STREAMEGO: STREAMING EXOCENTRIC-TO-EGOCENTRIC VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Exocentric-to-egocentric (Exo2Ego) video generation transforms third-person recordings into first-person viewpoints, unlocking immersive AR/VR experiences and scalable demonstration synthesis for embodied robotics. Extending it to a streaming setting is expected to further enable live applications. However, existing methods suffer from degraded manipulation regions, overfitting and confinement to fixed-length clips with prohibitive latency. To address these limitations, we present StreamEgo, a real-time framework for unbounded, generalizable Exo2Ego video generation. To resolve intricate manipulation dynamics while supporting the streaming pipeline, a streaming dual-branch prior provides pixel-aligned guidance for translating appearance and skeletal kinematics. We also introduce a region-aware loss curriculum to improve generalization and stabilize multi-objective diffusion training. Furthermore, we distill the model into a chunk-wise autoregressive model, augmented with a region-aware corruption mechanism and a first-chunk curriculum, enabling high-quality, generalizable real-time streaming Exo2Ego generation at 24.3 FPS on a single H100 GPU.

1 INTRODUCTION

Exocentric-to-egocentric (Exo2Ego) video generation (Liu et al., 2024; Luo et al., 2024; Kang et al., 2026; Mahdi et al., 2025; Xu et al., 2025; Wang et al., 2026; Mahdi et al., 2026) converts third-person recordings into dynamic first-person viewpoints. For embodied AI, Exo2Ego unlocks massive web-scale video repositories as a scalable and diverse data source. Real-time streaming Exo2Ego further extends this capability to live applications such as AR-based tutorials and surgical demonstrations.

To achieve accurate Exo2Ego video generation, recent studies (Kang et al., 2026; Xu et al., 2025) formulate Exo2Ego by decomposing the egocentric canvas into three functional regimes (Fig. 1:**Left**): the hand-object interaction area ($\mathcal{H}$), the shared view ($\mathcal{S}$), and the unseen view ($\mathcal{U}$). By leveraging geometric reprojections from the exocentric view to construct appearance priors, existing methods (Kang et al., 2026; Shi et al., 2026; Wang et al., 2026) achieve robust alignment and satisfactory visual fidelity in $\mathcal{S}$.

However, they fall short across the remaining domains. Firstly, accurate hand-object interaction remains challenging in $\mathcal{H}$, as projective geometric priors break down under complex foreground dynamics, while existing mask- (Xu et al., 2025) or pose-based (Park et al., 2026) remedies are fragile due to their reliance on ground-truth first-frame or noisy hand pose estimates (Fig. 1:**Right**, red box). Meanwhile, existing approaches suffer from benchmark memorization on paired datasets (Grauman et al., 2024), artificially replicating ground-truth details in $\mathcal{U}$ and impairing generalization (Fig. 1:**Right**, yellow box). Finally, existing methods rely on full-sequence offline scanning for appearance priors (Shi et al., 2026; Kang et al., 2026; Wang et al., 2026) and full-sequence bidirectional attention, implicitly assuming access to future frames. This non-causal paradigm is fundamentally incompatible with streaming, where the future is unknown (Fig. 1:**Right**, 0.08 FPS).

In this paper, we propose StreamEgo, a real-time, streaming, and generalizable framework for Exo2Ego video generation. To resolve fine-grained manipulation in $\mathcal{H}$ under the streaming setting, we establish a pixel-aligned conditioning scheme via streaming dual-branch priors. A dynamic ego prior maintains an incremental 3D world cache for the static background and tracks moving entities to render ego-aligned appearance anchors, while a human pose prior provides structural guidance for hand interactions. To improve the quality in $\mathcal{U}$, we propose a region-aware training curriculum.

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

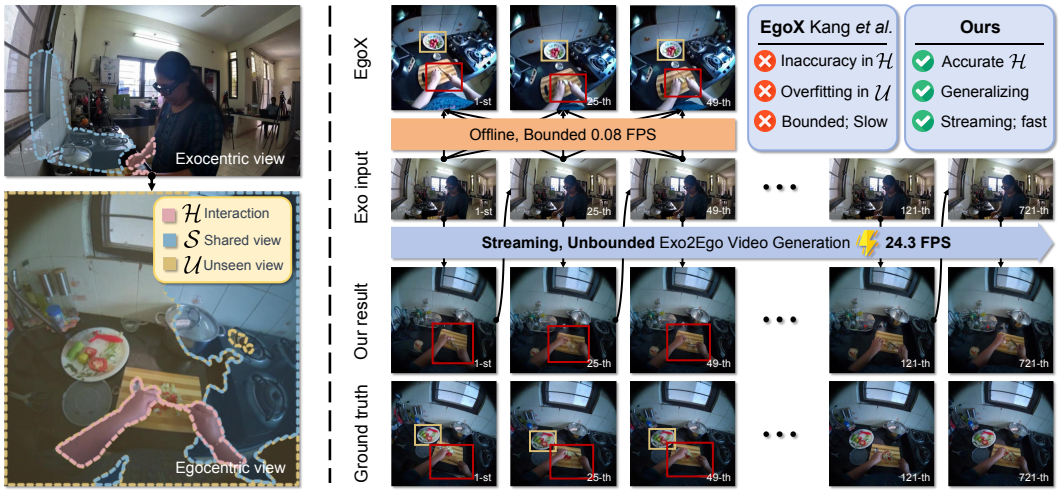

Figure 1: Streaming exocentric-to-egocentric (Exo2Ego) video generation. **Left:** Tri-regional decomposition of the task of Exo2Ego. **Right:** Comparison of our method with state-of-the-art Exo2Ego method EgoX (Kang et al., 2026) on Ego-Exo4D dataset (Grauman et al., 2024).

It progressively shifts the optimization focus from $\mathcal{S}$ to $\mathcal{U}$ and then to $\mathcal{H}$. This curriculum stabilizes training while improving generalization. To support streaming under a causal setting, we adapt a chunk-wise autoregressive architecture, where each chunk is generated only from past chunks, and distill the diffusion backbone into a high-throughput causal student with 3 steps. This removes the reliance on full-sequence offline scanning and bidirectional attention. Meanwhile, we propose a region-aware corruption mechanism and a first-chunk curriculum to improve the quality for the streaming rollouts. As a result, StreamEgo can process unbounded exocentric video streams and generate high-quality egocentric videos at 24.3 FPS on a single H100 GPU.

Our main contributions are summarized as follows:

- We present StreamEgo, the first streaming framework for Exo2Ego video generation, reformulating the task from full-sequence generation to causal generation.
- We introduce a streaming dual-branch prior and a region-aware training curriculum, which respectively improve interaction fidelity in $\mathcal{H}$ and generalization in $\mathcal{U}$.
- We introduce a region-aware corruption mechanism and a first-chunk curriculum, enhancing the visual quality for StreamEgo, achieving high-quality real-time Exo2Ego generation.

2 Related Work

2.1 Exocentric-to-Egocentric Video Generation

Paired Exo-Ego datasets (Grauman et al., 2024; Kwon et al., 2021; Sener et al., 2022) have catalyzed exocentric-to-egocentric synthesis. Early attempts simplify the setup via multi-view inputs (Liu et al., 2024; Mahdi et al., 2025), initial ego guidance (Xu et al., 2025), or static image translation (Park et al., 2026). Concurrently, novel-view synthesis methods inject camera encodings (Bai et al., 2025a;b; Li et al., 2025a; Zhang et al., 2026; Yu et al., 2025) or condition on rendered 3D/4D geometry (Ren et al., 2025; Cao et al., 2025; Bahmani et al., 2026; Lin et al., 2026), but falter under extreme perspective shifts. Recent single-exo works incorporate geometry guidance (Kang et al., 2026) or semantic grounding (Wang et al., 2026), yet uniform canvas supervision causes benchmark memorization and hand drift across short, fixed-length clips. In contrast, StreamEgo addresses these problems with streaming dual-branch priors and region-aware curriculum scheduling, enabling generalizable, real-time streaming synthesis.

2.2 Streaming Video Generation Models

To break the temporal limits of short-clip diffusion, training-free noise rescheduling (Qiu et al., 2024; Kim et al., 2024) and training-based autoregressive conditioning (Chen et al., 2024; Ruhe et al., 2024; Sun et al., 2025; Song et al., 2025; Huang et al., 2025) have been proposed to mitigate exposure bias. Subsequent frameworks scale temporal length via predictive modelling, long-context architectures, and causal forcing (Zhang et al., 2025b; Gu et al., 2025; Liu et al., 2025a; Li et al., 2025c; Liu et al., 2025b; Zhu et al., 2026). However, these methods focus on text-to-video generation rather than asymmetric cross-view streaming. Built upon streaming long-video designs and few-step distillation paradigms (Yuan et al., 2026; Zhao et al., 2023; Yin et al., 2024b;a; 2025), StreamEgo achieves chunk-wise autoregressive modelling for the Exo2Ego task.

2.3 Video World Models for Embodied Learning

Video world models facilitate embodied learning via interactive simulation (Yang et al., 2023), visual planning (Du et al., 2023; Zhou et al., 2024), and synthetic demonstrations (Bharadhwaj et al., 2024). Subsequent methods couple video with action prediction (Li et al., 2025b; Zhu et al., 2025), adapt video priors via inverse dynamics (Feng et al., 2025), or incorporate causal history for online control (Li et al., 2026a). Although cross-view conversion can furnish training data for egocentric world models (Tran et al., 2026), synthesizing continuous, structurally faithful streams remains challenging. StreamEgo bridges this gap by streaming third-person observations into structurally consistent egocentric trajectories, providing scalable first-person training data for embodied AI.

3 Method

Given an input exocentric video stream along with known exocentric and target egocentric camera intrinsics, and the target egocentric camera trajectory, the goal for StreamEgo is to synthesize the corresponding egocentric video from the target viewpoint with the same frame rate as the input.

We develop StreamEgo in three stages: formulating streaming priors (Sec. 3.1), building the base system with the priors (Sec. 3.2), and making the system real-time (Sec. 3.3).

3.1 Streaming Dual-Branch Priors

Exo2Ego is intrinsically underconstrained because of extreme perspective disparities. This challenge is worse in $\mathcal{H}$, where motion is rapid and occlusions are frequent. The model therefore benefits from pixel-aligned conditions that provide visual context and structural guidance. At the same time, practical online deployment requires these conditions to be built causally and in a streaming manner. To this end, we introduce streaming dual-branch priors (Fig. 2**a**): (1) a streaming dynamic ego prior for appearance cues, paired with (2) a streaming human pose prior for structural guidance.

Streaming dynamic ego prior. Unlike existing methods (Kang et al., 2026; Shi et al., 2026) that rely on offline, full-video inputs with prohibitive latency, we introduce a streaming geometric conditioning scheme to construct causal pixel-aligned visual priors in real-time.

As is shown in Fig. 2**a**, we decouple the environment into persistent static background surfaces and transient interaction geometry, following CamDirector (Shi et al., 2026). Using the stationary exocentric camera, each rectified pixel corresponds to a fixed physical ray in world space. We maintain an incremental, ray-indexed 3D world cache that stores the first valid depth and color observed along each ray. The cache updates at sparse keyframes. MoGe-2 (Wang et al., 2025b) predicts metric depth, while SAM3.1 (Meta AI, 2026) segments dynamic entities to isolate the interaction region. Static rays outside dynamic masks are committed to the cache and remain fixed once initialized. Depth is propagated via backward optical flow between keyframes.

