Decoupling Action from Egocentric Observation for World Simulation

Yue Ma[1] **Pengjie Song**[1] **Xinyu Wang**[2] **Yi He**[2]
Zeqian Long[3] **Fangneng Zhan**[1] **Kaichen Zhou**[4] **Hongyu Liu**[1]
Hongfa Wang[2] **Peihao Li**[5] **Haoyang Huang**[5] **Nan Duan**[5] **Qifeng Chen**[1,*]
[1]HKUST [2]Tsinghua University [3]Stanford University
[4]MIT [5]JD Explore Academy
*Corresponding Author

Project Page: https://ego-act.github.io/

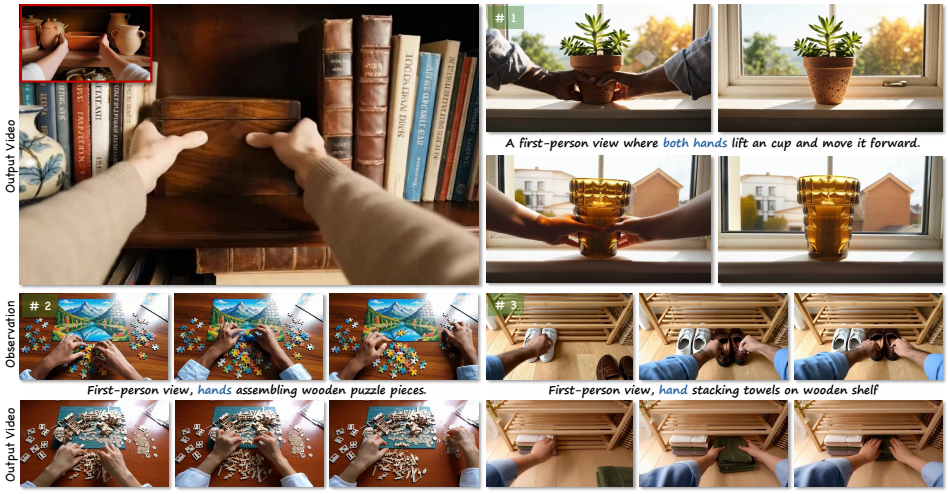

Figure 1: **Showcase of our method.** Given an egocentric manipulation video as observation, EgoACT decouples the transferable action semantics and generates new interaction videos that faithfully reproduce the reference action while maintaining physically plausible hand-object interactions in the target scene.

Abstract

Egocentric action transfer aims to decouple the semantic action from egocentric observations and reproduce them in new visual contexts, enabling scalable world simulation, embodied policy learning, and robotic data generation. Existing approaches to transferring actions rely on condition-guided video generation, which converts the observation video into explicit geometric controls (*e.g.*, hand poses or meshes) to drive synthesis. However, these methods introduce geometric estimation errors, making it difficult to preserve physically plausible interactions when transferred to new scenes. Alternatively, motion transfer methods directly extract motion patterns from the observation video, yet motion-level features alone cannot encode rich contact dynamics, frequently leading to severe hand structural collapse and implausible contact layout. To address both limitations, we present EgoACT, a test-time framework for egocentric action transfer that operates directly on the denoising process of video diffusion models without requiring additional

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

geometric estimators. EgoACT uses Velocity-guided Structure Anchoring to stabilize reference-consistent hand-object structure in the early denoising stage, and Sparse Correspondence Calibration to refine reliable local correspondences in the mid-to-late denoising stages. Together, these two components preserve transferable action semantics while improving temporal coherence and interaction realism. We further establish EgoActionBench, a benchmark for evaluating action preservation, visual quality, and hand-object plausibility across diverse egocentric manipulation scenarios. Experiments show that EgoACT generates more coherent and physically plausible interaction videos than strong baselines.

1 Introduction

Egocentric action transfer aims to decouple the semantic action from egocentric observations and reproduce it in new visual scenes, enabling scalable world simulation, embodied policy learning, and robotic data generation. The core challenge is to decouple *what action is being performed* from *how the current scene happens to look*, so that the interaction logic, including contact transitions, object responses, and coordinated hand-object dynamics , can be faithfully reproduced in novel scenarios.

Existing approaches to transferring egocentric actions are predominantly formulated as condition-guided video generation [19, 29, 67, 73, 80, 83]. These methods first convert an observation video into explicit geometric controls, such as keypoints, hand poses, or mesh sequences [8, 21, 70], and then force the target video to follow these extracted conditions. However, these methods rely on accurate geometric estimation and inevitably introduce external errors into the generation process. When transferred to a new scene, the injected geometric controls cannot guarantee physically valid interactions (Fig. 2). An alternative line of work, motion transfer [35, 43, 45], sidesteps explicit geometric estimation by extracting and replaying motion patterns directly from the observation video. While this avoids estimator-induced errors, motion-level features alone cannot encode the rich contact dynamics inherent in egocentric manipulation. These methods frequently suffer from severe hand structural collapse, and implausible contact layout(Fig. 2).

To address this, instead of hard-copying sparse geometric structures or replaying raw motion from an observation video, we aim to extract the transferable semantic action from an egocentric observation and use it to synthesize a new interaction video under a different visual realization. This setting asks the model to preserve the *interaction logic* behind the motion, including contact timing, object response, and coherent hand-object coordination, rather than the observation appearance or exact geometric condition.

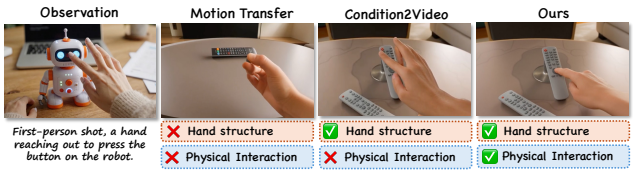

Figure 2: **Motivation.** Motion transfer suffers from hand structural collapse and lacks physical interaction. Condition-guided generation preserves hand shape but produces physically implausible contact. Our method maintains both hand structure and realistic interaction.

In this work, we present EgoACT, a test-time action transfer framework for egocentric world simulation. We observe that egocentric action transfer faces two major failure modes during generation. First, *hand-object structural collapse* often occurs in the early denoising stage, where action guidance can distort hands, objects, and contact geometry before a stable interaction layout is formed. Second, *physically unrealistic interaction* can still appear even when visible motion is roughly aligned, because local correspondences, contact timing, and object responses may drift in later denoising stages. To address these challenges, EgoACT combines two complementary test-time refinements. *Velocity-guided Structure Anchoring* (VSA) stabilizes early denoising by preserving reference-consistent hand-object structure, while *Sparse Correspondence Calibration* (SCC) refines mid-to-late denoising by correcting reliable local correspondences around interaction regions.

Beyond the method, we establish EgoActionBench, a benchmark for evaluating egocentric action transfer. EgoActionBench contains 350 egocentric manipulation videos covering diverse interaction categories, object types, hand configurations, and scene layouts. It evaluates not only visual fidelity, but also action preservation, temporal coherence, and hand-object interaction plausibility.

Experiments on EgoActionBench show that EgoACT produces more coherent transferred actions and more plausible hand-object interactions than strong baselines, supporting the effectiveness of our formulation. Our contributions are summarized as follows:

- We analyze the challenges of egocentric action transfer and propose EgoACT, a test-time action transfer framework that decouples transferable action semantics from observation visual appearance without additional training.
- Technically, we introduce *Velocity-guided Structure Anchoring* and *Sparse Correspondence Calibration*, which respectively address hand-object structural collapse and physically unrealistic interaction.
- We establish EgoActionBench, a benchmark for evaluating egocentric action transfer comprehensively, and conduct extensive experiments that support the effectiveness of our formulation and method.

2 Related Work

Video Motion Transfer. Motion transfer aims to synthesize novel videos by preserving reference motion dynamics, requiring effective decoupling of spatial appearance and temporal motion. Early works [12, 38, 41, 43, 47, 48, 56, 57, 67, 73, 74, 82, 83] rely on explicit control signals, which demand large-scale annotations and substantial computation. Consequently, recent diffusion models shift toward implicit control, employing either parameter-efficient tuning [23, 50, 68, 84] or inference-time motion extraction [11, 20, 35, 39, 42, 46, 64, 71, 75, 76, 77]. However, these frameworks primarily target 3D U-Nets [3]. As models evolve toward Diffusion Transformers (DiTs) [44] with fully coupled 3D attention, traditional methods relying on explicit spatio-temporal separation become difficult to apply. Recent DiT-based adaptations explore attention flow optimization [45], parameter-efficient tuning [1, 40], temporal kernel encoding [52], and streamlined extraction [43]. However, these methods still focus primarily on visible motion transfer, while egocentric action transfer requires preserving contact transitions and hand-object dynamics.

Interaction-Centric Video Generation. Recent advances in diffusion have increasingly pushed generative video models toward the paradigm of video world modeling [4, 5, 18], where the goal is to capture not only visual appearance but also the underlying dynamics and interactions of the physical world. Some works [27, 37, 65, 81] address interactive manipulation video generation by introducing structured interaction signals to improve controllability and physical plausibility. Methods such as Mask2IV [27] and VHOI [81] leverage mask trajectories or keypoints to guide fine-grained interactions. Other approaches incorporate more explicit control cues or geometric priors, such as InterDyn [2], GenHSI [31], and ZeroHSI [28]. EgoSim [13] and Hand2World [66] further introduce egocentric conditioning or structured world representations to enhance interaction realism and consistency. Despite these advances, existing methods largely rely on external condition estimators [8, 21, 70] to guide the generation of physically consistent interaction videos, which limit practical application. In contrast, we propose the EgoACT that enhances interaction consistency without requiring external estimators, enabling more robust interaction world modeling.

3 Method

Given an egocentric manipulation video x^{ref} that demonstrates a source action, egocentric action transfer aims to synthesize a new video x^{gen} that preserves the transferable action semantics while allowing the visual realization to change. The transferred signal should not be limited to visible hand displacement, but should also retain the underlying interaction logic, including contact transitions, object response, and coordinated hand-object dynamics. In this section, we first discuss the motivation and key challenges of egocentric action transfer in Sec. 3.1. We then introduce Velocity-guided Structure Anchoring (VSA) in Sec. 3.2, which stabilizes the global hand-object structure during the early denoising stage. Finally, we present Sparse Correspondence Calibration (SCC) in Sec. 3.3, which refines local temporal correspondences in the mid-to-late denoising stages.

