

E^3C : Video Generation with 3D Environmental Memory and Ego-Exo Human Pose Control

Qiao Gu^{1,2*}, Lingni Ma¹, Adam W Harley¹, Richard Newcombe¹, Florian Shkurti², and Julian Straub¹

¹ Meta Reality Labs, Redmond, WA 98052, USA

² University of Toronto, Toronto, ON M5S 1A1, Canada

q.gu@mail.utoronto.ca, {lingni.ma, aharley, newcombe}@meta.com, florian@cs.toronto.edu, jstraub@meta.com

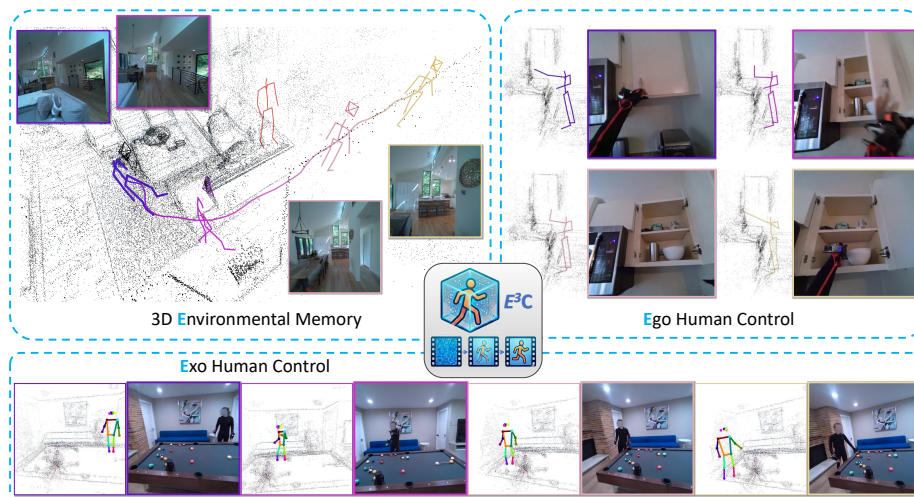


Fig. 1: We present E^3C , a controllable egocentric video generation model that unifies 3D Environmental memory with Ego-Exo human pose Control. E^3C generates egocentric videos that maintain 3D scene consistency (top left), follow ego-body motion (top right), and adhere to exo-human motion (bottom). All images are from generated videos. Project page: e3c-videogen.github.io.

Abstract. Controllable and physically grounded egocentric video generation is essential for embodied agents to reason about how their own and others’ actions manifest and change the world. Compared to generic video synthesis, egocentric generation is especially challenging: the camera is tightly coupled to the actor, leading to rapid viewpoint changes and frequent self-occlusions; the underlying actions are subtle, articulated, and often only partially visible; and both the people and the scene state must evolve consistently with the specified controls. We present E^3C , a controllable video diffusion framework for egocentric generation that builds structured and compact conditions disentangling persistent scene structure from human-driven dynamics. From context frames, E^3C

* Work done during internship at Reality Labs, Meta.

constructs a semi-dense point cloud-based 3D memory and augments each point with appearance descriptors from video-VAE features. Rendering this memory into target viewpoints produces conditioning aligned with the target frames. Human dynamics are modeled separately. The observed people in the scene are controlled by skeleton renderings (exo human control), while the camera wearer is specified by their 3D body joints and 6DoF wrist motion (ego human control). To preserve ego human control when the wearer’s body parts are invisible, we introduce an ego motion encoder that produces persistent cross-attention tokens. Experiments on Nymeria show that E^3C improves visual fidelity, camera-motion accuracy, object consistency, and ego & exo human control over strong baselines, while also enabling intuitive scene editing.

Keywords: Egocentric Vision · Video Generation

1 Introduction

Embodied agents benefit from being able to imagine how candidate actions will change future first-person observations [17, 32]. A controllable and physically grounded egocentric video generator can serve as a rollout simulator [5, 54, 73]: given candidate future human controls from a planner, a user, or a motion prior, it visualizes the resulting first-person video for the camera wearer, nearby people, and the surrounding scene. This capability matters in embodied planning, human-robot interaction, AR/VR, and controlled data generation, where action and perception are tightly coupled [14, 16, 70].