At frame t, the retrieved static background points $\mathcal{A}_t$ and the current dynamic foreground points $\mathcal{D}_t$ are unified in world space. The ego-aligned prior P_t is synthesized via:

$$P_t = \mathcal{R}_{\mathcal{C}_{g,t}}\left(T_{g,t}^{-1}(\mathcal{A}_t \cup \mathcal{D}_t)\right), \tag{1}$$

where $T_{g,t}$ is the transformation from target egocentric coordinates to world space, and $\mathcal{R}_{\mathcal{C}_{g,t}}$ is a calibrated splatting renderer parameterized by target camera intrinsics $\mathcal{C}_{g,t}$. This formulation provides stable anchors for region $\mathcal{S}$ and dynamic visual cues for contacts in $\mathcal{H}$ in a streaming fashion.

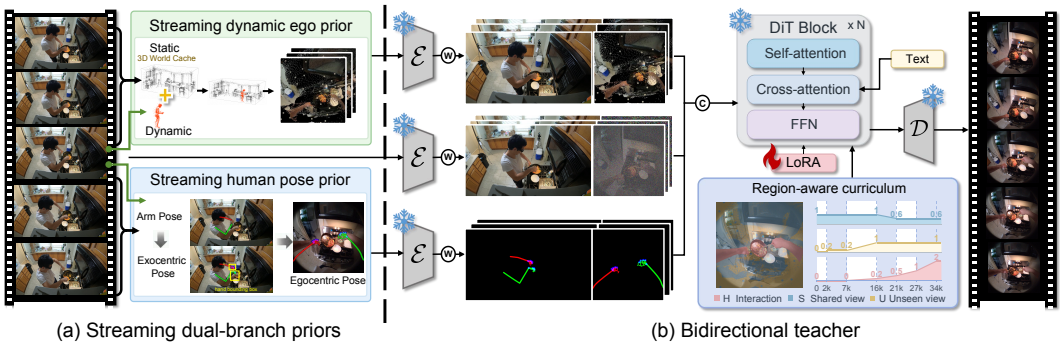

Figure 2: Overview of the bidirectional teacher. **(a)** Causal construction of the streaming dual-branch priors, comprising the dynamic ego prior and human pose prior. **(b)** Conditioning assembly and fine-tuning of the base Wan model (Team Wan et al., 2025) under our region-aware curriculum.

Streaming human pose prior. While dynamic ego priors supply visual context, accurately synthesizing complex interactions in $\mathcal{H}$ requires explicit structural guidance (Park et al., 2026). Prior attempts either rely on ground-truth egocentric first frames for interaction masks (Xu et al., 2025), or condition directly on isolated hand poses (Park et al., 2026); the latter induces severe spatial misalignment on the egocentric canvas because monocular depth estimation on small, dynamic hand regions is highly ill-posed. Inspired by human proprioception, we introduce skeletal constraints by kinematically bonding hand estimation to the upper-body pose.

At sampled exocentric frames, SMPLer-X (Cai et al., 2023) estimates and ViTPose (Xu et al., 2022) refines full-body topology and upper-limb joints. Wrist joints derived from the upper-body chain define zoom-in bounding boxes for WiLoR (Potamias et al., 2025) to recover wrist-relative finger articulation. For each hand $s \in \{\mathrm{L}, \mathrm{R}\}$ (left and right) at frame t, the predicted j-th local finger joint $\widehat{J}^s_{t,j}$ is kinematically anchored to the refined 3D wrist joint position $B^s_{t,\mathrm{wrist}}$:

$$J^s_{t,j} = B^s_{t,\mathrm{wrist}} + \left(\widehat{J}^s_{t,j} - \widehat{J}^s_{t,\mathrm{root}}\right), \qquad s \in \{\mathrm{L}, \mathrm{R}\}, \tag{2}$$

where $\widehat{J}^s_{t,\mathrm{root}}$ is the canonical root position of the predicted hand model. This locks the hand's global coordinate to the physical arm chain $B^s_{t,\mathrm{wrist}}$ while preserving local finger articulation. The resulting upper-limb skeletons are causally interpolated and rasterized into both exocentric and egocentric views as skeletal guidance maps (Q^e_t, Q^g_t) using calibrated camera parameters $\mathcal{C}$, with joint confidence scores modulating stroke intensities to suppress hallucinated joints.

3.2 Bidirectional Teacher

With the streaming dual-branch priors established, we construct the teacher model with bidirectional attention to master Exo2Ego (Liu et al., 2024; Luo et al., 2024; Kang et al., 2026; Mahdi et al., 2025; Xu et al., 2025; Wang et al., 2026; Mahdi et al., 2026). As illustrated in Fig. 2**b**, all visual inputs and condition priors are encoded into latent representations via a frozen video VAE. Following EgoX (Kang et al., 2026), we concatenate the streaming priors with video latents along the channel dimension for pixel-aligned conditioning, while concatenating the clean exocentric video and the noisy egocentric latent along the width dimension. The composite latents are then fed into a pretrained Wan backbone. The model is trained using LoRA (Hu et al., 2022) with our proposed region-aware training curriculum.

Region-aware loss weight scheduling. Uniform supervision across the egocentric canvas is suboptimal due to disparate learning difficulties across functional domains: $\mathcal{S}$ offers stable visual anchors, $\mathcal{U}$ requires plausible completion vulnerable to overfitting, and $\mathcal{H}$ exhibits complex dynamics that destabilize early training. We therefore propose a region-aware curriculum (Fig. 2**b**): first anchoring cross-view alignment in $\mathcal{S}$, gradually introducing completion in $\mathcal{U}$, and finally ramping up supervision in $\mathcal{H}$ to refine manipulation fidelity.

Let $r_{\mathcal{S}}(p), r_{\mathcal{U}}(p), r_{\mathcal{H}}(p) \in \{0, 1\}$ denote binary regional indicators on the egocentric latent grid Ω_g. At training progress $u \in [0, 1]$, the unnormalized spatial weight at position p is linearly combined as

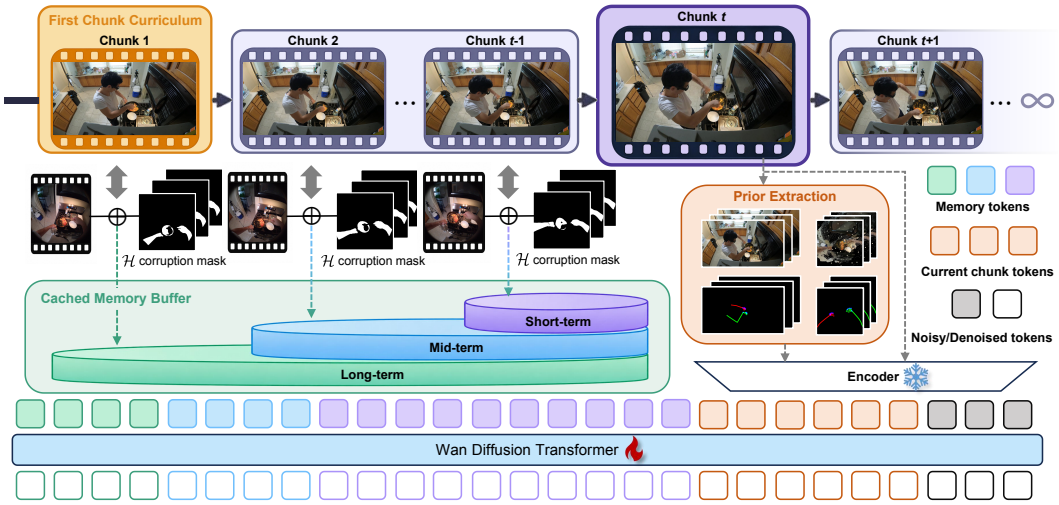

Figure 3: Training pipeline for the streaming student model. The k-th egocentric chunk is denoised by the Wan DiT (Team Wan et al., 2025) conditioned on its dual-branch priors and the memory buffer. We apply the proposed region-aware corruption and first-chunk curriculum during training.

$a_u(p) = \sum_{r\in\{\mathcal{S},\mathcal{U},\mathcal{H}\}} w_r(u) r_r(p)$, where $w_r(u)$ follows the region-aware curriculum. To stabilize gradient scales, we normalize $a_u(p)$ across the canvas into $\omega_u(p) = a_u(p)/(\frac{1}{|\Omega_g|}\sum_{q\in\Omega_g} a_u(q) + \delta)$ ($\delta > 0$). Denoting the flow-matching prediction error as $e_\sigma(p)$, our training objective is:

$$\mathcal{L} = \mathbb{E}\left[\frac{1}{|\Omega_g|}\sum_{p\in\Omega_g} \omega_u(p) e_\sigma(p)\right]. \tag{3}$$

3.3 Real-Time Autoregression

To enable streaming from our bidirectional teacher, we adapt the framework into a causal, chunk-wise autoregressive student distilled for few-step inference. As illustrated in Fig. 3, the video stream is processed chunk by chunk; the k-th chunk (which denotes the current chunk) is denoised via intra-chunk bidirectional attention conditioned on current-chunk priors (Sec. 3.1) and a historical memory buffer $\mathcal{M}_k$ which maintains exclusively clean egocentric latents from past chunks and compresses distant context via multi-term memory patchification (Yuan et al., 2026). To address Exo2Ego task-specific challenges in autoregressive training, we introduce two dedicated strategies: *region-aware corruption* to mitigate exposure bias and motion drift in $\mathcal{H}$, and a *first-chunk curriculum* to handle the initial chunk without historical conditioning, improving the student's rollout robustness in $\mathcal{U}$.

Region-aware corruption. The exposure bias induced by teacher forcing is highly uneven across regions: $\mathcal{S}$ is continuously anchored through the proposed streaming dynamic ego prior, resilient to drift. By contrast, the dynamic and complex nature of region $\mathcal{H}$ makes it susceptible to self-occlusions and motion blur, under which the streaming priors can degrade. In such unconstrained cases, synthesis in $\mathcal{H}$ refers to historical context and thus prone to error accumulation.

We inject stochastic perturbations into $\mathcal{H}$ on historical latents $z_t^g \in \mathcal{M}_k$:

$$\tilde{z}_t^g = z_t^g + b_t m_t^H \odot \epsilon_t, \tag{4}$$

where m_t^H is the binary $\mathcal{H}$ mask, $b_t \sim \text{Bernoulli}(p)$ acts as a stochastic trigger, and zero-mean Gaussian noise ϵ_t injects additive perturbation. This mechanism forces the student to ground hand dynamics in current condition priors rather than over-relying on potentially degraded history.

First-chunk curriculum. Because region $\mathcal{U}$ lacks direct exocentric observations, subsequent chunks rely entirely on previously synthesized ego frames for consistency. Thus, rollout quality in $\mathcal{U}$ heavily depends on the cold-start synthesis of the first chunk ($k = 1, \mathcal{M}_1 = \emptyset$). Any initial geometric artifacts in region $\mathcal{U}$ propagate downstream as erroneous history, triggering degradations.

We therefore introduce a progressive training curriculum on chunk sampling. We initially train exclusively on the first chunk to anchor high-fidelity inpainting from visual priors alone. As training proceeds, its sampling probability decays linearly, shifting focus toward continuation chunks. This design ensures robust visual initialization before exposing the model to autoregressive training.

