContextFlow enriches attention context for object insertion, replacement, and removal (Chen et al., 2025). These methods nevertheless process predefined clips with iterative, often bidirectional computation.

Training-based editors learn a direct mapping from a source video and an instruction to an edited result. InsV2V uses synthetic paired videos (Cheng et al., 2023), and InsViE-1M scales this direction with a curated million-pair corpus (Wu et al., 2025b). VACE unifies editing, inpainting, and reference-guided creation (Jiang et al., 2025), while UNIC and EditVerse cast heterogeneous generation and editing tasks as in-context token sequences (Ye et al., 2025; Ju et al., 2025). LIVE transfers image-editing priors to video, Kiwi-Edit combines instruction and reference guidance, and TIDE distinguishes heterogeneous conditions with task-specific token embeddings (Wang et al., 2026a; Lin et al., 2026; Liu et al., 2026b). JoyAI-Video-Edit and EditaLive further extend trained editors toward open-ended or live streaming (Xiao et al., 2026; Li et al., 2026d). Our setting instead requires causal, training-free adaptation of a frozen streaming generator.

D.3 Egocentric Video Generation and Editing

Egocentric synthesis has progressed from simulation-based data generation to learned visual world models. EgoGen couples a virtual agent's perception and motion in a closed-loop simulator (Li et al., 2024). GEM predicts RGB and depth under controls for ego-motion, human pose, and scene content (Hassan et al., 2025), while EgoVid-5M and EgoDreamer provide large-scale text and kinematic supervision for action-controllable generation (Wang et al., 2025). Whole-body-conditioned prediction models how 3D pose trajectories affect future observations (Bai et al., 2025), and PlayerOne

transfers exocentric human motion into long-horizon egocentric simulation (Tu et al., 2025). These methods primarily synthesize future visual dynamics from action, pose, or scene conditions rather than preserve the interaction recorded in a source video.

Editing a recorded hand–object interaction requires the requested change to remain compatible with source hand motion, occlusion, and object dynamics. HOI-Swap adapts hand–object configuration in an edited keyframe and propagates it through motion-conditioned generation (Xue et al., 2024). EgoEdit constructs a curated benchmark and trains bidirectional and real-time streaming variants with explicit emphasis on hand preservation and temporal stability (Li et al., 2026b). Our task differs in two respects: the underlying interaction must remain fixed, and no task-specific editing supervision is available. We therefore regulate source and target information according to soft token responsibilities and maintain a fixed source-addressed appearance reference across causal blocks.