Egocentric video generation is especially challenging because the camera is physically coupled to the actor. Every reach, turn, and step changes not only the body’s configuration but also the viewpoint itself, yielding rapid camera motion, frequent self-occlusion, and large appearance changes across time. At the same time, the underlying actions are complex, subtle, and highly articulated, and their visual consequences may be occluded, delayed, or occur partially or entirely outside the camera’s field of view. To be useful, a first-person video generator must also make both the humans and the environment respond coherently to the input controls. If the wearer turns, reaches, or walks past another person, the generated frames should update the camera view, the visible body, nearby people, and the persistent scene layout in a mutually consistent way. These properties quickly expose the limits of control based only on text prompts, reference images, or other 2D visual signals.

Recent work addresses important parts of this problem. Explicit camera-control methods improve viewpoint steering [22, 24, 62, 63, 71]; 3D-informed generators use geometry or spatial memory to preserve world consistency under viewpoint changes [7, 27, 36, 50, 51, 64, 72]; and egocentric models show that body pose is a powerful signal for predicting first-person visual futures [5, 48, 66]. Yet the key ingredients remain disconnected. Existing egocentric pose-conditioned models do not explicitly maintain a persistent memory of the 3D environment, while 3D-consistent video generators generally target generic scenes and do not

model ego self-occlusion together with nearby people. As a result, fully controllable egocentric generation that is both geometrically grounded and human-aware remains unresolved.

We introduce E^3C , a controllable egocentric video generation model that unifies 3D Environmental memory with Ego-Exo human pose Control, as illustrated in Fig. 1. Our key idea is to build structured and compact conditions that disentangle persistent scene structure from human-driven dynamics. We represent the static world with a semi-dense point cloud-based 3D memory reconstructed from context frames, and render this memory into the target future viewpoints to obtain conditioning aligned with the target future frames. The semi-dense point clouds can be obtained from simultaneous localization and mapping (SLAM) or structure-from-motion (SfM) algorithms, which reconstruct 3D points in high-gradient image regions. To further allow the model to infer accurate image texture, we augment each 3D point with appearance descriptors from video-VAE features. In indoor egocentric scenes, dynamics are largely human-driven, so we explicitly control ego-exo motion and leave induced object changes to the video model. To model human dynamics, including both *exo human* dynamics, i.e., other people observed by the wearer, and *ego human* dynamics, i.e., the wearer’s own body motion, we render every dynamic 3D skeleton into the conditioning camera viewpoints. Since the wearer’s own body is only partially observed in these views, we also introduce ego pose encoding tokens to capture the full (amodal) egocentric body motion. During generation, 3D memory anchors sparse view-aligned structure, pose controls steer human motion, and the diffusion prior fills missing appearance, people, and dynamics.

We build E^3C on a pretrained video diffusion model, VACE [28, 58], finetuned and evaluated on Nymeria [42]. E^3C strongly outperforms 3D-aware and egocentric baselines on visual fidelity, camera-motion adherence, object consistency, and ego/exo human control. Because E^3C ’s conditioning is explicit, it also supports intuitive scene editing, such as removing objects from the 3D memory or changing the motion of surrounding people. Our contributions are:

- We propose E^3C , a controllable egocentric video diffusion framework that jointly models the 3D environment and ego and exo human motion.
- We introduce a semi-dense point cloud-based 3D environmental memory, augmented with per-point appearance features, enabling geometrically grounded generation under strong egocentric camera motion.
- We design an ego pose encoder that converts 3D body joints and wrist poses into persistent cross-attention tokens, improving ego-motion adherence under self-occlusion and out-of-view body configurations.
- We demonstrate state-of-the-art results on the Nymeria dataset, and show that explicit memory and pose conditioning enable intuitive scene editing.