Few-step distillation. We adapt the few-step Distribution Matching Distillation from Helios (Yuan et al., 2026) to distill our causal teacher G_θ into a few-step student generator G_ϕ:

$$\begin{aligned} \mathcal{L}_{\text{student}} &= \mathcal{L}_{\text{DMD}}(G_\phi; \mu_{\text{real}}, \mu_{\text{fake}}) + \lambda_G \mathcal{L}^G_{\text{adv}}(G_\phi, D_\psi), \\ \mathcal{L}_{\text{critic}} &= \mathcal{L}_{\text{reg}}(D_\psi) + \lambda_D \mathcal{L}^D_{\text{adv}}(G_\phi, D_\psi) + \lambda_R \mathcal{L}_{\text{R1}}(D_\psi), \end{aligned} \tag{5}$$

where $\mathcal{L}_{\text{DMD}}$ aligns the generative distribution, and the adversarial objective with discriminator D_ψ sharpens fine details. We refer the readers to the Appendix for more details.

4 Experiments

4.1 Implementation details

Training. We train both our bidirectional teacher and autoregressive student on 4,400 curated clips from Ego-Exo4D (Grauman et al., 2024), with each clip spanning 5 seconds at 24 FPS. Exocentric inputs are set to a resolution of 832×480, while generated egocentric outputs are 480×480. The bidirectional teacher (Sec. 3.2) is initialized from Wan2.2-TI2V-5B (Team Wan et al., 2025), and the student model (Sec. 3.3) is initialized directly from this trained teacher. Detailed data curation procedures and training configurations are provided in the Appendix.

Evaluation data. We evaluate quantitatively on Ego-Exo4D (Grauman et al., 2024) using both the Unseen (unseen scene during training) and New Action (seen scene with unseen actions) splits as well as H2O (Kwon et al., 2021) (with ground truth calibrated to the FOV of Grauman et al. (2024)). We evaluate only qualitatively on Assembly101 (Sener et al., 2022), as its egocentric views provide only monochromatic ground truth within a single indoor scene.

Evaluation metrics. We evaluate performance across four groups: (1) *Image metrics*: $\mathcal{S}$-PSNR, $\mathcal{S}$-MSSIM (Wang et al., 2003), $\mathcal{S}$-LPIPS measure fidelity in $\mathcal{S}$, $\mathcal{U}$-CLIP-I measure naturalness and visual plausibility of $\mathcal{U}$ and PCK@0.05 (Yang & Ramanan, 2012) and Motion-Consistency (MC) measures hand position accuracy and motion accuracy in $\mathcal{H}$. Full-image PSNR and SSIM are also provided for reference. (2) *Video metrics*: FVD evaluate distribution discrepancy and inter-frame temporal continuity. Reference-free VBench (Zhang et al., 2025a) dimensions temporal flickering (TF) and motion smoothness (MS) evaluates temporal quality and motion quality. (3) *Object metrics*: Following the matching pipeline in EgoX (Kang et al., 2026) using SAM 2 and DINOv3, we report center location error and bounding-box IoU. Additionally, we measure generation efficiency by inference FPS. (4) *Efficiency*: Inference FPS.

Baselines. We compare against Exo2Ego baselines, including EgoX (Kang et al., 2026) and Exo2Ego-V (Liu et al., 2024), alongside representative camera-control models: TrajCrafter (Yu et al., 2025), Wan-VACE (Jiang et al., 2025), and Vista4D (Lin et al., 2026). For a fair comparison, all baselines are fine-tuned on our training dataset. Other related works (Xu et al., 2025; Mahdi et al., 2025; Wang et al., 2026; Mahdi et al., 2026) are omitted since they are not open-source.

4.2 Qualitative and quantitative comparison

4.2.1 Short video

For a fair evaluation, Ours (teacher), Ours (student), and all baseline methods are compared on short video clips truncated to identical temporal durations in Table. 1. In addition, Ours (long) reports scores computed from our student model running for 30-second continuous rollouts initialized from the same starting timestamps across the test splits.

As shown in Table 1 and Figure 4, StreamEgo achieves superior performance across quantitative metrics and visual comparisons.: (1) Across image- and object-level metrics, our framework aligns closely with the task objectives of Exo2Ego: establishing high visual fidelity in region $\mathcal{S}$, maintaining plausible naturalness and diversity region $\mathcal{U}$, and ensuring precise hand articulation and motion

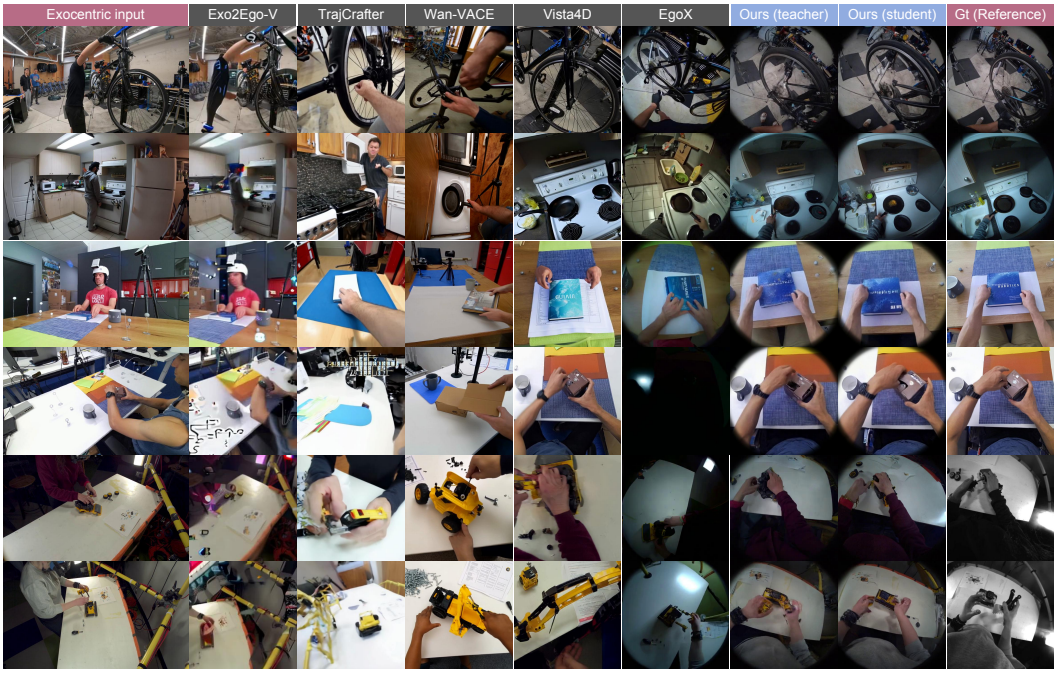

Figure 4: Visual comparison from Grauman et al. (2024) ((UNSEEN)), Kwon et al. (2021), and Sener et al. (2022) on synchronized key frames. Please refer to our supplementary video.

Table 1: Comparison on the NEW ACTION and UNSEEN from Ego-Exo4D (Grauman et al., 2024) and H2O (Kwon et al., 2021). **Bold**: the best. Underline: second best.

	Method	Image Metrics								Object Metrics		Video Metrics			Efficiency
		$\mathcal{S}$-PSNR ↑	$\mathcal{S}$-MSSIM ↑	$\mathcal{S}$-LPIPS ↓	$\mathcal{U}$-CLIP-I ↑	PCK@0.05 ↑	MC ↑	PSNR ↑	SSIM ↑	Loc Err ↓	IoU ↑	FVD ↓	TF ↑	MS ↑	FPS ↑
NEW ACTION (new actions, seen envs.)	Exo2Ego-V	13.36	0.384	0.569	0.774	14.86%	16.93%	14.53	0.381	0.326	0.074	622.47	0.960	0.966	3.92
	TrajCrafter	10.91	0.347	0.617	0.783	3.60%	16.79%	12.55	0.349	0.232	0.116	568.92	0.960	0.980	0.30
	Wan-VACE	10.42	0.383	0.624	0.817	5.83%	18.37%	13.23	0.430	0.230	0.120	505.89	**0.989**	**0.994**	0.96
	Vista4D	12.07	0.412	0.465	0.801	11.58%	15.92%	13.42	0.435	0.135	0.249	353.83	0.983	0.992	0.07
	EgoX	15.75	0.489	0.580	0.888	25.91%	34.84%	16.06	0.561	0.130	0.357	184.24	0.977	0.990	0.08
	Ours (teacher)	**18.35**	**0.603**	0.372	**0.907**	**49.86%**	**56.26%**	**17.51**	**0.603**	**0.076**	**0.478**	**111.97**	0.983	**0.994**	0.97
	Ours (student)	18.12	0.599	**0.366**	0.891	47.61%	53.12%	17.45	0.593	0.099	0.462	128.52	0.984	0.993	**24.3**
	Ours (long)	18.16	0.588	0.365	0.893	47.59%	53.48%	17.47	0.588	0.118	0.447	135.62	0.981	0.991	**24.3**
UNSEEN (new actions, new envs.)	Exo2Ego-V	13.25	0.439	0.597	0.679	11.24%	13.80%	12.70	0.418	0.447	0.003	1283.50	0.971	0.976	3.92
	TrajCrafter	10.73	0.295	0.607	0.782	2.75%	12.65%	12.31	0.306	0.386	0.041	818.63	0.966	0.984	0.30
	Wan-VACE	10.04	0.352	0.649	0.825	5.41%	17.22%	12.21	0.378	0.412	0.038	1023.96	**0.995**	**0.996**	0.96
	Vista4D	11.58	0.401	0.576	0.840	9.37%	12.08%	12.64	0.435	0.144	0.230	783.54	0.990	0.994	0.07
	EgoX	13.36	0.406	0.577	0.879	11.07%	15.41%	14.41	0.456	0.311	0.092	438.79	0.981	0.992	0.08
	Ours (teacher)	**18.18**	0.576	0.405	**0.901**	**40.24%**	**43.76%**	**17.19**	0.544	**0.106**	**0.456**	138.91	0.980	0.995	0.97
	Ours (student)	18.04	**0.583**	0.423	0.894	37.49%	41.13%	17.08	**0.548**	0.113	0.436	**136.48**	0.986	0.990	**24.3**
	Ours (long)	17.88	0.569	**0.394**	0.890	37.22%	41.70%	16.84	0.537	0.119	0.427	144.33	0.988	0.994	**24.3**
Kwon et al. (2021) (in-the-wild.)	Exo2Ego-V	12.18	0.404	0.510	0.666	12.05%	15.21%	12.20	0.395	0.452	0.019	861.67	0.973	0.971	3.92
	TrajCrafter	12.98	0.379	0.465	0.774	20.60%	62.96%	13.15	0.391	0.362	0.163	724.07	0.961	0.980	0.30
	Wan-VACE	14.78	0.348	0.322	0.804	27.35%	68.66%	14.65	0.376	0.310	0.183	662.87	**0.995**	**0.996**	0.96
	Vista4D	18.39	0.644	0.181	0.916	68.57%	52.99%	18.31	0.640	0.072	0.741	161.45	0.990	0.994	0.07
	EgoX	14.23	0.465	0.377	0.710	35.10%	40.28%	15.00	0.468	0.205	0.513	479.52	0.911	0.929	0.08
	Ours (teacher)	**23.59**	**0.724**	**0.106**	**0.931**	**89.22%**	**95.78%**	**21.35**	**0.688**	**0.041**	**0.876**	**89.94**	0.993	0.994	0.97
	Ours (student)	23.42	0.713	0.118	0.926	84.40%	93.14%	20.92	0.680	0.045	0.859	93.47	0.993	0.994	**24.3**
	Ours (long)	23.21	0.713	0.118	0.929	81.80%	89.71%	21.05	0.680	0.056	0.814	94.17	0.993	0.994	**24.3**

consistency in region $\mathcal{H}$. (2) As evidenced by both the metrics (PCK@0.05 and MC in Table. 1) and visual results, StreamEgo substantially outperforms baseline methods in hand-motion accuracy, thanks to the proposed streaming dual-branch prior providing appearance and structural guidance for region $\mathcal{H}$. (3) Camera control models (Yu et al., 2025; Jiang et al., 2025; Lin et al., 2026) largely fail due to their underlying assumption of modest viewpoint changes. Besides, existing Exo2Ego methods like Kang et al. (2026) suffer from viewpoint rotational drifts and degrade severely in out-of-distribution cases. In contrast, our method demonstrates markedly superior cross-domain generalization on in-the-wild datasets (Kwon et al., 2021; Sener et al., 2022), validating the effectiveness of our region-aware curriculum in coordinating multi-region optimization. (4) In terms of efficiency, our distilled student model achieves real-time streaming speeds while maintaining competitive quality against the teacher model, underscoring its potential for practical applications.