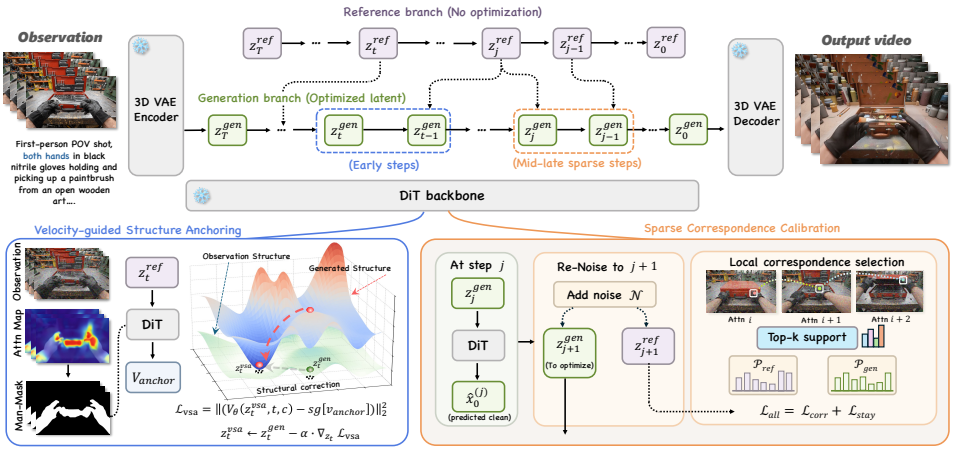

Figure 3: **Overview of the proposed framework.** Given an egocentric manipulation video, our framework uses a reference branch and an optimized generation branch for test-time action transfer. We apply Velocity-guided Structure Anchoring (VSA) in the early denoising steps to preserve hand-object structure, and Sparse Correspondence Calibration (SCC) in the mid-to-late steps to refine local correspondences and improve interaction realism.

3.1 Motivation

Egocentric action transfer requires the model to transfer the semantic action from a reference egocentric observation while synthesizing a new interaction video under a different visual realization. Consistent with the discussion in Sec. 1, this setting is difficult because the model must preserve the interaction logic behind the motion rather than simply reproduce visible hand displacement. We focus on two failure modes that commonly appear during generation:

Hand-object Structural Collapse. Explicit condition guidance usually converts the reference video into geometric controls, such as keypoints, hand poses, or mesh sequences, and then forces the generation process to follow these controls. Although this can preserve coarse motion, it is fragile in egocentric manipulation, where hands are frequently occluded, distorted by perspective, and tightly coupled with objects. Noisy or distribution-mismatched controls can over-constrain the early denoising trajectory, leading to broken fingers, distorted hands, collapsed objects, or implausible contact geometry.

Physically Unrealistic Interaction. Even when visible motion is roughly aligned, the generated video may still fail to preserve the physical logic of manipulation, such as when contact is established, how the object responds, and whether local hand-object correspondences remain temporally consistent. This problem becomes especially apparent in the mid-to-late denoising stages, where fine-grained appearance and contact details are formed.

To address these two failure modes, EgoACT performs two complementary latent optimization steps during inference. Velocity-guided Structure Anchoring (VSA) is applied in the early denoising stage to keep the generation close to reference-consistent hand-object structure, reducing structural collapse before the interaction layout is formed. Sparse Correspondence Calibration (SCC) is then applied in the mid-to-late denoising stages to correct reliable local correspondences around interaction regions, improving temporal consistency and hand-object realism without additional training.

Both modules use a frozen reference branch and an optimized generation branch. At each guided denoising step t, the clean reference latent z_0^{ref} is mapped to the same noise level by the forward process, $z_t^{\text{ref}} = (1-\sigma_t)z_0^{\text{ref}} + \sigma_t\epsilon$, and is used only to extract stop-gradient guidance targets. We update only the generation latent z_t^{opt}. The reference branch is evaluated with a null-text prompt, while the generation branch uses the target prompt.

3.2 Velocity-guided Structure Anchoring

To prevent the hand-object structural layout from being severely derailed during the initial motion injection, we propose **V**elocity-guided **S**tructure **A**nchoring (VSA) at the early guidance steps. In egocentric manipulation, the source observation does not merely provide a motion template; it also encodes implicit physical knowledge about feasible hand-object contacts, relative object motion, and interaction timing. Directly injecting motion cues without preserving this knowledge can produce trajectories that follow the reference displacement but violate the physical structure of the manipulation process. VSA therefore uses the reference branch as a structure-and-physics anchor at the same noise level, aligning the generation trajectory with both the kinematic layout and the interaction dynamics implied by the observation video.

Specifically, instead of relying purely on spatial loss, we extract the velocity prediction from the reference branch at the same noise level $v_{\text{anchor}} = \text{sg}\big[v(z_t^{\text{ref}})\big]$, where $V_\theta(\cdot)$ denotes the velocity prediction of the diffusion model parameterized by θ at noise level t and $\text{sg}[\cdot]$ denotes the stop-gradient operation. Given the velocity prediction $v_{\text{opt}} = v(z_t^{\text{opt}})$ of the currently optimized generation latent, VSA acts as a spatially weighted velocity regression.

The rationale behind using velocity as the structural carrier is that, in flow-based diffusion sampling, the velocity prediction defines the local tangent direction of the denoising trajectory on the latent video manifold. For egocentric manipulation, this tangent direction implicitly captures how hand poses, object states, and their spatial relations should evolve from the current noisy state. Therefore, aligning v_{opt} with v_{anchor} does not directly copy the reference latent; instead, it steers the current generation to follow a reference-consistent local flow. This preserves the freedom of the generated content while encouraging its structural features to move toward the physically plausible manifold indicated by z_t^{ref}.

To avoid applying the velocity anchor uniformly over the entire frame, we introduce an Attention Motion Flow (AMF) [45] induced soft structural mask $\mathcal{M}(\mathbf{x})$. The intuition is that regions with stronger reference motion, such as moving hands and manipulated objects, should receive stronger structural anchoring, while static backgrounds should still retain a weak constraint to preserve global coherence. Let $\mathcal{A}_{\text{ref}}(\mathbf{x})$ denote the reference AMF at spatial location $\mathbf{x}$, and define the structural saliency score as its magnitude:

$$\mathcal{S}(\mathbf{x}) = \|\mathcal{A}_{\text{ref}}(\mathbf{x})\|_2. \tag{1}$$

We normalize $\mathcal{S}(\mathbf{x})$ into $[0, 1]$ denoted as $\hat{\mathcal{S}}(\mathbf{x})$ and apply a power transform controlled by γ to adjust the concentration of the mask. With a minimum weight bound $m_{\min}$, the final soft mask is formulated as:

$$\mathcal{M}(\mathbf{x}) = \hat{\mathcal{S}}(\mathbf{x})^\gamma (1 - m_{\min}) + m_{\min}. \tag{2}$$

The resulting $\mathcal{M}(\mathbf{x})$ is defined only over spatial locations and is shared across all frames during VSA optimization. Using soft masks and velocity anchors provides structural guidance. The final VSA objective is a mask-weighted velocity regression over N spatial-temporal latent positions:

$$\mathcal{L}_{\text{VSA}} = \frac{1}{N} \sum_{\mathbf{x}} \mathcal{M}(\mathbf{x}) \left\| v_{\text{opt}}(\mathbf{x}) - v_{\text{anchor}}(\mathbf{x}) \right\|_2^2. \tag{3}$$

During early sampling, by gradient descent $z_t^{\text{opt}} \leftarrow z_t^{\text{opt}} - \alpha \nabla_{z_t^{\text{opt}}} \mathcal{L}_{\text{VSA}}$, the latent is directly updated, effectively stabilizing the global structural layout.

3.3 Sparse Correspondence Calibration

While VSA stabilizes the macroscopic structure in the early stage of denoising, cross-frame correspondences may still drift during the following denoising process. Existing global optimization strategies, such as test-time optimization [25, 30, 69], usually regularize the noisy latent with dense reconstruction, feature, or motion-level objectives. In egocentric manipulation, however, the visual content and action semantics are tightly coupled: hands, objects, and contact states jointly determine whether a trajectory remains valid. Global objectives tend to dilute these fine-grained vision-action constraints into a single optimization target, which may over-correct the latent and unintentionally alter object appearance, hand-object relations, or the intended action.

To address this, we introduce **S**parse **C**orrespondence **C**alibration (SCC), which operates within a valid diffusion path and aligns the local correspondence distributions between the reference and

generated branches at equivalent noise levels. By optimizing only the reliable sparse matching support induced by the reference branch, SCC provides a more targeted correction signal for local interaction-level correspondence drift while avoiding unnecessary updates on ambiguous or irrelevant regions.

Egocentric manipulation videos exhibit locally continuous motion between adjacent frames: a hand or object token is unlikely to jump to an arbitrary global position within one denoising step. Based on this observation, SCC searches for correspondences only within a local displacement window $\mathcal{N}(u) = \{u + \Delta \mid \Delta \in [-r, r]^2\}$ around each query token u. Specifically, for each adjacent frame pair $(i, i+1)$, SCC uses the query token from the current frame and the key token from the next frame, i.e., $q_i(u)$ and $k_{i+1}(u + \Delta)$, to measure cross-frame correspondence. In the dual branch, we score each candidate displacement by the scaled query-key similarity:

$$s_i(u, \Delta) = \frac{\langle q_i(u), k_{i+1}(u + \Delta) \rangle}{\sqrt{d}}, \quad u + \Delta \in \mathcal{N}(u). \tag{4}$$

We then keep the top-k displacements with the largest reference scores as $\Omega_k^i(u)$. This reference-induced support specifies where reliable local matches are expected to occur. Both the reference and generated correspondence distributions are normalized over the same support:

$$\mathcal{P}_i(\Delta \mid u) = \mathrm{softmax}\left(\frac{s_i(u, \Delta)}{\tau}\right), \quad \Delta \in \Omega_k^i(u). \tag{5}$$

Recognizing that not all reference tokens provide reliable matching cues, we assign a confidence weight $w_{\mathrm{conf}}^i(u) = 1 - \frac{H_i(u)}{\log k}$ based on the entropy $H_i(u)$ of the reference distribution, and normalize it as $\bar{w}_i(u) = w_{\mathrm{conf}}^i(u) / \sum_{u'} w_{\mathrm{conf}}^i(u')$. The correspondence alignment is then supervised by a confidence-weighted cross-entropy loss:

$$\mathcal{L}_{\mathrm{corr}} = -\sum_i \sum_u \bar{w}_i(u) \sum_{\Delta \in \Omega_k^i(u)} \mathcal{P}_{\mathrm{ref},i}(\Delta \mid u) \log \mathcal{P}_{\mathrm{gen},i}(\Delta \mid u). \tag{6}$$

This objective encourages the generated branch to place high probability on the same local displacements preferred by the reference branch. The confidence weight further makes sharp and reliable reference correspondences dominate the calibration, while suppressing ambiguous tokens with high-entropy matching distributions. In this way, SCC optimizes the cross-frame matching structure rather than directly pulling the latent toward a dense visual target, which helps correct local correspondence drift while preserving the generated content.