2 Related Work

Video Generation with Camera and Pose Controls. A broad family of controllable video generators improves adherence to user intent by conditioning diffusion models on explicit motion cues, including camera trajectories and

articulated poses. Recent work has made camera motion a first-class control signal for text/image-to-video generation, using learned controllers or plug-and-play guidance to follow user-specified paths [22, 24, 62, 71]. To reduce drift under viewpoint changes, several methods further inject geometry-aware camera conditioning—especially in image-to-video and diffusion-transformer settings—leading to stronger 3D-consistent camera control [3, 4, 37, 67, 80]. Meanwhile, unified motion control frameworks begin to jointly control camera and actor/object motion using 3D or 6D pose representations [7, 52]. However, most existing controllers target exocentric or generic domains and do not explicitly handle egocentric self-occlusion and the tight coupling between the wearer’s body motion and camera motion. *E³C* addresses this gap by conditioning on exo 2D skeleton renderings and ego 3D body and wrist poses, and introducing pose tokenization for persistent cross-attention that preserves ego human control even when the body is partially out of view.

3D-aware Video Generation and Spatial Memory. Maintaining long-horizon spatial consistency under large viewpoint changes remains challenging for video diffusion models, and recent methods increasingly inject explicit 3D structure into the generation process. One line of work conditions video generators on camera and geometric cues—through camera-controllable architectures or explicit 3D supervision [77]—or through rendering-based conditioning from reconstructed scene representations [10, 27, 51]. In particular, Gen3C [51] constructs a depth-derived point-cloud cache and renders it along a target trajectory to guide synthesis, improving camera control and mitigating geometric drift. Complementary to one-shot caches, spatial memory mechanisms maintain an explicit world state over time, retrieving and updating geometry to avoid attending to the full history [34, 36, 59, 64, 79]. Related efforts promote 3D consistency via multi-view video generation for 4D content [57, 65, 74] or training-time alignment of diffusion representations with geometric priors [11, 12]. While these advances focus on scene consistency and camera control, they typically assume a free-flying camera and do not explicitly target egocentric body control or multi-human interactions. *E³C* adapts the 3D-memory paradigm to egocentric generation by building a semi-dense point-based map enriched with per-point appearance features from video latents, providing spatiotemporally aligned, viewpoint-aware conditioning for future synthesis.

Egocentric Video Generation and Ego–Exo Settings. Recent progress in egocentric video generation has been driven by large-scale data and diffusion-based generators, with benchmarks such as EgoVid-5M providing paired video–action supervision and strong baselines for controllable ego synthesis [60]. Building on these resources, egocentric “world models” increasingly condition future prediction on actions or kinematics: PEVA uses whole-body pose trajectories for egocentric video prediction [5], EgoControl enables pose-controllable egocentric generation via 3D full-body signals [48], and EgoTwin explores joint generation of egocentric video and motion [66]; complementary work studies long-horizon consistency via long-context diffusion mechanisms [75]. In parallel, ego–exo settings investigate cross-view generation and translation, including exo-to-ego synthesis

and extensions [29,39,41,43], as well as ego-to-exo generation [40] and cross-view ego future prediction from exo observations [68]. Different from translation-centric formulations or pose-only conditioning, E^3C targets future egocentric generation from past observations while *jointly* leveraging explicit 3D memory and controllable ego-exo pose signals, aiming to improve both scene coherence and motion adherence in complex real-world egocentric recordings.

3 Method

E^3C conditions a latent video diffusion transformer (DiT) [49] on view-aligned renderings of a 3D environmental memory and on ego-exo human pose controls, as illustrated in Fig. 2. In this section, we first introduce the problem setup (Sec. 3.1), and then describe the latent diffusion backbone (Sec. 3.2), the 3D memory and view-aligned rendering (Sec. 3.3), and finally ego body pose conditioning (Sec. 3.4).