Figure 5: Streaming long-video generation with the autoregressive student on Grauman et al. (2024) (UNSEEN) and Kwon et al. (2021). Please refer to our supplementary video for better viewing.

4.2.2 LONG VIDEO

As shown in Fig. 5, our student operates as a causal streaming model, taking input chunk-by-chunk and generating the corresponding egocentric chunk with only a single-chunk warm-up latency (*i.e.*, emitting the first egocentric chunk as the second exocentric chunk arrives. With chunk size 33 frames, this corresponds to roughly 1.375 seconds.). The generated long videos demonstrate sustained visual quality, smooth temporal coherence, and minimal flicker, remaining highly comparable to the teacher model in both viewpoint consistency and hand-motion accuracy (Table 1: Ours (Long)). Moreover, the model generalizes robustly to in-the-wild recordings such as H2O (Kwon et al., 2021). Furthermore, generation quality remains stable throughout the rollouts. This resilience stems from the strong initial transfer inherited from the bidirectional teacher, combined with continuous, chunk-wise pixel-aligned anchors in region $\mathcal{S}$, while our first-chunk curriculum and region-aware corruption effectively mitigate exposure bias across regions $\mathcal{U}$ and $\mathcal{H}$.

4.3 ABLATION STUDIES

As summarized in Table 2, we evaluate key components against the full teacher and student models. For the teacher, we isolate the loss curriculum (*w/o* HSU) and condition priors: removing the streaming human pose prior (*w/o pose*), replacing our body-guided pose prior with isolated hand poses (Park et al., 2026) (Ew pose) or SAM-based hand-and-object predictions (Xu et al., 2025) (HOI pose), and removing the streaming dynamic prior (*w/o* prior), substituting our streaming dynamic prior with EgoX's offline ego prior (E prior) lacking foreground segmentation. For the student, we ablate region-aware history corruption (*w/o* H corruption) and the first-chunk curriculum (*w/o* first chunk) to verify their roles in stable autoregressive rollout.

Human pose prior. Removing the streaming human pose prior leads to position inaccuracy and motion degradation in $\mathcal{H}$ (PCK@0.05 and MC in Table. 2; missing left hand in Fig. 6: *w/o* pose), which can not be recovered by adopting isolated hand pose estimation (Park et al., 2026) with inaccurate projected hand positions (Fig. 6: Ew pose). Substituting our prior with HOI mask predictions (Xu et al., 2025) produces blurred manipulations since it can not define the structure of $\mathcal{H}$.

Dynamic ego prior. Removing the dynamic ego prior eliminates pixel-aligned appearance anchors, causing severe visual degradation in $\mathcal{S}$ and noticeable viewpoint drift. Replacing it with EgoX's prior removes appearance guidance in $\mathcal{H}$, leading to degraded hand quality (Fig. 6: E prior).

Loss scheduling. Replacing our curriculum with fixed uniform weights leads to inaccurate motion in $\mathcal{H}$, alongside degradations in $\mathcal{S}$ and $\mathcal{U}$ (Fig. 6: blurry hands, distorted plate with unnatural lighting).

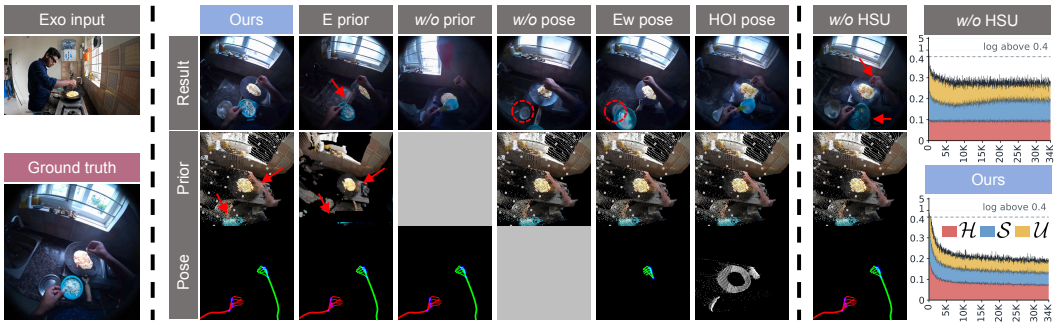

Figure 6: Ablation comparison on Grauman et al. (2024). **Left:** exocentric input and its egocentric ground truth. **Middle:** ablations results for the streaming dual-branch prior. **Right:** ablation results and training log for the region-aware curriculum. Red circles and arrows mark failure regions.

Table 2: Ablation study on Ego-Exo4D (Grauman et al., 2024). Teacher variants isolate conditioning and loss scheduling; student variants isolate autoregressive training strategies. **Bold: The best**.

Variant	Image Metrics						Object Metrics		Video Metrics		
	$\mathcal{S}$-PSNR ↑	$\mathcal{S}$-SSIM ↑	$\mathcal{S}$-LPIPS ↓	$\mathcal{U}$-CLIP-I ↑	PCK@0.05 ↑	MC ↑	Loc Err ↓	IoU ↑	FVD ↓	TF ↑	MS ↑
Full teacher	**18.18**	**0.576**	**0.405**	**0.901**	**40.24%**	**43.76%**	**0.106**	**0.456**	**138.91**	0.980	**0.995**
w/o pose	18.10	0.564	0.430	0.893	16.75%	19.27%	0.110	0.448	284.59	0.978	0.991
Ew pose	18.14	0.572	0.422	0.887	18.03%	29.64%	0.115	0.452	289.04	0.980	0.988
HOI pose	17.61	0.558	0.444	0.880	17.59%	24.80%	0.172	0.341	275.50	0.968	0.985
w/o prior	13.57	0.396	0.588	0.714	15.71%	18.35%	0.255	0.126	645.42	0.959	0.962
E prior	18.05	0.546	0.411	0.890	36.16%	40.38%	0.198	0.366	144.45	0.980	0.991
w/o HSU	16.88	0.481	0.514	0.864	29.22%	31.37%	0.222	0.139	301.76	0.977	0.989
Full student	**17.88**	**0.569**	**0.394**	**0.890**	**37.22%**	**41.70%**	**0.119**	**0.427**	**144.33**	**0.988**	**0.994**
w/o H corruption	17.84	0.560	0.412	0.875	31.27%	35.84%	0.142	0.400	203.48	0.984	0.992
w/o first-chunk	17.32	0.557	0.395	0.837	36.90%	38.73%	0.294	0.273	397.10	0.981	0.987

The loss curves (Fig. 6:**right**) also reveals that *w/o* HSU hinders convergence in $\mathcal{H}$ and $\mathcal{S}$ (*e.g.* lower $\mathcal{S}$-PSNR and MC in Table. 2) while potentially encourages overfitting in $\mathcal{U}$. In contrast, our curriculum stabilize training and enhance generalization.

Region-aware corruption. Disabling the region-aware corruption leaves $\mathcal{H}$ vulnerable to exposure bias between teacher-forced training and free-running rollout. As is shown in Fig. 7, the hand position drifts, leading to degradation to $\mathcal{H}$-region image metrics in Table. 2.

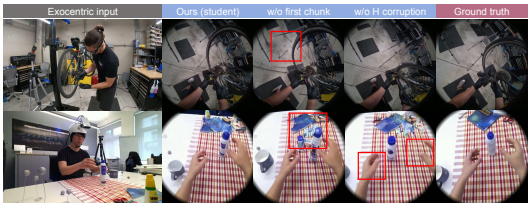

Figure 7: Ablation comparisons for region-aware corruption and first-chunk curriculum.

First-chunk curriculum. Removing the first-chunk curriculum degrades yields blur and distortion in $\mathcal{U}$ (Fig. 7: distorted book and bottles; $\mathcal{U}$-CLIP-I in Table. 6). This stems from two interrelated factors: synthesizing the initial chunk is inherently harder without historical reference, and any resulting artifacts may propagate, compounding errors across subsequent chunks.

5 Conclusion

In this work, we presented StreamEgo, a streaming framework enabling unbounded and generalizable exocentric-to-egocentric video generation. To overcome the degraded interactions, overfitting, and fixed-length latency bottlenecks of existing approaches, we proposed three core techniques: streaming dual-branch priors for structurally guided hand-object synthesis, a region-aware curriculum that stabilizes training across functional canvas domains, and a chunk-wise autoregressive adaptation with few-step distillation for causal rollout, together with a region-aware corruption mechanism and a first-chunk curriculum for enhancing quality. Achieving generalizable streaming generation at 24.3 FPS on a single H100 GPU, StreamEgo provides a practical pathway to repurpose vast third-person demonstration recordings into first-person trajectories for embodied robotics.

AI USE STATEMENT

In this work, we used generative AI tools only for the recommended disclosure task of editing the research paper to improve readability and polish writing. We have not used generative AI tools for any tasks with required disclosure, and all other recommended disclosure tasks are not applicable to this work. We have reviewed all AI-assisted edits. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Sherwin Bahmani, Tianchang Shen, Jiawei Ren, Jiahui Huang, Yifeng Jiang, Haithem Turki, Andrea Tagliasacchi, David B. Lindell, Zan Gojcic, Sanja Fidler, Huan Ling, Jun Gao, and Xuanchi Ren. Lyra: Generative 3D scene reconstruction via self-distillation with video diffusion models. In *International Conference on Learning Representations (ICLR)*, 2026.

Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, and Di Zhang. ReCamMaster: Camera-controlled generative rendering from a single video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14834–14844, 2025a.

Jianhong Bai, Menghan Xia, Xintao Wang, Ziyang Yuan, Xiao Fu, Zuozhu Liu, Haoji Hu, Pengfei Wan, and Di Zhang. SynCamMaster: Synchronizing multi-camera video generation from diverse viewpoints. In *International Conference on Learning Representations (ICLR)*, 2025b.

Homanga Bharadhwaj, Debidatta Dwibedi, Abhinav Gupta, Shubham Tulsiani, Carl Doersch, Ted Xiao, Dhruv Shah, Fei Xia, Dorsa Sadigh, and Sean Kirmani. Gen2act: Human video generation in novel scenarios enables generalizable robot manipulation. In *Conference on Robot Learning (CoRL)*, 2024.

Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Yanjun Wang, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, et al. SMPLer-X: Scaling up expressive human pose and shape estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 11454–11468, 2023.