To restrict the correction magnitude and prevent the latent from deviating too far from the original sampling path, a stay regularizer $L_{\mathrm{stay}} = \|z_{j+1}^{\mathrm{opt}} - \mathrm{sg}[z_{j+1}]\|_2^2$ is added. The total objective is:

$$\mathcal{L}_{\mathrm{SCC}} = \lambda_{\mathrm{c}} \mathcal{L}_{\mathrm{corr}} + \lambda_{\mathrm{s}} \mathcal{L}_{\mathrm{stay}}. \tag{7}$$

By minimizing L_{SCC}, the model efficiently rectifies shape drift in interactive regions while regularizing the latent to stay close to the original sampling path.

4 Experiments

4.1 Implementation Details

In our experiments, we implement EgoACT on top of a pre-trained video diffusion transformer [62] and keep all model parameters frozen during inference. Unless otherwise specified, videos are sampled with 50 denoising steps at a resolution of 832×480 and 81 frames. During inference, we apply the sliding-window AMF guidance in the first 20% of the outer denoising process to stabilize the hand-object layout, and use SCC in the mid-to-late denoising stage from step 35 to 45 to calibrate local correspondences. Each guided outer step performs a 10-step latent-only optimization with AdamW, where the learning rate is linearly decayed from 0.003 to 0.002. More implementation details are provided in the supplementary material.

EgoActionBench. We establish EgoActionBench to evaluate egocentric action transfer in action-centric world simulation. EgoActionBench is built from online egocentric manipulation videos, public manipulation datasets [9, 10, 17], and generated videos from powerful video diffusion models [15,

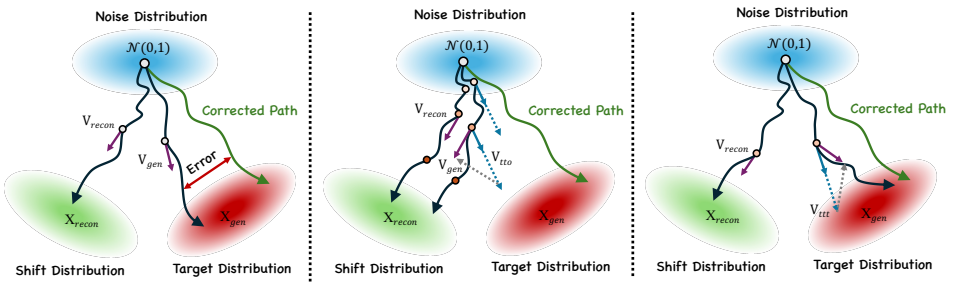

Figure 4: **Illustration of the proposed SCC.** (a) Direct transfer injects the reference action without correction, which may lead to errors in weak hand-object interactions. (b) Applying correction throughout the whole denoising process over-constrains early generation, making it difficult to decouple transferable action from source observation. (c) Our SCC performs sparse correspondence calibration only in the mid-to-late denoising stages, where the action layout has been established and local interaction details can be refined, improving temporal consistency and hand-object realism.



Figure 5: **Gallery of proposed method.** EgoACT can transfer a wide range of egocentric manipulation videos, covering diverse interaction categories, object types, and motion patterns. The generated videos preserve plausible hand-object structure, contact dynamics, and temporal coherence while adapting the reference action to new visual contexts.

26, 62]. It contains 350 high-quality observation videos and 1,400 evaluation prompts, covering 287 indoor and 63 outdoor scenes, with 164 single-hand and 186 two-hand interactions. The actions include grasping, opening, pouring, placing, tool use, and handover across diverse objects and interaction patterns. To ensure the visual quality of benchmark, we filter out clips with excessive ego-camera motion, severe blur, heavy occlusion, objects leaving the frame, unclear hand-object contact, incomplete actions, or corrupted frames. For each retained video, GPT-5 generates an action-centric source caption, from which we derive 4 target prompts by changing the scene, object appearance, or background while preserving the original action semantics. All prompts are manually checked to avoid reference-specific appearance leakage. Each sample is annotated with scene type, hand type, action category, and object category.

4.2 Comparison with Baselines

Qualitative comparison. We compare EgoACT with representative egocentric manipulation and interaction generation methods, including VACE [24], Mask2IV [27], CosHand [58], InterDyn [2], and DiTFlow [45]. As shown in Fig. 7, previous methods exhibit different failure patterns in egocentric manipulation transfer. Mask2IV and CosHand better preserve hand shape but struggle with contact consistency, often generating implausible object responses such as objects remaining static despite clear grasping motion. InterDyn and DiTFlow introduce visible structural artifacts when hands and objects are tightly coupled, leading to floating contacts or broken spatial relations. In contrast, EgoACT preserves a more stable hand-object layout while maintaining temporally coherent contact

Table 1: **Comparison with state-of-the-art methods**. **Red** and **Blue** denote the best and second best results, respectively.

Method	Quantitative Metrics				Vbench Metrics			
	Text Sim.↑	Motion Fid.↑	Temp. Cons.↑	App. Cons.↑	Sub. Cons.↑	Back. Cons.↑	Aes. Qual.↑	Motion Smooth.↑
InterDyn [2]	0.2645	**0.2434**	0.9673	**0.9265**	0.9105	0.9069	0.4273	0.9865
Mask2IV [27]	0.2604	0.1910	0.9749	0.8565	0.8929	0.9495	0.4673	0.9820
CosHand [58]	0.2602	0.2227	0.9277	0.9133	0.8381	0.9201	0.4226	0.8784
DiTFlow [45]	0.2579	0.2167	**0.9750**	0.8839	0.8856	0.9325	**0.5065**	0.9868
VACE [24]	**0.2624**	0.2389	0.9571	0.9149	**0.9439**	**0.9507**	0.4647	**0.9881**
Ours	**0.2746**	**0.2750**	**0.9866**	**0.9617**	**0.9732**	**0.9719**	**0.5633**	**0.9898**

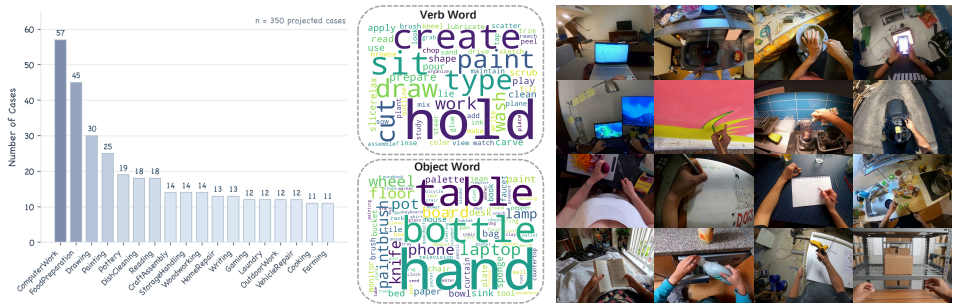

Figure 6: **The overview of EgoActionBench.** EgoActionBench contains 350 reference manipulation videos and 1,400 target prompts, covering indoor/outdoor scenes, single-/two-hand interactions, and diverse manipulation categories. The benchmark evaluates whether generated videos preserve the reference action semantics while adapting to new visual contexts.

and physically plausible object response throughout the interaction sequence. These qualitative results confirm that VSA suppresses early structural collapse by anchoring the hand-object skeleton, while SCC improves local correspondence consistency in the later denoising stages to produce more realistic contact dynamics.

Quantitative comparison. We quantitatively compare EgoACT with VACE [24], Mask2IV [27], CosHand [58], InterDyn [2], and DiTFlow [45] on EgoActionBench. For fair comparison, all methods are evaluated on the same set of reference manipulation videos and target prompts. Following prior works, we adopt four task-level metrics: (a) *Text Similarity*, the average CLIP similarity between the target prompt and generated frames; (b) *Motion Fidelity*, the tracklet similarity between reference and generated videos; (c) *Temporal Consistency*, the average feature similarity between consecutive frames; and (d) *Appearance Consistency*, measuring how well the generated video preserves the target visual appearance.

User Study. To account for the limitations of automatic metrics in capturing real-world preferences, we conduct a user study with 20 volunteers on 50 randomly selected test cases. The volunteers categorize each result into one of four levels: excellent, good, pass, or fail, based on appearance consistency, text alignment, temporal consistency, and motion fidelity. The final data is presented in Fig. 8. Our method outperforms others in both automated metrics and human subjective preferences.

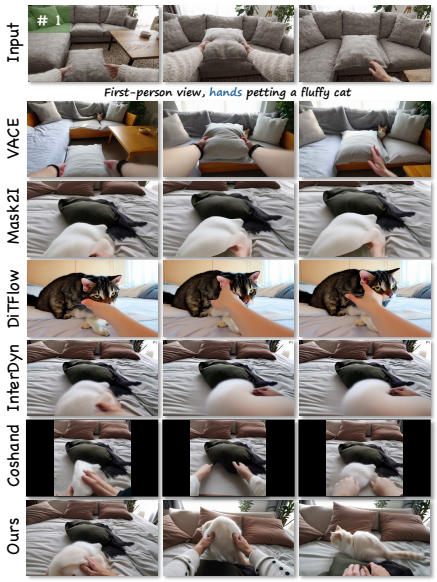

Figure 7: **Qualitative comparison.** Compare with previous methods, our method preserves stable hand-object layout and object response.