3.1 Problem Overview

We study conditional egocentric video generation. Given m context RGB frames $\mathbf{x}_{1:m} \in \mathbb{R}^{m \times H \times W \times 3}$ and a text prompt τ , we synthesize n frames $\hat{\mathbf{y}}_{1:n} \in \mathbb{R}^{n \times H \times W \times 3}$ that are consistent with the observed environment and follow user-specified human motion controls. The corresponding GT target frames are denoted as $\mathbf{y}_{1:n} \in \mathbb{R}^{n \times H \times W \times 3}$. The context frames are often past observations, but can be any set of relevant frames. From the context clip, we construct a 3D environmental memory \mathcal{M} and render it along a target camera trajectory $\mathbf{T}_{1:n}$ to condition our generative model on the 3D environment and the camera’s path through it. In egocentric videos, $\mathbf{T}_{1:n}$ is typically given by the future headset pose trajectory; at inference, it can be user-specified. We control human motion with ego motion $\mathbf{p}_{1:n}^e$ and exo motion $\mathbf{p}_{1:n}^x$: ego human motion consists of 3D body joints $\mathbf{p}_{1:n}^{e,s}$ and 6DoF wrist poses $\mathbf{p}_{1:n}^{e,w}$, and exo human motion is specified as 2D skeleton maps in the image plane. We additionally condition our model on the last observed frame \mathbf{x}_m , serving as a high-detail visual cue. Overall, we learn:

$$p_{\theta}(\hat{\mathbf{y}}_{1:n} \mid \mathbf{x}_m, \tau, \mathcal{M}, \mathbf{T}_{1:n}, \mathbf{p}_{1:n}^e, \mathbf{p}_{1:n}^x). \tag{1}$$

3.2 Latent Video Diffusion Preliminaries

E^3C builds on VACE [28, 58], a pretrained latent video diffusion model. Let $\mathcal{E}(\cdot)$ and $\mathcal{D}(\cdot)$ denote a pretrained video VAE encoder and decoder. We encode the target clip into latents $\mathbf{z}_0 = \mathcal{E}(\mathbf{y}_{1:n})$ and apply a flow-matching noising process [15, 38]:

$$\mathbf{z}_t = (1 - \sigma_t) \mathbf{z}_0 + \sigma_t \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad t \in \{1, \dots, T\}. \tag{2}$$

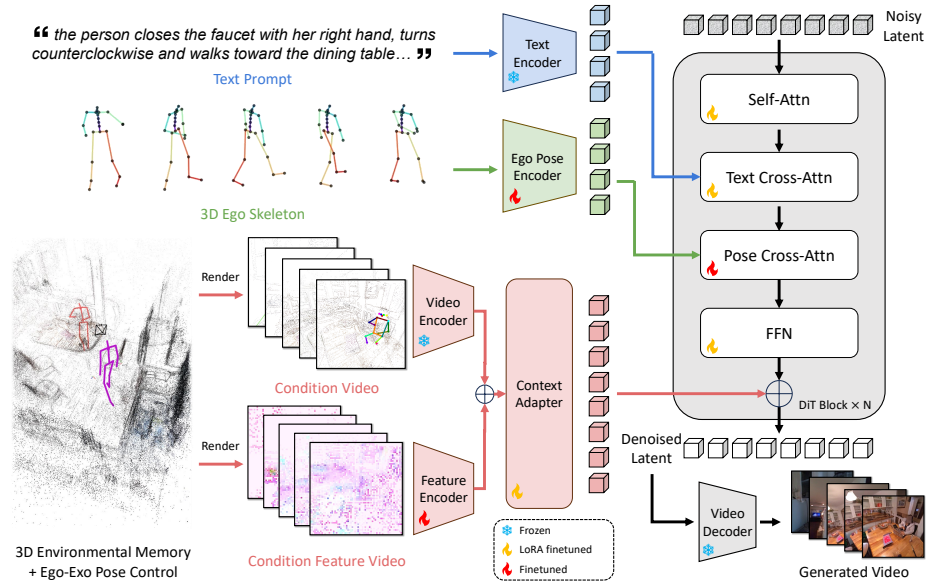


Fig. 2: Overview of E^3C . A latent video DiT iteratively denoises noisy video latents conditioned on a text prompt, view-aligned renderings of a semi-dense point cloud-based 3D memory augmented with VAE appearance features, and ego-exo human pose controls. The rendered memory is injected via the context adapter, while pose controls are provided both as pose drawings in the conditioning video and as persistent ego pose tokens for cross-attention.