Chenjie Cao, Jingkai Zhou, Shikai Li, Jingyun Liang, Chaohui Yu, Fan Wang, Xiangyang Xue, and Yanwei Fu. Uni3C: Unifying precisely 3D-enhanced camera and human motion controls for video generation. In *SIGGRAPH Asia Conference Papers*, 2025.

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

Yilun Du, Mengjiao Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Joshua B. Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Yao Feng, Hengkai Tan, Xinyi Mao, Chendong Xiang, Guodong Liu, Shuhe Huang, Hang Su, and Jun Zhu. Vidar: Embodied video diffusion model for generalist manipulation, 2025. arXiv:2507.12898.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-Exo4D: Understanding skilled human activity from first- and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19383–19400, 2024.

Yuchao Gu, Weijia Mao, and Mike Zheng Shou. Long-context autoregressive video modeling with next-frame prediction, 2025. arXiv:2503.19325.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

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

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. VACE: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17191–17202, 2025.

Taewoong Kang, Kinam Kim, Dohyeon Kim, Minho Park, Junha Hyung, and Jaegul Choo. EgoX: Egocentric video generation from a single exocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Jihwan Kim, Junoh Kang, Jinyoung Choi, and Bohyung Han. FIFO-Diffusion: Generating infinite videos from text without training. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 89834–89868, 2024.

Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2O: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10118–10128, 2021.

Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Luyao Zhang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Boyu Zhou, Xing Zhu, Mingyu Ding, Yujun Shen, and Yinghao Xu. Causal world modeling for robot control. In *Robotics: Science and Systems (RSS)*, 2026a.

Ruilong Li, Brent Yi, Junchen Liu, Hang Gao, Yi Ma, and Angjoo Kanazawa. Cameras as relative positional encoding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025a.

Shuang Li, Yihuai Gao, Dorsa Sadigh, and Shuran Song. Unified video action model. In *Robotics: Science and Systems (RSS)*, 2025b. doi: 10.15607/RSS.2025.XXI.074.

Wuyang Li, Wentao Pan, Po-Chien Luan, Yang Gao, and Alexandre Alahi. Stable video infinity: Infinite-length video generation with error recycling. In *International Conference on Learning Representations (ICLR)*, 2025c.

Zhixuan Li, Sai Kumar Dwivedi, Filip Maric, Carlos Chacon, Nadine Bertsch, Filippo Arcadu, Tomas Hodan, Michael Ramamonjisoa, Peter Wonka, Akash Zhao, et al. EgoPoseFormer v2: Accurate egocentric human motion estimation for AR/VR. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026b.

Kuan Heng Lin, Zhizheng Liu, Pablo Salamanca, Yash Kant, Ryan Burgert, Yuancheng Xu, Koichi Namekata, Yiwei Zhao, Bolei Zhou, Micah Goldblum, Paul Debevec, and Ning Yu. Vista4D: Video reshooting with 4D point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 32671–32682, 2026.

Jia-Wei Liu, Weijia Mao, Zhongcong Xu, Jussi Keppo, and Mike Zheng Shou. Exocentric-to-egocentric video generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 136149–136172, 2024.

Jinlai Liu, Jian Han, Bin Yan, Hui Wu, Fengda Zhu, Xing Wang, Yi Jiang, Bingyue Peng, and Zehuan Yuan. InfinityStar: Unified spacetime AutoRegressive modeling for visual generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025a.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. In *International Conference on Learning Representations (ICLR)*, 2025b.

Mi Luo, Zihui Xue, Alex Dimakis, and Kristen Grauman. Put myself in your shoes: Lifting the egocentric perspective from exocentric videos. In *European Conference on Computer Vision (ECCV)*, pp. 407–425. Springer, 2024.

Mohammad Mahdi, Yuqian Fu, Nedko Savov, Jiancheng Pan, Danda Pani Paudel, and Luc Van Gool. Exo2EgoSyn: Unlocking foundation video generation models for exocentric-to-egocentric video synthesis, 2025. arXiv:2511.20186.

Mohammad Mahdi, Nedko Savov, Danda Pani Paudel, and Luc Van Gool. From synchrony to sequence: Exo-to-ego generation via interpolation. In *European Conference on Computer Vision (ECCV)*, volume 17016 of *Lecture Notes in Computer Science*, pp. 134–150. Springer Nature Switzerland, 2026. doi: 10.1007/978-3-032-37383-0_8.

Meta AI. SAM 3.1: Object multiplex. `https://github.com/facebookresearch/sam3/blob/main/RELEASE_SAM3p1.md`, 2026. Release notes.

Junho Park, Andrew Sangwoo Ye, and Taein Kwon. EgoWorld: Translating exocentric view to egocentric view using rich exocentric observations. In *International Conference on Learning Representations (ICLR)*, 2026.

Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. WiLoR: End-to-end 3D hand localization and reconstruction in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12242–12254, 2025.

Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. FreeNoise: Tuning-free longer video diffusion via noise rescheduling. In *International Conference on Learning Representations (ICLR)*, 2024.

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. GEN3C: 3D-informed world-consistent video generation with precise camera control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6121–6132, 2025.

David Ruhe, Jonathan Heek, Tim Salimans, and Emiel Hoogeboom. Rolling diffusion models. In *International Conference on Machine Learning (ICML)*, volume 235, pp. 42818–42835, 2024.

Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21064–21074, 2022.

Zhihao Shi, Kejia Yin, Weilin Wan, Yuhongze Zhou, Yuanhao Yu, Xinxin Zuo, Qiang Sun, and Juwei Lu. CamDirector: Towards long-term coherent video trajectory editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 32683–32692, 2026.

Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion. In *International Conference on Machine Learning (ICML)*, 2025.

Mingzhen Sun, Weining Wang, Gen Li, Jiawei Liu, Jiahui Sun, Wanquan Feng, Shanshan Lao, SiYu Zhou, Qian He, and Jing Liu. AR-Diffusion: Asynchronous video generation with auto-regressive diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7364–7373, 2025.

Team Wan et al. Wan: Open and advanced large-scale video generative models, 2025. arXiv:2503.20314.

Danny Tran, Roberto Martín-Martín, and Kristen Grauman. EgoExo-WM: Unlocking exo video for ego world models, 2026. arXiv:2605.15477.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. VideoMAE v2: Scaling video masked autoencoders with dual masking. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14549–14560, 2023.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. CUT3R: Continuous 3D perception model with persistent state. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10510–10522, 2025a.

Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. MoGe-2: Accurate monocular geometry with metric scale and sharp details. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 40377–40408, 2025b.

Shengze Wang, Michael Stengel, Tianye Li, Seonwook Park, Amrita Mazumdar, Koki Nagano, Alex Trevithick, and Shalini De Mello. Grounded-Exo2Ego: Structured semantic grounding for robust exocentric-to-egocentric video generation, 2026. arXiv:2608.20534.

Z. Wang, E.P. Simoncelli, and A.C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pp. 1398–1402, 2003. doi: 10.1109/ACSSC.2003.1292216.

Jilan Xu, Yifei Huang, Baoqi Pei, Junlin Hou, Qingqiu Li, Guo Chen, Yuejie Zhang, Rui Feng, and Weidi Xie. EgoExo-Gen: Ego-centric video prediction by watching exo-centric videos. In *International Conference on Learning Representations (ICLR)*, pp. 14694–14711, 2025.

Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 38571–38584, 2022.

Chenhongyi Yang, Anastasia Tkach, Shreyas Hampali, Linguang Zhang, Elliot J. Crowley, and Cem Keskin. EgoPoseFormer: A simple baseline for stereo egocentric 3D human pose estimation. In *European Conference on Computer Vision (ECCV)*, 2024.

Sherry Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Leslie Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. In *International Conference on Learning Representations (ICLR)*, 2023.

Yi Yang and Deva Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2878–2890, 2012.

Tianwei Yin, Michael Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T. Freeman. Improved distribution matching distillation for fast image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 47455–47487, 2024a.

Tianwei Yin, Michael Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6613–6623, 2024b.

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

Mark Yu, Wenbo Hu, Jinbo Xing, and Ying Shan. TrajectoryCrafter: Redirecting camera trajectory for monocular videos via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 100–111, 2025.

Shenghai Yuan, Yuanyang Yin, Zongjian Li, Xinwei Huang, Xiao Yang, and Li Yuan. Helios: Real real-time long video generation model, 2026. arXiv:2603.04379.

Cheng Zhang, Boying Li, Meng Wei, Yan-Pei Cao, Camilo Cruz Gambardella, Dinh Phung, and Jianfei Cai. Unified camera positional encoding for controlled video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 38027–38037, 2026.

Fan Zhang, Shulin Tian, Ziqi Huang, Yu Qiao, and Ziwei Liu. Evaluation agent: Efficient and promptable evaluation framework for visual generative models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 7561–7582, 2025a.

Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025b.

Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. UniPC: A unified predictor-corrector framework for fast sampling of diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 49842–49869, 2023.

Siyuan Zhou, Yilun Du, Jiaben Chen, Yandong Li, Dit-Yan Yeung, and Chuang Gan. RoboDreamer: Learning compositional world models for robot imagination. In *International Conference on Machine Learning (ICML)*, volume 235, pp. 61885–61896, 2024.

Chuning Zhu, Raymond Yu, Siyuan Feng, Benjamin Burchfiel, Paarth Shah, and Abhishek Gupta. Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets. In *Robotics: Science and Systems (RSS)*, 2025.

Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. In *International Conference on Machine Learning (ICML)*, 2026.

A Data Curation

A.1 Ego-Exo4D training corpus

Ego-Exo4D (Grauman et al., 2024) supplies synchronized calibrated exocentric (GoPro) and first-person (Aria) RGB streams and is the sole source of our training data. We extract candidate windows of five seconds at 24 fps (121 frames, endpoints inclusive) on a one-second stride from manipulation-oriented activities, *e.g.* cooking, bicycle repair, medical procedures, and musical performance, rotating the Aria stream to its upright orientation before export.

Selection proceeds from inexpensive metadata checks to visual checks. We discard takes with missing calibration or trajectory data, and retain windows whose mean exo-to-ego distance is at most 3 m, whose ego camera lies inside the exocentric frustum in at least 85% of frames, and whose ego trajectory is sufficiently stable (quality score ≥ 0.7, angular-velocity standard deviation ≤ 0.25 rad/s, mean linear velocity ≤ 0.4 m/s). The visual screen then requires an actor that is large and central (person height $\geq 45\%$ of the image in $\geq 70\%$ of sampled frames), at least one visible hand in 80% of sampled frames, adequate Laplacian sharpness in both views, and bounded optical-flow magnitude. Surviving clips are ranked by quality score; non-maximum suppression within each take and exocentric camera enforces a 1.5 s minimum separation, and per-take caps and activity quotas further suppress near-duplicates while broadening activity coverage. A final captioning pass records the action and flags poor subject orientation, hidden hands, and severe exocentric occlusion. The resulting corpus contains $4{,}400$ training clips and a 100-clip validation set.

A.2 Evaluation splits

Exo2Ego models (Kang et al., 2026; Mahdi et al., 2025; Wang et al., 2026; Mahdi et al., 2026) must generalize rather than memorize their paired training data. Our evaluation is therefore designed to be *hard by construction*: from the same source inventory we construct two 500-clip splits that are disjoint from training along the two axes that a paired dataset makes easiest to exploit.