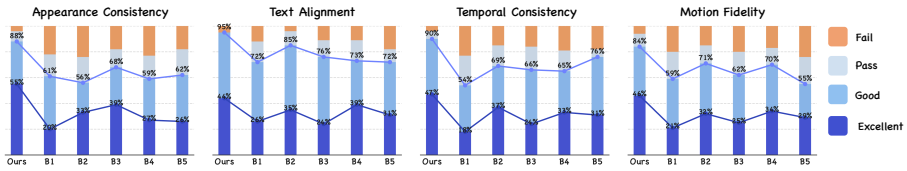

Figure 8: **User Study.** We compare our method against state-of-the-art baselines. Our method is consistently preferred by human evaluators, achieving the highest excellent/good ratio across all evaluated aspects.

VBench Metrics. We further adopt four VBench metrics for a broader assessment of video quality. *Subject Consistency* evaluates whether the main subject identity is preserved across frames. *Background Consistency* measures the coherence of static scene regions. *Aesthetic Quality* scores the visual appeal of generated videos, and *Motion Smoothness* evaluates inter-frame motion continuity. As shown in Table 1, EgoACT achieves strong performance across all metrics, which further validates the effectiveness of our method in producing high-quality, coherent, and semantically aligned egocentric manipulation videos.

4.3 Ablation Study

We ablate the two core components of EgoACT under the same experimental setup as the main comparison to validate their individual contributions.

Effectiveness of VSA. As shown in Table 2 and Fig. 9, removing VSA leads to notable degradation in hand generation quality. Without velocity-guided anchoring in early denoising, the generation trajectory lacks structural constraints from the reference branch, causing frequent hand distortion, broken fingers, and implausible contact geometry. With VSA enabled, the reference velocity field steers the latent toward a structurally consistent manifold, preventing structural collapse before the interaction skeleton is established.

Table 2: **Quantitative ablation.** **Red**: best, **Blue**: second best.

Method	Text Sim.↑	Motion Fid.↑	Temp. Cons.↑	App. Cons.↑
W/o VSA	**0.2689**	**0.2442**	0.9683	0.9312
W/o SCC	0.2641	0.2398	**0.9714**	**0.9487**
Ours	**0.2746**	**0.2750**	**0.9866**	**0.9617**

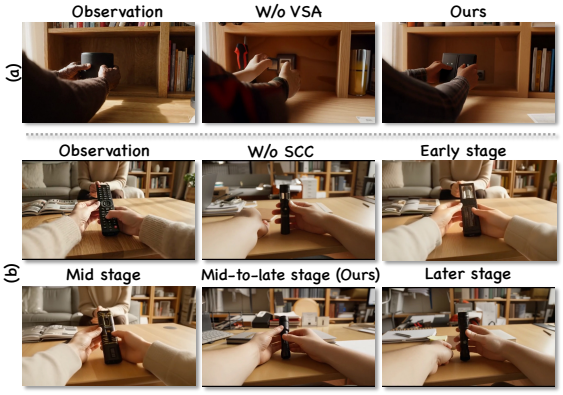

Figure 9: **Visual ablation.** VSA suppresses structural collapse; SCC improves local correspondence.

Effectiveness of SCC. Disabling SCC preserves coarse hand motion but degrades fine-grained interaction quality. Without correspondence calibration, local hand-object correspondences drift across frames, resulting in contact jitter and delayed object response (Table 2, Fig. 9). We apply SCC in the mid-to-late denoising stages (steps 35–45): early-stage correction over-constrains the latent and disrupts action-appearance decoupling, while late-stage intervention has limited effect as fine details have solidified. This window allows SCC to refine correspondences after the global layout is stabilized by VSA, yielding more plausible interactions.

5 Conclusion

We present EgoACT, a test-time action transfer framework for egocentric world simulation that decouples transferable action semantics from source visual appearance without additional training. EgoACT addresses two critical failure modes in egocentric manipulation transfer: hand-object structural collapse in early denoising and physically unrealistic interaction in later stages. Velocity-guided

Structure Anchoring (VSA) stabilizes the global hand-object layout by aligning the generation trajectory with reference-consistent velocity fields during early denoising, while Sparse Correspondence Calibration (SCC) refines local correspondences around interaction regions in the mid-to-late stages to improve contact consistency and object response plausibility. We also introduce EgoActionBench, a benchmark with 350 egocentric manipulation videos and 1,400 evaluation prompts spanning diverse scenes, hand configurations, and interaction categories. Extensive experiments demonstrate that EgoACT outperforms existing methods in both quantitative metrics and human evaluation, producing more coherent action transfer with plausible hand-object interactions.

References

[1] Rameen Abdal, Or Patashnik, Ivan Skorokhodov, Willi Menapace, Aliaksandr Siarohin, Sergey Tulyakov, Daniel Cohen-Or, and Kfir Aberman. Dynamic concepts personalization from single videos. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–9, 2025.

[2] Rick Akkerman, Haiwen Feng, Michael J Black, Dimitrios Tzionas, and Victoria Fernández Abrevaya. Interdyn: Controllable interactive dynamics with video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12467–12479, 2025.

[3] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22563–22575, 2023.

[4] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Leo Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1(8):1, 2024.

[5] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024.

[6] Luca Eyring, Shyamgopal Karthik, Alexey Dosovitskiy, Nataniel Ruiz, and Zeynep Akata. Noise hypernetworks: Amortizing test-time compute in diffusion models. *arXiv preprint arXiv:2508.09968*, 2025.

[7] Yossi Gandelsman, Yu Sun, Xinlei Chen, and Alexei Efros. Test-time training with masked autoencoders. *Advances in Neural Information Processing Systems*, 35:29374–29385, 2022.

[8] Liuhao Ge, Yujun Cai, Junwu Weng, and Junsong Yuan. Hand pointnet: 3d hand pose estimation using point sets. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8417–8426, 2018.

[9] GenRobot. Gen-egodata. https://huggingface.co/datasets/genrobot2025/Gen-EgoData, 2025. Accessed: 2026-04-28.

[10] GenRobot. 10kh-realomni-opendata. https://huggingface.co/datasets/genrobot2025/10Kh-RealOmin-OpenData, 2025. Accessed: 2026-04-28.

[11] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023.

[12] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

[13] Jinkun Hao, Mingda Jia, Ruiyan Wang, Xihui Liu, Ran Yi, Lizhuang Ma, Jiangmiao Pang, and Xudong Xu. Egosim: Egocentric world simulator for embodied interaction generation. *arXiv preprint arXiv:2604.01001*, 2026.

[14] Haoran He, Jiajun Liang, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, and Ling Pan. Scaling image and video generation via test-time evolutionary search. *arXiv preprint arXiv:2505.17618*, 2025.

[15] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

[16] Yining Hong, Beide Liu, Maxine Wu, Yuanhao Zhai, Kai-Wei Chang, Linjie Li, Kevin Lin, Chung-Ching Lin, Jianfeng Wang, Zhengyuan Yang, et al. Slowfast-vgen: Slow-fast learning for action-driven long video generation. *arXiv preprint arXiv:2410.23277*, 2024.

[17] Ryan Hoque, Peide Huang, David J Yoon, Mouli Sivapurapu, and Jian Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video. *arXiv preprint arXiv:2505.11709*, 2025.

[18] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

[19] Bin Hu, Yanwen Ma, Jiehui Huang, Ziliang Zhang, Haoning Wu, Ruicheng Zhang, Yaokun Li, Zijun Wang, Yuechen Zhang, Chun-Mei Tseng, et al. Physeditworld: A large-scale dataset toward physics-editable world models. *arXiv preprint arXiv:2606.26694*, 2026.

[20] Teng Hu, Jiangning Zhang, Ran Yi, Yating Wang, Hongrui Huang, Jieyu Weng, Yabiao Wang, and Lizhuang Ma. Motionmaster: Training-free camera motion transfer for video generation. *arXiv preprint arXiv:2404.15789*, 2024.

[21] Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. Depthcrafter: Generating consistent long depth sequences for open-world videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2005–2015, 2025.

[22] Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, et al. Vipe: Video pose engine for 3d geometric perception. *arXiv preprint arXiv:2508.10934*, 2025.

[23] Hyeonho Jeong, Jinho Chang, Geon Yeong Park, and Jong Chul Ye. Dreammotion: Space-time self-similar score distillation for zero-shot video editing. In *European Conference on Computer Vision*, pages 358–376. Springer, 2024.

[24] 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*, pages 17191–17202, 2025.

[25] Sunwoo Kim, Minkyu Kim, and Dongmin Park. Test-time alignment of diffusion models without reward over-optimization. *arXiv preprint arXiv:2501.05803*, 2025.

[26] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

[27] Gen Li, Bo Zhao, Jianfei Yang, and Laura Sevilla-Lara. Mask2iv: Interaction-centric video generation via mask trajectories. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 6091–6099, 2026.

[28] Hongjie Li, Hong-Xing Yu, Jiaman Li, and Jiajun Wu. Zerohsi: Zero-shot 4d human-scene interaction by video generation. *arXiv preprint arXiv:2412.18600*, 2024.

[29] Jiahua Li, Zixu John, Tom Zhong, Fuping Wu, Tianhao Xu, Ruicheng Zhang, Jun Zhou, and Zunnan Xu. Fight world model: Customizable combat through visual interaction proxies. *arXiv preprint arXiv:2610.40607*, 2026.

[30] Yafu Li, Xuyang Hu, Xiaoye Qu, Linjie Li, and Yu Cheng. Test-time preference optimization: On-the-fly alignment via iterative textual feedback. *arXiv preprint arXiv:2501.12895*, 2025.

[31] Zekun Li, Rui Zhou, Rahul Sajnani, Xiaoyan Cong, Daniel Ritchie, and Srinath Sridhar. Genhsi: Controllable generation of human-scene interaction videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 138–149, 2026.

[32] Long Lian, Baifeng Shi, Adam Yala, Trevor Darrell, and Boyi Li. Llm-grounded video diffusion models. *arXiv preprint arXiv:2309.17444*, 2023.

[33] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International conference on machine learning*, pages 6028–6039. PMLR, 2020.

[34] Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision*, 133(1):31–64, 2025.

[35] Pengyang Ling, Jiazi Bu, Pan Zhang, Xiaoyi Dong, Yuhang Zang, Tong Wu, Huaian Chen, Jiaqi Wang, and Yi Jin. Motionclone: Training-free motion cloning for controllable video generation. *arXiv preprint arXiv:2406.05338*, 2024.

[36] Fangfu Liu, Hanyang Wang, Yimo Cai, Kaiyan Zhang, Xiaohang Zhan, and Yueqi Duan. Video-t1: Test-time scaling for video generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 18671–18681, 2025.

[37] Kun Liu, Qi Liu, Xinchen Liu, Jie Li, Yongdong Zhang, Jiebo Luo, Xiaodong He, and Wu Liu. Hoigen-1m: A large-scale dataset for human-object interaction video generation. *arXiv preprint arXiv:2503.23715*, 2025.

[38] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4117–4125, 2024.

[39] Yue Ma, Kunyu Feng, Zhongyuan Hu, Xinyu Wang, Yucheng Wang, Mingzhe Zheng, Xuanhua He, Chenyang Zhu, Hongyu Liu, Yingqing He, et al. Controllable video generation: A survey. *arXiv preprint arXiv:2507.16869*, 2025.

[40] Yue Ma, Yulong Liu, Qiyuan Zhu, Ayden Yang, Kunyu Feng, Xinhua Zhang, Zhifeng Li, Sirui Han, Chenyang Qi, and Qifeng Chen. Follow-your-motion: Video motion transfer via efficient spatial-temporal decoupled finetuning. *arXiv preprint arXiv:2506.05207*, 2025.

[41] Yue Ma, Yulong Liu, Qiyuan Zhu, Ayden Yang, Kunyu Feng, Xinhua Zhang, Zhifeng Li, Sirui Han, Chenyang Qi, and Qifeng Chen. Follow-your-motion: Video motion transfer via efficient spatial-temporal decoupled finetuning. *arXiv preprint arXiv:2506.05207*, 2025.

[42] Yue Ma, Jiangming Wang, Yucheng Wang, Xilai Wang, Zhiyuan Li, Xinyu Wang, Hongyu Liu, Ruofan Liang, Songchun Zhang, Yuxuan Xue, et al. Livelight: Real-time streaming video relighting with interactive control. *arXiv preprint arXiv:2608.01771*, 2026.

[43] Yue Ma, Zhikai Wang, Tianhao Ren, Mingzhe Zheng, Hongyu Liu, Jiayi Guo, Mark Fong, Yuxuan Xue, Zixiang Zhao, Konrad Schindler, et al. Fastvmt: Eliminating redundancy in video motion transfer. *arXiv preprint arXiv:2602.05551*, 2026.

[44] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

[45] Alexander Pondaven, Aliaksandr Siarohin, Sergey Tulyakov, Philip Torr, and Fabio Pizzati. Video motion transfer with diffusion transformers. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22911–22921, 2025.

[46] 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*, pages 15932–15942, 2023.

[47] Xiangfei Qiu, Jilin Hu, Lekui Zhou, Xingjian Wu, Junyang Du, Buang Zhang, Chenjuan Guo, Aoying Zhou, Christian S. Jensen, Zhenli Sheng, and Bin Yang. TFB: Towards comprehensive and fair benchmarking of time series forecasting methods. In *Proc. VLDB Endow.*, pages 2363–2377, 2024.

[48] Xiangfei Qiu, Xingjian Wu, Yan Lin, Chenjuan Guo, Jilin Hu, and Bin Yang. DUET: Dual clustering enhanced multivariate time series forecasting. In *SIGKDD*, pages 1185–1196, 2025.

[49] Vignav Ramesh and Morteza Mardani. Test-time scaling of diffusion models via noise trajectory search. *arXiv preprint arXiv:2506.03164*, 2025.

[50] Yixuan Ren, Yang Zhou, Jimei Yang, Jing Shi, Difan Liu, Feng Liu, Mingi Kwon, and Abhinav Shrivastava. Customize-a-video: One-shot motion customization of text-to-video diffusion models. In *European Conference on Computer Vision*, pages 332–349. Springer, 2024.

[51] Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. *Advances in neural information processing systems*, 33:11539–11551, 2020.

[52] Qingyu Shi, Jianzong Wu, Jinbin Bai, Jiangning Zhang, Lu Qi, Yunhai Tong, and Xiangtai Li. Decouple and track: Benchmarking and improving video diffusion transformers for motion transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10995–11005, 2025.

[53] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35:14274–14289, 2022.

[54] Yang Shu, Zhangjie Cao, Chenyu Wang, Jianmin Wang, and Mingsheng Long. Open domain generalization with domain-augmented meta-learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9624–9633, 2021.

[55] Raghav Singhal, Zachary Horvitz, Ryan Teehan, Mengye Ren, Zhou Yu, Kathleen McKeown, and Rajesh Ranganath. A general framework for inference-time scaling and steering of diffusion models. *arXiv preprint arXiv:2501.06848*, 2025.

[56] Yiren Song, Cheng Liu, Yuxin Jiang, and Mike Zheng Shou. Streamingeffect: Real-time human-centric video effect generation. *arXiv preprint arXiv:2605.17019*, 2026.

[57] Yiren Song, Wangzi Yao, Haofan Wang, and Mike Zheng Shou. Vista: Triplet-supervised video style transfer with diffusion transformers. *arXiv preprint arXiv:2605.17312*, 2026.

[58] Sruthi Sudhakar, Ruoshi Liu, Basile Van Hoorick, Carl Vondrick, and Richard Zemel. Controlling the world by sleight of hand. In *European Conference on Computer Vision*, pages 414–430. Springer, 2024.

[59] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pages 9229–9248. PMLR, 2020.

[60] Chengzhuo Tong, Ziyu Guo, Renrui Zhang, Wenyu Shan, Xinyu Wei, Zhenghao Xing, Hongsheng Li, and Pheng-Ann Heng. Delving into rl for image generation with cot: A study on dpo vs. grpo. *arXiv preprint arXiv:2505.17017*, 2025.

[61] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238, 2024.

[62] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[63] Renhao Wang, Yu Sun, Arnuv Tandon, Yossi Gandelsman, Xinlei Chen, Alexei A Efros, and Xiaolong Wang. Test-time training on video streams. *Journal of Machine Learning Research*, 26(9):1–29, 2025.

[64] Xinyu Wang, Chongbo Zhao, Fangneng Zhan, and Yue Ma. Liveedit: Towards real-time diffusion-based streaming video editing. In *European Conference on Computer Vision*, pages 631–649. Springer, 2026.

[65] Yucheng Wang, Zedong Wang, Yuetong Wu, Yue Ma, and Dan Xu. Care-edit: Condition-aware routing of experts for contextual image editing. *arXiv preprint arXiv:2603.08589*, 2026.

[66] Yuxi Wang, Wenqi Ouyang, Tianyi Wei, Yi Dong, Zhiqi Shen, and Xingang Pan. Hand2world: Autoregressive egocentric interaction generation via free-space hand gestures. *arXiv preprint arXiv:2602.09600*, 2026.

[67] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024.

[68] Jianzong Wu, Xiangtai Li, Yanhong Zeng, Jiangning Zhang, Qianyu Zhou, Yining Li, Kai Chen, and Yunhai Tong. Motionbooth: Motion-aware customized text-to-video generation. *Advances in Neural Information Processing Systems*, 37:34322–34348, 2024.

[69] Tsung-Han Wu, Long Lian, Joseph E Gonzalez, Boyi Li, and Trevor Darrell. Self-correcting llm-controlled diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6327–6336, 2024.

[70] Yuxi Xiao, Qianqian Wang, Shangzhan Zhang, Nan Xue, Sida Peng, Yujun Shen, and Xiaowei Zhou. Spatialtracker: Tracking any 2d pixels in 3d space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20406–20417, 2024.

[71] Zeqi Xiao, Yifan Zhou, Shuai Yang, and Xingang Pan. Video diffusion models are training-free motion interpreter and controller. *Advances in Neural Information Processing Systems*, 37:76115–76138, 2024.

[72] Enze Xie, Junsong Chen, Yuyang Zhao, Jincheng Yu, Ligeng Zhu, Chengyue Wu, Yujun Lin, Zhekai Zhang, Muyang Li, Junyu Chen, et al. Sana 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer. *arXiv preprint arXiv:2501.18427*, 2025.

[73] Jinbo Xing, Menghan Xia, Yuxin Liu, Yuechen Zhang, Yong Zhang, Yingqing He, Hanyuan Liu, Haoxin Chen, Xiaodong Cun, Xintao Wang, et al. Make-your-video: Customized video generation using textual and structural guidance. *IEEE transactions on visualization and computer graphics*, 31(2):1526–1541, 2024.

[74] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pages 399–417. Springer, 2024.

[75] Zunnan Xu, Zhihong Chen, Yong Zhang, Yibing Song, Xiang Wan, and Guanbin Li. Bridging vision and language encoders: Parameter-efficient tuning for referring image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17503–17512, 2023.

[76] Zunnan Xu, Yukang Lin, Haonan Han, Sicheng Yang, Ronghui Li, Yachao Zhang, and Xiu Li. Mambatalk: Efficient holistic gesture synthesis with selective state space models. *Advances in Neural Information Processing Systems*, 37:20055–20080, 2024.

[77] Hidir Yesiltepe, Tuna Han Salih Meral, Connor Dunlop, and Pinar Yanardag. Motionshop: Zero-shot motion transfer in video diffusion models with mixture of score guidance. *arXiv preprint arXiv:2412.05355*, 2024.

[78] Zhengyang Yu, Akio Hayakawa, Masato Ishii, Qingtao Yu, Takashi Shibuya, Jing Zhang, and Yuki Mitsufuji. Autorefiner: Improving autoregressive video diffusion models via reflective refinement over the stochastic sampling path. *arXiv preprint arXiv:2512.11203*, 2025.

[79] Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. *Advances in neural information processing systems*, 35:38629–38642, 2022.

[80] Ruicheng Zhang, Guangyu Chen, Zunnan Xu, Zihao Liu, Zhizhou Zhong, Mingyang Zhang, Jun Zhou, and Xiu Li. Robostereo: Dual-tower 4d embodied world models for unified policy optimization. *arXiv preprint arXiv:2603.12639*, 2026.

[81] Wanyue Zhang, Lin Geng Foo, Thabo Beeler, Rishabh Dabral, and Christian Theobalt. Vhoi: Controllable video generation of human-object interactions from sparse trajectories via motion densification. *arXiv preprint arXiv:2512.09646*, 2025.

[82] Yabo Zhang, Xinpeng Zhou, Yihan Zeng, Hang Xu, Hui Li, and Wangmeng Zuo. Framepainter: Endowing interactive image editing with video diffusion priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18121–18131, 2025.