New Action. These clips depict action categories *absent* from the training corpus, but are recorded in environments (scenes, rooms, and capture setups) that the model has seen. They isolate whether the model has learned to translate a manipulation from the exocentric view, as opposed to recalling how a familiar action looked in a familiar room. Because region $\mathcal{U}$ in these clips can often be completed from memorized scene context, this split principally stresses fidelity in $\mathcal{H}$.

Unseen. These clips are recorded in environments that never occur in training, and their actions are likewise held out. No part of the egocentric canvas, including the currently unobserved region $\mathcal{U}$, can be recovered by recalling a training scene, so the model must infer $\mathcal{U}$ from the exocentric evidence alone. This split therefore penalizes benchmark memorization.

Both splits are constructed at the level of capture takes, so no clip shares a take, scene, or subject with the training set.

A.3 H2O: an in-the-wild evaluation set

Quantitative evaluation on Ego-Exo4D (Grauman et al., 2024), even on our held-out splits, remains within the distribution of a single capture protocol. We therefore additionally evaluate on H2O (Kwon et al., 2021), an in-the-wild dataset collected with different hardware, framing, and subject population. To keep the evaluation broad rather than convenient, we deliberately maximize coverage along three axes rather than selecting a small homogeneous subset: we retain the widest available range of exocentric camera viewpoints, participant identities, and indoor scene types.

Concretely, the screening of H2O proceeds as follows. We require synchronized exocentric and egocentric streams at a native resolution of 1280×720 with per-camera intrinsics and a complete egocentric pose trajectory covering at least 90% of the sequence, discarding any sequence for which a camera stream or its calibration is incomplete. We then apply a visual screen requiring adequate sharpness in both views and non-trivial inter-frame motion, which removes static or over-blurred windows, followed by the same temporal-NMS and diversity machinery used for Ego-Exo4D, with per-sequence caps on the number of clips drawn from any single camera or any single sequence. Crucially, the caps are set high enough that all four exocentric viewpoints and all six tabletop scene

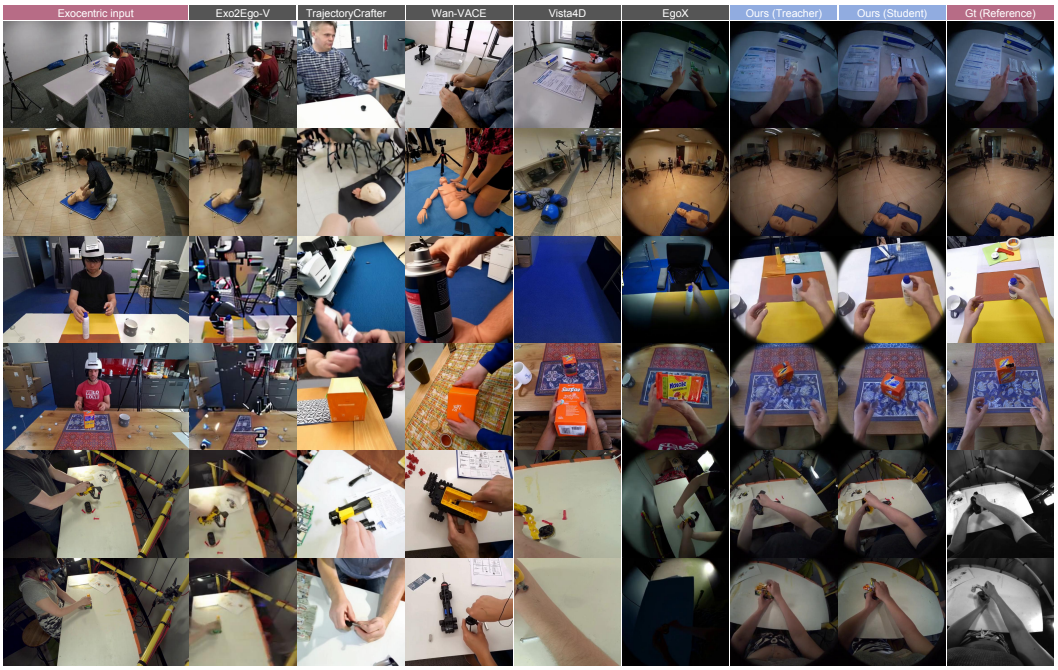

Figure 8: Additional qualitative comparison on synchronized key frames from Ego-Exo4D (UNSEEN) (Grauman et al., 2024), H2O (Kwon et al., 2021), and Assembly101 (Sener et al., 2022).

configurations remain represented in the final set, and the retained clips span every subject group and dozens of distinct action labels. We deliberately avoid a per-action cap so that rarer manipulations are not truncated away, which yields a many-action evaluation rather than a convenient few-class one.

Because H2O is not natively compatible with our setting, we apply two calibration steps. First, its original 30 fps streams are temporally resampled to the 24 fps rate at which both our model and Ego-Exo4D operate, so that a chunk spans the same duration and frame count across datasets. Second, and more fundamentally, H2O's egocentric ground truth is captured with a rectilinear camera that shares the exocentric sensor's geometry and a horizontal field of view of roughly $90°$, whereas Ego-Exo4D's ego view is a wide-angle fisheye. Because our ego canvas, our region decomposition, and our metrics are all defined with respect to the ego intrinsics, feeding H2O's original ego footage at its native calibration would change what $\mathcal{H}$, $\mathcal{S}$, and $\mathcal{U}$ mean and make the comparison to the Ego-Exo4D numbers meaningless. We therefore reproject H2O's egocentric ground truth onto a virtual Aria fisheye camera built from the calibrated intrinsics of the Ego-Exo4D ego sensor at 480×480, so that the ego trajectory occupies the same portion of the canvas and the same region masks remain valid. The reprojection is a geometric resampling only: no colour correction is applied, and the circular aperture of the virtual sensor is filled with the corresponding black corners.

With these two steps, we yield a diverse dataset from H2O (Kwon et al., 2021) with 850 clips, which functions as a genuine cross-dataset test: neither its exocentric viewpoints, nor its participants, nor its camera model, nor its scene layouts match the data on which the model was trained.

B ADDITIONAL RESULTS

B.1 QUALITATIVE RESULTS

Fig. 8 extends the comparison of Fig. 4 to additional clips. The trends are consistent across datasets and splits. Camera-control models (Yu et al., 2025; Jiang et al., 2025; Lin et al., 2026) fails because they work under the assumption of modest viewpoint changes. Exo2Ego baselines (Liu et al., 2024; Kang et al., 2026) recover $\mathcal{S}$ reasonably well but degrade sharply in $\mathcal{H}$ under self-occlusion, and on

UNSEEN and Kwon et al. (2021); Sener et al. (2022) they exhibit the viewpoint drift or even degrade into darkness. Our method maintains both hands, the object state, and the scene geometry. **Please refer to the supplementary video for temporal results, which we consider the primary evidence for streaming performance.**

B.2 USER STUDY

Following the evaluation protocol of prior work (Kang et al., 2026), we conducted a blind user study to evaluate visual fidelity and perceptual coherence across challenging UNSEEN videos from Grauman et al. (2024) and in-the-wild scenarios (Kwon et al., 2021; Sener et al., 2022). A total of 16 participants assessed shuffled, anonymized outputs generated by our method and competing baselines across 10 UNSEEN scene and 20 in-the-wild sequences. For each comparative trial, participants were asked to select the best-performing result according to four criteria: (1) **Overall Quality**: Which video delivers the highest perceptual realism and visual fidelity? (2) **Viewpoint Consistency**: Which result best reflects the true target egocentric camera trajectory and perspective transfer? (3) **Background Consistency**: Which method most faithfully reconstructs and preserves the observable background structure? (4) **Hand-Object Interaction Accuracy**: Which model yields the most physically plausible and accurate articulated hand movements and manipulation dynamics?

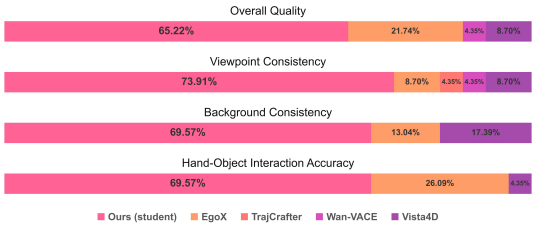

Figure 9: Results for user study questions (1)-(4).

To further examine potential boundary artifacts and temporal drift inherent to chunk-wise causal generation, we designed a dedicated evaluation for video continuity: (5) **Temporal Continuity**: Participants compared our streaming outputs against non-streaming baselines to evaluate whether noticeable stuttering, flickering, or abrupt transitions occur across chunk boundaries. Participants were asked to rate our results against offline methods (Kang et al., 2026) on temporal smoothness and the level of content/viewpoint drift: 5 points (highly consistent throughout with no perceptible drift), 4 points (minor but visually acceptable background drift), and 3 points (moderate drift and flickers), 2 (noticeable drift and flickers that hinders perception) and 1 (severe drift and flickers).

As shown in Fig. 9, our approach received the majority of votes across all comparative dimensions, validating both its high fidelity in Exo2Ego transfer, interaction accuracy and cross-domain generalization. Furthermore, our streaming student scored an average of **4.6/5.0** on extended rollouts, corroborating that our chunk-wise causal architecture sustains long-horizon temporal stability.

C END-TO-END LATENCY

Setup. We analyze the runtime of the deployed streaming student. Model loading—transferring weights to the GPU and initializing the prior estimators—is a one-time cost and is excluded from all figures below. The student consumes one chunk at a time; a chunk spans $L = 33$ video frames at 24 fps, *i.e.* 1.375 s of video. Each stage is measured on the device stated in Table 3. Our end-to-end latency is measured on H100 GPU.

Per-chunk compute. Within a chunk the stages are ordered by data dependency, and the order matters for a streaming implementation. Building the two priors comes first: the dynamic ego prior requires depth, segmentation, and cache updates, while the human pose prior requires body and hand estimation. The two are independent branches and run concurrently, so the stage costs the time of the slower branch. *Both are prerequisites for their own VAE encoding*, and the encoded priors in turn must be available before the diffusion transformer can denoise, since they enter the model as pixel-aligned conditioning channels. Encoding of the exocentric chunk and of the prior streams is independent and is therefore issued as one batched call. VLM inference depends only on the raw exocentric frames, so it is independent of the priors and the VAE, and runs alongside them. Decoding depends on denoising and follows it. Table 3 reports the measured cost of each stage.

Execution pipeline. Rather than executing stages serially, our system pipelines conditioning, generation, and decoding asynchronously across chunks (Fig. 10). When exocentric chunk k arrives, its

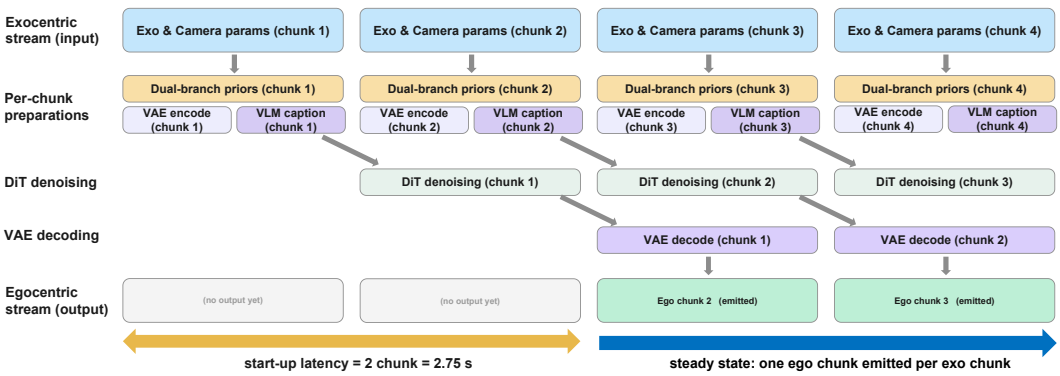

Figure 10: Streaming schedule of the student. Each incoming exocentric chunk triggers a fixed sequence of stages; the corresponding egocentric chunk is emitted when the *next* exocentric chunk arrives, giving a constant start-up latency of one chunk (1,375 ms).