[83] Min Zhao, Rongzhen Wang, Fan Bao, Chongxuan Li, and Jun Zhu. Controlvideo: Adding conditional control for one shot text-to-video editing. *arXiv preprint arXiv:2305.17098*, 2(3), 2023.

[84] Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jia-Wei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. Motiondirector: Motion customization of text-to-video diffusion models. In *European Conference on Computer Vision*, pages 273–290. Springer, 2024.

[85] Le Zhuo, Liangbing Zhao, Sayak Paul, Yue Liao, Renrui Zhang, Yi Xin, Peng Gao, Mohamed Elhoseiny, and Hongsheng Li. From reflection to perfection: Scaling inference-time optimization for text-to-image diffusion models via reflection tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15329–15339, 2025.

A Related Work

Test-Time Visual Generation. Test-time methods enhance pre-trained models during inference without retraining [34, 59, 63]. Existing approaches primarily follow two paradigms. First, test-time scaling improves quality through iterative candidate search [14, 36], preference optimization [60, 61], or VLM-guided reflection [72, 85]. However, this process requires prohibitive computational costs and shows marginal improvements for continuous diffusion models due to inherent architectural constraints [49, 55]. Second, test-time optimization refines outputs via auxiliary parameter updates [6, 7, 16, 33, 51, 54, 78, 79], prompt tuning [53], or layout-guided latent optimization [25, 30, 32, 69], but often relies on complex instance-specific training or global objectives. In contrast, our test-time refinements target the hand-object structure and local correspondences needed for egocentric action transfer.

B More Implementation Details

We use WAN-2.1 [62] as the base video generation model and keep all model parameters frozen during inference. For ablation study, all ablation variants are evaluated under the same setting as the main comparison, with only the specified component removed or modified. The latent video is initialized with Gaussian noise of shape $1 \times 16 \times (\frac{F-1}{4}+1) \times \frac{H}{8} \times \frac{W}{8}$, where F is internally rounded to $4k+1$. We enable latent tiling in the VAE space with `tile_size=`$(30, 52)$ and `tile_stride=`$(15, 26)$, which produces a per-frame DiT token grid of $h = \frac{H}{8}$ by $w = \frac{W}{8}$.

AMF-based structural mask. As described in the main text (Sec. 3.2), AMF is used solely to construct the soft structural mask $\mathcal{M}$ for VSA, not as a separate optimization objective. For sliding-window AMF computation, we use a tile grid of 3×4 tokens, temporal span $s_f = 5$, local search window $l = 21$ with half-width 10, and temperature $\tau = 1.0$. The AMF magnitude from the reference branch is computed over temporal offsets with linear weights decayed from 1.0 to 0.8 across offsets, normalized by the number of offsets and tokens, and then converted into the structural saliency mask via Eq. 1–2. The actual optimization objective is the mask-weighted velocity regression $\mathcal{L}_{\text{VSA}}$ (Eq. 3), and gradients are applied only to the latent variable. To reduce computational cost during the inner optimization loop, we recompute Q/K features and tile-level AMF every 3 steps and reuse the cached mask on intermediate steps.

Optimization. At each guided diffusion step t, we form the reference latent by adding step-consistent noise to the cached clean latent and run a null-text forward pass to extract query/key features from the 15th DiT block. Each guided step performs 10 latent-only AdamW updates with a linear learning-rate decay from 0.003 to 0.002. The VSA weight, SCC correspondence weight, and stay regularization weight are set to 15.0, 1.0, and 0.05, respectively.

Inference time. On an NVIDIA A100 GPU, the average inference time for a 81-frame video of resolution 832×480 is approximately 413 seconds, which includes the base diffusion process and the additional test-time optimization steps. The VSA module adds around 23 seconds, while the SCC module contributes an additional 72 seconds. The overall time is still within a practical range for test-time refinement, given the improved quality and consistency of the generated videos.

C More Details about Attention Motion Flow

Attention Motion Flow (AMF) [43] extracts implicit motion cues from the self-attention layers of a video diffusion transformer. The key observation is that cross-frame query-key affinity in DiT attention naturally encodes patch-level correspondences: if a token in one frame attends strongly to a token in a neighboring frame, their spatial offset provides an estimate of the local motion. Instead of using external trackers or optical-flow estimators, we compute AMF directly from the internal attention features of the frozen generation model. To instantiate this intuition, we compute token-level correspondences from local query-key affinities within self-attention.

Local Attention-based Correspondence. For a spatial token u in frame i, we search its correspondence in a nearby frame j within a local displacement window $\mathcal{N}(u) = \{u + \Delta \mid \Delta \in [-r, r]^2\}$. Let

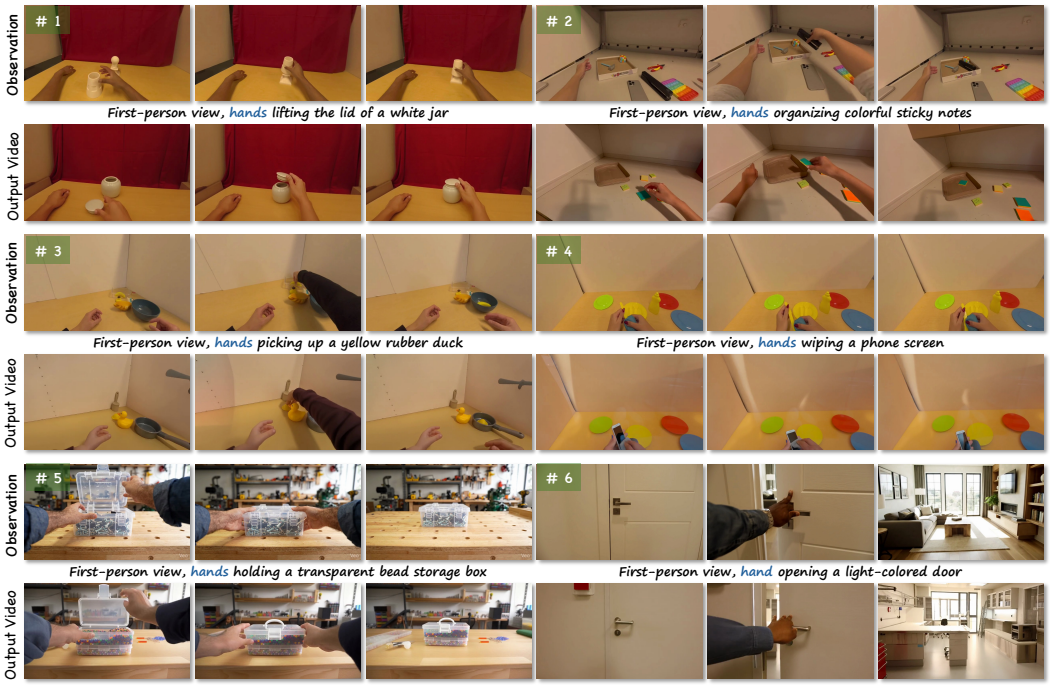

Figure 10: More visual results.

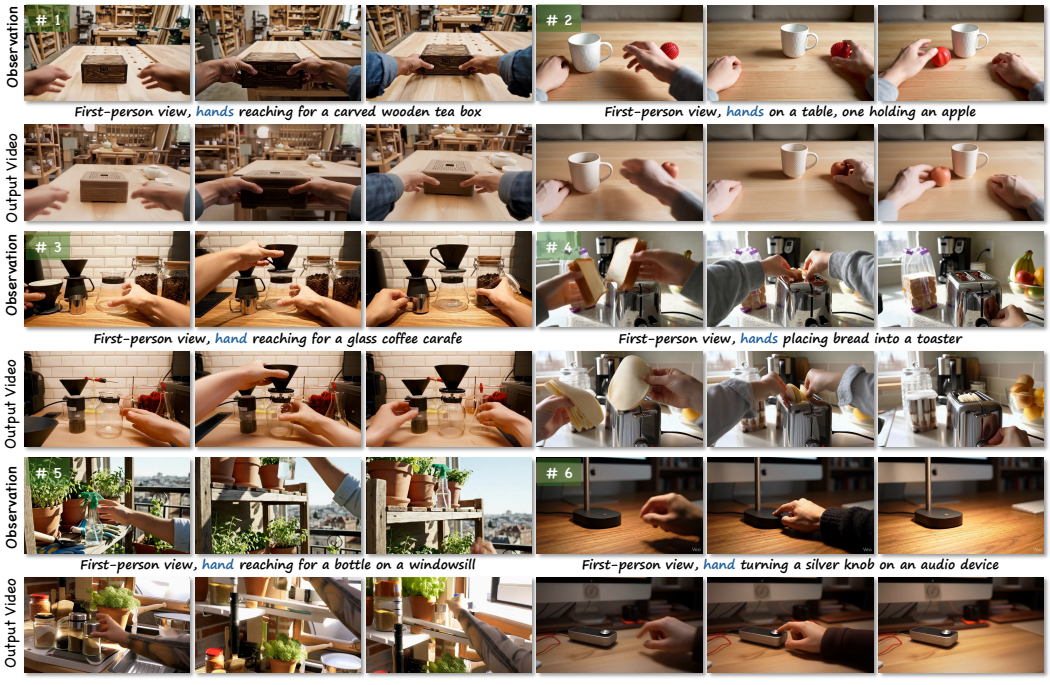

Figure 11: More visual results.