Table 3: Per-chunk stage costs of the streaming student (33 frames, 1,375 ms of video). Stages in the same *parallel group* are independent and can be executed concurrently; the dependency column states what each stage must wait for.

Stage	Device	Depends on	Par.	Latency
(1) Streaming dual-branch priors	H100	—	A	0.58 s
(2) VAE encode, exo chunk + priors	H100	(1)	B	0.30 s
(3) VLM action prompt (6 keyframes)	H100	—	B	0.23 s
(4) Denoising, 3 steps (student)	H100	(2), (3)	C	1.36 s
(5) VAE decode, ego chunk	H800	(4)	D	0.91 s

dual-branch priors and VAE latents are prepared concurrently; because the VLM branch runs faster than VAE encoding (0.23 s vs. 0.58 s), its latency is entirely hidden. Once chunk $k{+}1$ arrives, the student denoises chunk k conditioned on these precomputed priors and historical context to produce clean egocentric latents (1.36 s). Finally, as chunk $k{+}2$ arrives, the VAE decodes the clean latents into RGB frames (0.91 s). Because individual stages execute within the chunk duration (1.375 s) and run in parallel, steady-state throughput sustains interactive frame rates, introducing an end-to-end streaming latency of exactly two chunks.

This is what makes the few-step distillation of Sec. 3.3 necessary rather than merely convenient. The bidirectional teacher requires 50 steps and runs at 0.97 fps, so a single chunk would need roughly 34 s and no amount of pipelining could hide it. Distilling the student to 3 steps raises the generation throughput to 24.3 fps on one H100, which is the figure reported in Table 1.

D Implementation Details

D.1 Streaming dual-branch priors

Both priors of Sec. 3.1 are chunk-causal estimators rather than offline preprocessing: each consumes the exocentric chunk that has just arrived together with a bounded internal state and emits exactly the prior frames the current chunk requires. The dynamic ego prior exploits the fact that the exocentric camera is stationary, so every pixel of the rectified exocentric image corresponds to a *fixed ray in world space*, and only which ray observes a surface changes over time. We therefore keep two memories with different lifetimes: a ray-indexed static cache $\mathcal{A}_t = \{d_t^A(p), c_t^A(p), s_t(p)\}$ holding the first depth and colour observed along each ray, and a short-lived dynamic state holding the previous frame's masks, depth, and flow reference. The cache holds $|\Omega_e|$ entries (624×360 in deployment) irrespective of stream length, and is written only at keyframes, and only where the observation is valid, free of dynamics, and new: background revealed after an occlusion is filled in when it becomes observable, and an accepted entry is never revisited, so the static branch ac-

cumulates no recursive drift—though the same rule freezes the first estimate and offers no way to invalidate it. Segmentation and monocular depth run on keyframes only (every 12 frames), and intervening frames propagate the previous state along backward optical flow, so depth is refreshed from a genuine estimate at most 11 frames old. Depth and masks are propagated, but the *colour* of dynamic surfaces always comes from the current frame, so hands and manipulated objects keep current appearance, and dynamic points are never written into $\mathcal{A}$. Static and dynamic surfaces are then projected into the target camera and resolved in a single depth buffer, with source points reconstructed in the kernel from ray and depth so that no point cloud reaches the host; projection follows the calibrated Aria Fisheye624 model, each point estimates a local footprint radius to suppress striping on grazing surfaces, and the two surface types compete on equal terms so that occlusion is respected rather than the dynamic layer being composited on top of the background.

The human pose prior supplies the structural guidance that pixel-aligned geometry cannot, since it does not by itself disambiguate which arm is which or how fingers are articulated. As estimating hands directly in the exocentric frame is unreliable—the regions are small and hands cross—we bond the hand to the body. Per sampled frame, SMPLer-X (Cai et al., 2023) estimates the body and ViTPose (Xu et al., 2022) refines the upper-limb joints, with the principal subject selected using the projection of the *given* egocentric camera centre and therefore no egocentric pixels. Detections are associated to the left and right wrists by a size-normalized linear assignment, which enforces anatomical sidedness even when arms cross or the subject turns, and the resulting wrist joints define the crop boxes for WiLoR (Potamias et al., 2025), recovering wrist-relative finger articulation; when no hand is detected, a proposal along the forearm carries a deliberately low confidence rather than being silently dropped. Predicted fingers are then rigidly anchored so the hand root coincides with the refined wrist, as in Eq. 2, fixing global position from the arm chain while preserving local articulation—the kinematic analogue of proprioception, and what prevents the spatial misalignment that isolated hand estimation produces on the egocentric canvas. Since only a few frames per chunk are estimated, the rest are interpolated, and a two-bone IK step re-solves each elbow afterwards to restore the limb lengths that interpolation perturbs. The skeleton is projected into both views with the calibrated camera parameters to yield (Q_t^e, Q_t^g), with joint confidence modulating stroke intensity so that speculative finger poses are not presented at full contrast.

Because the cache is $\mathcal{O}(|\Omega_e|)$ and the expensive networks run only on a few keyframes per chunk, the cost of a chunk does not grow with the number of chunks already processed, and the two independent branches can execute concurrently, so the stage costs the time of the slower one. The combined prior stage enters the end-to-end budget as the 0.58 s of Table 3.

D.2 Bidirectional Teacher: PromptMix

The bidirectional teacher receives dense visual conditions from the exocentric video and from the two streaming priors. If the text prompt is allowed to describe scene appearance, it becomes a shortcut: the model can satisfy the training objective by copying environmental wording into the egocentric canvas instead of reading the pixel-aligned prior, and small changes in that wording then alter the output even when the visual evidence is unchanged. PromptMix controls how much information is available through cross-attention by sampling, per training clip, one of four prompt variants:

- a *concise* description of the hand action and the manipulated object, with appearance adjectives (colours, materials, brands) stripped;
- a *detailed exocentric* description that adds a scene summary and richer exocentric context;
- a *full* description combining both;
- an *empty* string.

All variants are derived from exocentric annotations; none contains a description of the ground-truth egocentric frame. The variant is sampled per clip with probabilities 0.3, 0.5, 0.1, and 0.1 respectively, and the concise variant is used at validation and inference. The empty branch is a text conditioning dropout and is not used in the classifier-free-guidance sense: we sample with a guidance scale of 1.0 and never evaluate an unconditional branch.

The design intent is that the visual streams—above all the pixel-aligned dynamic ego prior—become the primary route by which scene layout and appearance reach the model, while text supplies only

coarse action semantics. PromptMix is a regularizer rather than a mechanism with an expected accuracy gain.

D.3 Streaming Adaptation

We adapt the bidirectional teacher into a causal, chunk-wise autoregressive student in three stages, following Helios (Yuan et al., 2026).

Overview. *Stage 1* converts the bidirectional backbone into an autoregressive generator by introducing causal attention across chunks. *Stage 2* reduces the number of noisy tokens per step through a spatial pyramid, which is what makes the per-chunk cost tractable. *Stage 3* distills the multi-step generator into a few-step student through distribution matching and adversarial training. The region-aware corruption mechanism and the first-chunk curriculum of Sec. 3.3 are both introduced in Stage 1 and are retained through Stages 2 and 3. Table 4 lists our training configuration of Stages 1 and 2; Table 5 details Stage 3.

Stage 1: causal adaptation. The student warm-starts from the bidirectional teacher. Attention remains bidirectional *within* a chunk while being strictly causal across chunks. A 121-frame clip encodes to 31 latents, which we partition into chunks spanning latents $(0, 9)$, $(9, 18)$, $(18, 27)$, and $(27, 31)$; the last chunk is shorter and its loss is rescaled by its length fraction so that it does not dominate the average.

The memory buffer stores egocentric latents only. The exocentric video and both priors are supplied fresh for the current chunk and are *discarded* rather than accumulated, so the conditioning cost does not grow with rollout length; distant history is compressed by multi-term memory patchification (Yuan et al., 2026) into short-, mid-, and long-term scales with histories of $(16, 2, 1)$ latent frames. Annotations marking prior validity and dynamic regions accompany the cached latents.

Two departure points from a global prompt are worth noting. First, we use a *per-chunk* prompt describing the action in the chunk currently being denoised, rather than one prompt for the whole clip; text cross-attention is restricted to the current chunk's tokens, so the current description is never re-applied to cached history. This matches the teacher's conditioning semantics—a description of the target being denoised—and allows the action language to follow a changing manipulation without leaking future-chunk text into earlier generations. Second, we regularize the conditioning path throughout training: with probability 0.1 the prior is dropped and with probability 0.1 it is corrupted with Gaussian noise ($\sigma = 0.25$), applied only to the egocentric half of the conditioning channels, which is always zeroed for the exocentric half.

Stage 1 is trained in three sub-phases, each re-loading the previous stage's terminal checkpoint with a fresh optimizer. The sequence implements the first-chunk curriculum and then the region-aware corruption of Sec. 3.3: sub-phase A trains on the history-free first chunk only (sampling probability 1.0); sub-phase B reduces that probability to 0.6 and admits continuation chunks; sub-phase C lowers it to 0.25 and switches on history corruption. Corruption applies to history frames only, and never to the first chunk, which has no history. It combines a frame-level operator (clean 0.3 / blur 0.6 / noise 0.1, the latter mixing with $\sigma \sim \mathcal{U}(0, 1/3)$) with a targeted operator restricted to the hand and interaction masks on the latent grid, which fires per frame with probability $\sim \mathcal{U}(0.5, 0.7)$ and, on firing, applies either stronger noise ($\sigma \sim \mathcal{U}(0.2, 0.5)$) or stronger blur. The targeted operator forces the student to resolve hand dynamics from the current priors rather than from a degraded history.

Stage 2: pyramid token compression. Denoising is divided into three spatial levels while each chunk retains its full temporal span. At 832×480 exocentric and 480×480 egocentric resolution the two views occupy different widths on the latent grid, so they are resized separately; interpolation is never allowed to blend pixels across the seam between them. At every level the exocentric half is clamped to its clean conditioning latent and the flow-matching target and loss apply to the egocentric half only. Training first emphasizes the intermediate level with sampling ratios $(1, 2, 1)$, then balances the levels with $(1, 1, 1)$ at a lower learning rate. Because the finest latent grid is not divisible by successive factors of two, we sample independent unit-variance Gaussian noise directly at each level rather than downsampling a single fine noise tensor, which would otherwise produce spatially varying variance; the finest transition additionally uses negatively correlated block noise to compensate for the correlation introduced by nearest-neighbor upsampling.

Table 4: Configuration of the causal adaptation (Stage 1) and of the pyramid token compression (Stage 2). Each sub-phase re-initializes the optimizer. † The Stage-1 sub-phases share a single LoRA configuration.