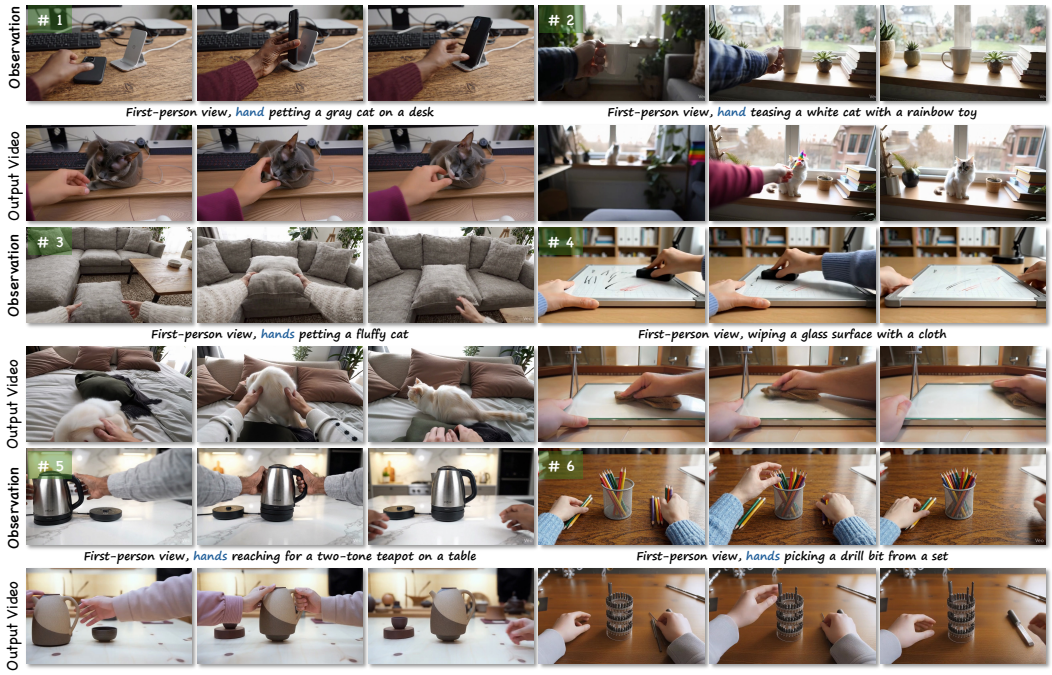

Figure 12: More visual results.

$q_i(u)$ be the query feature at token u and $k_j(u+\Delta)$ be the key feature at the candidate location. We compute the local attention score as:

$$s_{i\to j}(u,\Delta) = \frac{\langle q_i(u), k_j(u+\Delta)\rangle}{\sqrt{d}}, \quad \Delta \in [-r,r]^2, \tag{8}$$

where d is the feature dimension. The normalized local correspondence distribution is then

$$p_{i\to j}(u,\Delta) = \frac{\exp(s_{i\to j}(u,\Delta))}{\sum_{\Delta'\in[-r,r]^2}\exp(s_{i\to j}(u,\Delta'))}. \tag{9}$$

Following the soft motion extraction strategy in FastVMT, we use the expectation over local displacements to obtain a differentiable AMF vector:

$$\mathcal{A}_{i\to j}(u) = \sum_{\Delta\in[-r,r]^2} p_{i\to j}(u,\Delta)\Delta. \tag{10}$$

This sliding-window design reflects the local continuity of video motion and avoids unreliable global matching between unrelated tokens.

AMF-induced Structural Mask. In our VSA module, AMF is not used as a hard motion supervision signal; instead, it serves as a cue for locating interaction-relevant structures. We compute AMF from the reference branch and convert its magnitude into a structural saliency score:

$$\mathcal{S}(u) = \|\mathcal{A}_{\text{ref}}(u)\|_2, \tag{11}$$

where larger values usually indicate regions with stronger action dynamics, such as moving hands, manipulated objects, and hand-object contact areas. Intuitively, these regions carry most of the structural and physical constraints of egocentric manipulation, so they should receive stronger anchoring during optimization.

After normalizing the structural saliency score over with Eq. 2, we then apply a power reweighting to control the concentration of the mask:

$$\mathcal{M}(u) = \hat{\mathcal{S}}(u)^{\gamma}(1 - m_{\min}) + m_{\min}. \tag{12}$$

where, $\gamma > 1$ sharpens the mask and emphasizes the most motion-salient regions, while $\gamma < 1$ produces a flatter weighting profile. The lower bound $m_{\min}$ ensures that low-motion regions still contribute weakly to the velocity anchoring objective, which helps preserve global scene layout and prevents the optimization from overfitting only to highly dynamic tokens.

Therefore, the AMF-induced mask allocates optimization strength adaptively: dynamic interaction regions are emphasized, whereas static regions are softly regularized. This design allows VSA in Sec. 3.2 to focus on the action-critical structure without imposing unnecessarily strong constraints on unrelated background content.

D More Details about User Study

To complement automatic metrics, we conduct a user study with 20 volunteers on 50 randomly selected cases from EgoActionBench. For each evaluated case, volunteers are shown the reference egocentric manipulation video, the target text prompt, and generated videos from all compared methods. The method names are hidden, and the generated videos are presented in randomized order to reduce ordering bias.

Rating Dimensions. Annotators evaluate each generated result according to the following four criteria:

- **Appearance Consistency:** whether the generated video preserves a coherent target appearance across frames without obvious identity drift, texture inconsistency, or visual artifacts.
- **Text Alignment:** whether the generated video follows the target prompt, including the described scene, object appearance, and intended manipulation context.
- **Temporal Consistency:** whether the generated video remains stable over time without obvious flickering, jittering, ghosting, or frame-to-frame structural instability.
- **Motion Fidelity:** whether the generated video preserves the reference action pattern, including hand motion, contact timing, and object response, without severe drifting or motion collapse.

Scoring Scale. Each dimension is rated using a four-level ordinal scale: **Excellent** (fully satisfies the criterion with no obvious artifacts), **Good** (minor artifacts or deviations), **Pass** (noticeable but acceptable issues), and **Fail** (severe artifacts or failure to satisfy the criterion). The final statistics are obtained by aggregating ratings across all volunteers and evaluated cases.

E More Details about EgoActionBench

EgoActionBench focuses on whether the generated video preserves the interaction logic of the reference action, including contact transitions, object response, and coordinated hand-object dynamics. The main paper summarizes the data sources, category coverage, and basic filtering rules of EgoActionBench. The benchmark contains 350 reference videos, with 4 target prompts for each video, resulting in 1,400 evaluation prompts. It includes 287 indoor scenes and 63 outdoor scenes, with 164 single-hand and 186 two-hand manipulation cases. In terms of data composition, approximately 60% of the reference videos are sourced from real-world egocentric recordings (online videos and public manipulation datasets), while the remaining 40% are generated by video diffusion models to cover rarer manipulation categories and ensure diversity. Here we provide additional details about camera-motion filtering, prompt construction, and manual quality control.

Camera-motion filtering. Egocentric videos often contain strong head or body motion, which can make action transfer evaluation ambiguous: a method may appear to fail because the reference observation is dominated by camera shake rather than hand-object interaction. To reduce this issue, we estimate camera poses with ViPE [22] for candidate real videos and compute frame-to-frame camera translation and rotation changes. Clips with large pose jumps, unstable pose estimates, or camera motion that dominates the hand-object region are removed before manual inspection. This



Figure 13: **More details about EgoActionBench.** EgoActionBench covers diverse egocentric manipulation scenarios and is annotated by scene type, and object category.

filtering step keeps moderate egocentric viewpoint changes while excluding references where the manipulated object cannot be reliably tracked in the image plane.

Caption and prompt instructions. For each retained clip, we ask GPT-5 to produce an action-centric caption using the following instruction: *Describe the egocentric manipulation action in one sentence. Focus on the actor's hand motion, the manipulated object, the contact transition, and the object's response. Do not describe camera motion, video quality, identity, or irrelevant background details.* Based on this source caption, we further ask GPT-5 to generate target prompts with the instruction: *Rewrite the scene, object appearance, or background while preserving the same manipulation verb, contact order, and object response. Avoid copying reference-specific colors, textures, or scene details unless they are essential to the action.* This prompt design encourages evaluation cases that test action transfer rather than source appearance copying.

Manual quality control. After automatic captioning and prompt rewriting, we manually review each benchmark sample. We reject prompts that change the action semantics, introduce physically incompatible objects, leak reference-specific appearance details, or describe interactions that are not visible in the source clip. We also check whether the target prompt is feasible for the reference action; for example, actions involving pouring, opening, or handover require objects with compatible affordances. The remaining samples are annotated with difficulty tags according to contact complexity, object motion magnitude, and whether the action involves one or two hands. These tags are used only for analysis and are not provided to the generation methods.

F More Results

More Comparison with Baseline. We provide more visual comparisons with baselines in Fig. 15. Our method consistently produces more coherent action transfer and more plausible hand-object



Figure 14: **Failure case.** Our method may struggle with actions involving very fast hand or object motion, which can lead to motion blur, temporal aliasing, or loss of hand-object correspondence. This limitation is primarily due to the base video diffusion model, which was not specifically trained to handle rapid egocentric dynamics and tends to produce temporally smoothed outputs when the inter-frame displacement is large.

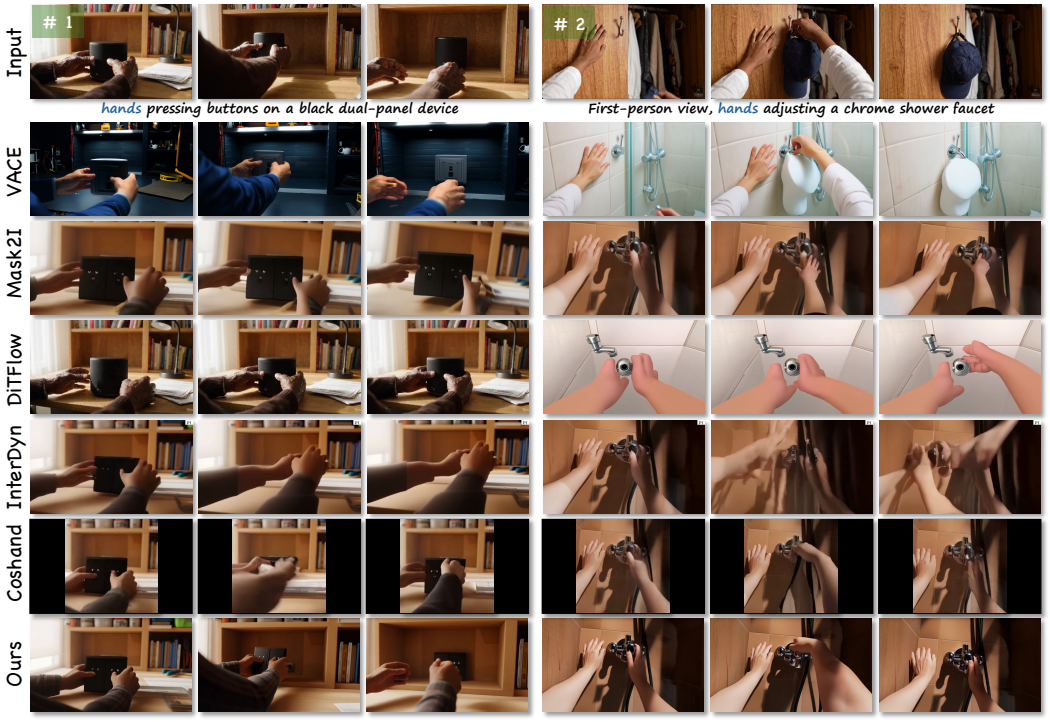

Figure 15: **More Comparison with Baseline.** Our method consistently produces more coherent action transfer and more plausible hand-object interactions than baselines, which often suffer from structural collapse, temporal inconsistency, or physically unrealistic interactions.

interactions than baselines, which often suffer from structural collapse, temporal inconsistency, or physically unrealistic interactions.

More Visual Results. Figures 10, 11, and 12 show more visual results of our method. The generated videos can be viewed in the demo video.

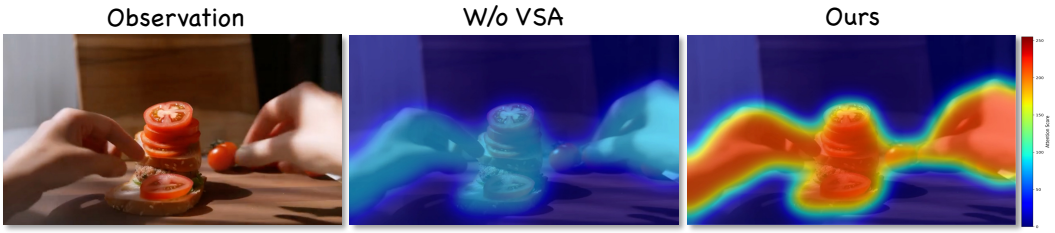

Figure 16: **Attention Visualization.** We visualize the cross-frame attention maps before and after enabling VSA during the early denoising stage. Without VSA, attention is dispersed across irrelevant regions, failing to capture coherent hand-object structure. With VSA enabled, the attention concentrates on interaction-relevant areas such as hand, object, and contact regions, indicating that the generation trajectory is better aligned with the reference structural layout. This confirms that VSA effectively anchors the hand-object skeleton through velocity-guided alignment, resulting in more structurally consistent attention patterns.

G Limitations and Future Work

Although EgoACT achieves strong results on a wide range of egocentric manipulation scenarios, it still struggles with actions involving very fast hand or object motion. As shown in Fig. 14, the generated videos may exhibit motion blur, temporal aliasing, or loss of hand-object correspondence in such cases. We attribute this limitation primarily to the base video diffusion model, which was not specifically trained to handle rapid egocentric dynamics and tends to produce temporally smoothed outputs when the inter-frame displacement is large. As video foundation models continue to advance in temporal resolution and motion modeling capacity, we expect this limitation to be naturally alleviated. In future work, we also plan to extend EgoACT to longer-horizon action sequences and multi-step task compositions, enabling more complex manipulation planning within the world simulation paradigm.

H Potential Social Impact

Our work studies egocentric action transfer for controllable video world simulation. The potential positive impacts include more efficient simulation for embodied AI, robotics, and human-object interaction research, where generating diverse interaction videos can reduce the cost of collecting physical demonstrations and enable safer testing before real-world deployment. The method may also support content creation and instructional video synthesis when used with appropriate consent and attribution. At the same time, the ability to generate realistic egocentric manipulation videos may be misused to create misleading visual content or to synthesize actions that did not actually occur. Since our method builds on pre-trained video generation models, it may also inherit biases, failure modes, or unsafe generation behaviors from the underlying model and source data. To mitigate these risks, released examples and benchmark assets should be filtered for sensitive or harmful content, and generated videos should be clearly identified as synthetic when used outside research settings. We also recommend applying existing provenance, watermarking, or content moderation tools when deploying such systems in user-facing applications.

Algorithm 1 EgoACT test-time latent optimization

Require: Reference clean latent x_0^{ref}, target prompt c, frozen video diffusion model V_θ, timesteps $\{t_j\}_{j=0}^{T-1}$, VSA step set $\mathcal{T}_{\mathrm{VSA}}$, SCC step set $\mathcal{T}_{\mathrm{SCC}}$, VSA iterations K_{VSA}, SCC iterations K_{SCC}, local window $\mathcal{N}$, top-k size k, learning rate η, SCC weights $\lambda_{\mathrm{c}}, \lambda_{\mathrm{s}}$

Ensure: Generated video x^{gen}

$z_T \sim \mathcal{N}(0, I)$
for $j = T-1$ to 0 **do**
  Predict velocity/noise for z_{t_j} with $V_\theta(\cdot, t_j, c)$
  $z_{t_j}^{\mathrm{opt}} \leftarrow z_{t_j}$
  Velocity-guided Structure Anchoring
  if $t_j \in \mathcal{T}_{\mathrm{VSA}}$ **then**
    $z_{t_j}^{\mathrm{ref}} \leftarrow q(z \mid x_0^{\mathrm{ref}}, t_j)$
    Extract reference Q/K features from $z_{t_j}^{\mathrm{ref}}$
    Compute reference AMF $\mathcal{A}_{\mathrm{ref}}$ and structural mask $\mathcal{M}$
    $v_{\mathrm{anchor}} \leftarrow \mathrm{sg}[V_\theta(z_{t_j}^{\mathrm{ref}}, t_j)]$
    for $n = 1$ to K_{VSA} **do**
      $v_{\mathrm{opt}} \leftarrow V_\theta(z_{t_j}^{\mathrm{opt}}, t_j, c)$
      $\mathcal{L}_{\mathrm{VSA}} \leftarrow N^{-1} \sum_{\mathbf{x}} \mathcal{M}(\mathbf{x}) \| v_{\mathrm{opt}}(\mathbf{x}) - v_{\mathrm{anchor}}(\mathbf{x}) \|_2^2$
      $z_{t_j}^{\mathrm{opt}} \leftarrow z_{t_j}^{\mathrm{opt}} - \eta \nabla_{z_{t_j}^{\mathrm{opt}}} \mathcal{L}_{\mathrm{VSA}}$
    end for
  end if
  $(z_{t_{j+1}}, \hat{x}_0^{(j)}) \leftarrow \mathrm{SchedulerStep}(z_{t_j}^{\mathrm{opt}}, t_j)$
  $z_{t_{j+1}}^{\mathrm{opt}} \leftarrow z_{t_{j+1}}$
  Sparse Correspondence Calibration
  if $t_j \in \mathcal{T}_{\mathrm{SCC}}$ **then**
    $z_{t_{j+1}}^{\mathrm{ref}} \leftarrow q(z \mid x_0^{\mathrm{ref}}, t_{j+1}; \epsilon_1)$
    $z_{t_{j+1}}^{\mathrm{opt}} \leftarrow q(z \mid \hat{x}_0^{(j)}, t_{j+1}; \epsilon_1)$
    Extract $\{q_{\mathrm{ref}}^i, k_{\mathrm{ref}}^{i+1}\}$ and $\{q_{\mathrm{gen}}^i, k_{\mathrm{gen}}^{i+1}\}$
    for (i, u) **do**
      $\Omega_k^i(u) \leftarrow \mathrm{TopK}_{\Delta \in \mathcal{N}(u)} \langle q_{\mathrm{ref}}^i(u), k_{\mathrm{ref}}^{i+1}(u + \Delta) \rangle / \sqrt{d}$
      $\mathcal{P}_{\mathrm{ref},i}, \mathcal{P}_{\mathrm{gen},i} \leftarrow \mathrm{softmax}_{\Delta \in \Omega_k^i(u)}(s_i / \tau)$
      $\bar{w}_i(u) \leftarrow (1 - H_i(u) / \log k) / \sum_{u'} w_{\mathrm{conf}}^i(u')$
    end for
    for $n = 1$ to K_{SCC} **do**
      $\mathcal{L}_{\mathrm{corr}} \leftarrow -\sum_i \sum_u \bar{w}_i(u) \sum_{\Delta \in \Omega_k^i(u)} \mathcal{P}_{\mathrm{ref},i}(\Delta \mid u) \log \mathcal{P}_{\mathrm{gen},i}(\Delta \mid u)$
      $\mathcal{L}_{\mathrm{stay}} \leftarrow \| z_{t_{j+1}}^{\mathrm{opt}} - \mathrm{sg}[z_{t_{j+1}}] \|_2^2$
      $\mathcal{L}_{\mathrm{SCC}} \leftarrow \lambda_{\mathrm{c}} \mathcal{L}_{\mathrm{corr}} + \lambda_{\mathrm{s}} \mathcal{L}_{\mathrm{stay}}$
      $z_{t_{j+1}}^{\mathrm{opt}} \leftarrow z_{t_{j+1}}^{\mathrm{opt}} - \eta \nabla_{z_{t_{j+1}}^{\mathrm{opt}}} \mathcal{L}_{\mathrm{SCC}}$
    end for
  end if
  $z_{t_{j+1}} \leftarrow z_{t_{j+1}}^{\mathrm{opt}}$
end for
$x^{\mathrm{gen}} \leftarrow \mathrm{VAE.decode}(z_0)$

return x^{gen}

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The claims are stated in the abstract and introduction and are supported by the method, experiments, and ablations.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: The paper discusses limitations and failure cases in the limitations section.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A].

Justification: The paper does not present new theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes].

   Justification: The paper and appendix describe the model, training setup, data processing, metrics, and evaluation protocol.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes].

Justification: We provide anonymized code, scripts, configurations, and data-access instructions in the supplementary material.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes].

Justification: Training details, data splits, hyperparameters, and evaluation settings are specified in the paper and appendix.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes].

Justification: We report mean results with variability estimates where applicable.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes].

   Justification: The paper reports the hardware, memory, runtime, and training cost used for the experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: We reviewed and followed the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes].

    Justification: The paper discusses beneficial uses, possible misuse, and mitigation measures.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes].

    Justification: The paper describes license-aware release, consent-based use, and restrictions against deceptive or non-consensual use.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes].

    Justification: Existing datasets, assets, code, and models are cited with license information where available.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes].

    Justification: The new benchmark and released artifacts are documented in the paper, appendix, and supplementary material.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A].

    Justification: The paper does not involve crowdsourcing or human-subject studies.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A].

    Justification: The paper does not involve human-subject studies, so IRB approval is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A].

    Justification: LLMs are not used as a core methodological component.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.