Configuration	Stage-1 A	Stage-1 B	Stage-1 C	Stage-2 init	Stage-2 post
Initialization	teacher	1A	1B	1C	2 init
Training steps	2,500	10,000	18,000	16,000	20,000
Learning rate	1×10^{-5}	1×10^{-5}	6×10^{-6}	2×10^{-5}	6×10^{-6}
LR schedule	constant w/ 100-step warmup	←	←	←	←
First-chunk probability	1.0	0.6	0.25	0.25	0.25
History corruption	off	off	on	on	on
Pyramid sample ratios	—	—	—	$(1, 2, 1)$	$(1, 1, 1)$
Global batch size	8	8	8	8	8
Optimizer	AdamW, β_1=0.9, β_2=0.95, ϵ=10^{-8}, weight decay 10^{-4}				
Gradient clipping	1.0	1.0	1.0	1.0	1.0
LoRA rank / α (q, k, v, o)†	256/256	←	←	←	←
LoRA rank / α (FFN)	128/128	←	←	←	←
Trainable modules	LoRA adapters, patch embedding, memory patchifiers				
Numerical precision	bf16	bf16	bf16	bf16	bf16
GPU usage	8×H800	8×H800	8×H800	8×H800	8×H800

Table 5: Configuration of the few-step distillation (Stage 3).

Configuration	Stage-3 ODE init	Stage-3 post
Initialization	Stage-2 post	Stage-3 ODE init
Training steps	3,759	2,250
Generator learning rate	2×10^{-6}	2×10^{-6}
Critic learning rate	—	4×10^{-7}
LR schedule	constant	constant
Optimizer	AdamW, β_1=0, β_2=0.999, wd 10^{-3}	←
Gradient clipping (generator / critic)	1.0 / —	1.0 / 10.0
ODE regression weight	80.0	—
Adversarial weight λ_G (generator)	—	5×10^{-2}
Adversarial weight λ_D (critic)	—	1×10^{-2}
R1 weight / σ	—	100.0 / 0.1
Critic update ratio (TTUR)	—	5
Discriminator start step	—	1,000
Denoising steps per level	2	2
Real-/fake-score guidance scale	1.0 / 1.0	1.0 / 1.0
Numerical precision	bf16	bf16
GPU usage	4×A800	4×A800

Stage 3: few-step distillation. Stage 3 removes the remaining inference cost. It proceeds in two sub-phases. The first is an ODE initialization: the Stage-2 model is treated as a frozen teacher, and a compact set of ODE solution pairs is collected from it, which the student fits by regression at all pyramid levels simultaneously. The second is full adversarial distillation, which follows a generalized Distribution Matching Distillation with an adversarial term (Eq. 5). We instantiate the real- and fake-score estimators from frozen teacher copies; the fake-score branch additionally carries trainable critic parameters and discriminator heads, and since the two estimators are identical at initialization, the critic must depart from the frozen teacher before it can supply a non-degenerate signal. Only the finest pyramid level is scored. Classifier-free guidance is disabled throughout—both estimators are evaluated conditionally—which removes the extra forward pass and closes the train/inference gap. The generator is updated once every five critic updates, and the discriminator heads are attached at four intermediate depths plus the output. Configurations are listed in Table 5.

D.4 BIDIRECTIONAL TEACHER TRAINING

The bidirectional teacher is initialized from Wan2.2-TI2V-5B (Team Wan et al., 2025) and trained for $34{,}000$ steps on 8 H800 GPUs with a global batch size of 8, at a constant learning rate of $2{\times}10^{-5}$, in bf16, with AdamW and ZeRO-2. LoRA ($r{=}\alpha{=}256$) is applied to the attention projections. Clips span 121 frames at 24 fps, with exocentric inputs at 832×480 and egocentric targets at 480×480. PromptMix probabilities are $0.3/0.5/0.1/0.1$ as in Sec. D.2, and the pose condition is dropped with probability 0.05.

The two conditioning branches enter the backbone as additional input channels of the patch embedding, concatenated with the noisy latent along the channel dimension. Because the backbone is warm-started from a pretrained model whose noisy-latent channels are already well scaled, the new branches are initialized as scaled copies of that block rather than from scratch; otherwise a randomly initialized branch would dominate the pretrained channels at step zero and destroy the warm start. We set the initialization scale of the dynamic ego prior branch to 0.5 and that of the human pose prior branch to 0.19 of the noisy-latent block. The scale of the pose branch is deliberately smaller than that of the appearance branch: the pose prior is a sparse, high-contrast skeleton map, and at parity it would inject a far larger activation than the dense appearance prior.

D.5 EVALUATION METRICS

Why region-aware metrics. A conventional pixel-level comparison over the whole egocentric canvas is the wrong instrument for this task, because the three regions have different correctness criteria. In $\mathcal{S}$ the view is shared with the exocentric camera, and the generated content must match the ground truth essentially exactly; a pixel metric is meaningful there. In $\mathcal{U}$ no observation exists by definition, so requiring pixel agreement with the ground truth penalizes legitimate and desirable diversity: a different but coherent completion of an unseen region is not an error. In $\mathcal{H}$ the situation is subtler. The hand and object are not always observable, and their egocentric appearance is not pixel-determined—a slightly different wrist roll, a shifted contact point, or a slightly different grasp all produce plausible images with low pixel similarity to the recorded frame. What matters in $\mathcal{H}$ is *semantic* agreement: the action performed, the positions of the hands, and the spatial relation between hand and object.

Region definition. We derive $\mathcal{H}$–$\mathcal{S}$–$\mathcal{U}$ region masks by combining geometric projection with segmentation on ground-truth egocentric videos. Following the streaming geometric construction in Sec. 3.1, we build a 3D world cache from the exocentric stream and project the cached points into each target egocentric view using the calibrated camera parameters. Pixels covered by valid projections form the initial shared-view region $\mathcal{S}'$, while the remaining pixels form the initial unseen-view region $\mathcal{U}'$. We then apply SAM 3.1 (Meta AI, 2026) to the ground-truth egocentric video, using words from text prompts to segment and track dynamic objects and human limbs; the union of these masks defines the interaction region $\mathcal{H}$. To resolve overlaps, we assign all pixels in $\mathcal{H}$ to the interaction region and set $\mathcal{S} = \mathcal{S}' \setminus \mathcal{H}$ and $\mathcal{U} = \mathcal{U}' \setminus \mathcal{H}$. The resulting three regions are mutually exclusive and together cover the entire egocentric image. These masks support the region-aware loss curriculum during training and region-specific metric computation during evaluation. They are not required at inference time, and generation does not require access to ground-truth egocentric videos.

Image metrics. *$\mathcal{S}$-PSNR* is the PSNR restricted to $\mathcal{S}$, with the mean squared error normalized by the number of pixels inside the mask rather than by the full canvas, so that small-region scores are not inflated. *$\mathcal{S}$-MSSIM* applies SSIM with an 11×11 Gaussian window ($\sigma{=}1.5$) whose local moments are renormalized by the in-mask Gaussian weight, so that pixels outside $\mathcal{S}$ never enter the local statistics; it therefore returns exactly 1.0 for a prediction that differs from the ground truth only in $\mathcal{U}$ or $\mathcal{H}$. *$\mathcal{S}$-LPIPS* uses the AlexNet LPIPS distance map, pooled with the same $\mathcal{S}$ mask; note that LPIPS is evaluated on the full frame pair, so its receptive field includes context outside $\mathcal{S}$ and the score is $\mathcal{S}$-weighted rather than $\mathcal{S}$-isolated.

$\mathcal{U}$-CLIP-I measures the naturalness and visual plausibility of the unseen region by CLIP image-embedding similarity between corresponding predicted and ground-truth frames. We deliberately use a similarity (plausibility) measure here rather than a fidelity measure, for the reason given above: $\mathcal{U}$ is unobservable, so the standard is coherence, not agreement with one realization.

PCK@0.05 measures hand pose accuracy in $\mathcal{H}$. Hand keypoints are detected independently in the prediction and in the ground truth using Yang & Ramanan (2012), matched hand-to-hand by a Hungarian assignment on wrist distance with a handedness term, and a joint is counted correct when its distance to the corresponding ground-truth joint is within 5% of the image diagonal. Keypoints belonging to a ground-truth hand with no matched prediction count as errors, so failing to render a hand is penalized rather than ignored.

Motion-Consistency (MC) measures whether the hands move as they do in the ground truth using Wang et al. (2023), rather than merely occupying the correct positions at each instant. It compares the per-frame displacement of matched keypoints between consecutive frames in the prediction and in the ground truth and reports the fraction whose motion agrees within the same 5%-of-diagonal tolerance used by PCK. This complements PCK: a model can place a hand correctly in every frame while moving it along the wrong path, which PCK alone cannot detect, and MC also captures the drift and jitter that accumulate over a long rollout.

Object metrics. We follow the matching pipeline of EgoX (Kang et al., 2026). Objects are segmented with SAM 2 and tracked across the clip; each object is described by a DINOv3 embedding of its cropped bounding box, and ground-truth and predicted objects are matched by one-to-one assignment on embedding cosine similarity subject to a similarity threshold. For each matched pair we report the *center location error*, the Euclidean distance between bounding-box centers, and the *bounding-box IoU*. Location error captures whether the object is placed correctly on the egocentric canvas, while IoU additionally captures whether its extent is right; both are reported in normalized image units.

Video metrics. *FVD* is the Fréchet distance between I3D feature distributions of predictions and ground truth, computed at the level of a split rather than per video. It captures both the visual distribution gap and, through the temporal features of I3D, inter-frame continuity. We additionally report two reference-free VBench dimensions. *Temporal Flickering (TF)* is the mean absolute difference between consecutive frames, mapped to $[0, 1]$ so that higher is better, and detects high-frequency temporal noise. *Motion Smoothness (MS)* measures the error of an off-the-shelf frame-interpolation model Zhang et al. (2025a) in reconstructing the actual frames from their neighbours, again mapped so that higher is better and thus penalizing motion that is not temporally coherent. Because a static output can score well on both, we read them alongside FVD and the object metrics rather than in isolation.

E LIMITATIONS

The vision motivating this work has two parts: streaming Exo2Ego generation, and the use of that generation to manufacture first-person training data for embodied agents at scale. We have addressed the former as far as the current task formulation allows, but the latter remains constrained by a single assumption that our method—and every existing Exo2Ego method (Luo et al., 2024; Liu et al., 2024; Mahdi et al., 2025; Xu et al., 2025; Kang et al., 2026; Park et al., 2026; Mahdi et al., 2026; Wang et al., 2026) we are aware of—inherits: the egocentric camera trajectory is given in advance.

This assumption is what restricts the applicability of Exo2Ego to the comparatively small set of recordings that are paired with tracked first-person capture, whereas the vast majority of third-person video on the internet has no associated egocentric trajectory. Removing the assumption is beyond the scope of this paper, since trajectory estimation is an orthogonal perception problem rather than a generation one, and since our conditioning and evaluation apparatus both take calibrated ego intrinsics and extrinsics as input.

We regard the obstacle as temporary rather than fundamental. Head and camera pose estimation has advanced rapidly, and streaming, causally evaluated models now exist for this purpose, including egocentric pose estimation models designed for on-device AR/VR deployment with sub-millisecond latency (Yang et al., 2024; Li et al., 2026b), and stateful reconstruction models that consume an image stream online and maintain a persistent, continuously updated scene state (Wang et al., 2025a). Once such estimators are accurate and stable enough to supply a plausible ego trajectory from in-the-wild third-person footage, they could be composed with our pipeline at negligible latency. Realizing the full vision is therefore primarily a matter of that component maturing, rather than of a change to the generative formulation presented here.