Here $\sigma_t \in [0, 1]$ is a monotonically increasing noise schedule over timesteps, and the corresponding flow-matching target is $\mathbf{u} = \boldsymbol{\epsilon} - \mathbf{z}_0$. A denoiser ϵ_θ predicts the flow-matching target conditioned on text and control signals:

$$\hat{\mathbf{u}} = \epsilon_\theta(\mathbf{z}_t, t; \tau, \mathcal{C}), \quad (3)$$

where ϵ_θ is implemented as a video DiT [49]. Here \mathcal{C} aggregates all non-text conditions, including the last context frame \mathbf{x}_m , view-aligned 3D memory renderings (Sec. 3.3), and ego-exo pose controls (Sec. 3.4). We train with a timestep-weighted regression objective:

$$\mathcal{L} = \mathbb{E}_{\mathbf{z}_0, t, \boldsymbol{\epsilon}} \left[w(t) \|(\boldsymbol{\epsilon} - \mathbf{z}_0) - \epsilon_\theta(\mathbf{z}_t, t; \tau, \mathcal{C})\|_2^2 \right]. \quad (4)$$

At inference, we sample from the reverse process starting from $\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and decode the final latent clip via $\mathcal{D}(\cdot)$. As shown in Fig. 2, text conditioning and ego pose conditioning enter through cross-attention, and view-aligned spatiotemporal conditions are injected through a context adapter network [28].

3.3 3D Spatial Memory for Consistent Video Generation

Rapid egocentric camera motion makes it difficult for video diffusion models to preserve a coherent 3D world over time. We address this by constructing a 3D

spatial memory \mathcal{M} from the context and rendering it into the target viewpoints as view-aligned conditioning. Concretely, we represent \mathcal{M} as a semi-dense point cloud \mathbf{P} with per-point color \mathbf{C} and latent appearance descriptors \mathbf{F} .

Memory construction. Given the m context frames $\mathbf{x}_{1:m}$, we run a SLAM algorithm [13] to build a semi-dense point cloud $\mathbf{P} \in \mathbb{R}^{N \times 3}$ of the environment. For each point in \mathbf{P} , we assign an RGB color $\mathbf{C} \in \mathbb{R}^{N \times 3}$ by sampling from the context frames where the point is observed. The point cloud is voxel-subsampled, and the color of each voxel is computed by mean pooling the colors of points within the voxel. Compared to video models that rely on monocular dense depth estimation [7, 51, 64], our semi-dense reconstruction provides metrically accurate point clouds and a temporally coherent semi-dense map. However, semi-dense SLAM reconstructs points primarily at high image-gradient locations (e.g., edges and textured regions), so the resulting point cloud concentrates on visually informative structure while leaving low-texture areas under-represented. See Fig. 2 for an example 3D memory.

Per-point appearance representation. The semi-dense map \mathcal{M} is sparse and may miss appearance details in textureless surfaces or partially occluded areas. To enrich this spatial memory, we augment each 3D point with local appearance features extracted from the input frames. We encode the context frames $\mathbf{x}_{1:m}$ using the video VAE encoder \mathcal{E} (Sec. 3.2), and extract from these feature maps per-point latent descriptors $\mathbf{F} \in \mathbb{R}^{N \times D}$. Similar to the construction process of \mathbf{P} and \mathbf{C} , we project each 3D point into the context frames’ feature maps, sample latent descriptors using bilinear interpolation, and aggregate multi-view samples by voxelizing the points and mean-pooling features within each voxel. During generation, these learned appearance features allow the diffusion model to extrapolate texture and fill regions not directly covered by the sparse geometry, while remaining grounded by the high-gradient structural support of the SLAM map.

View-aligned rendering and conditioning. Given a target camera trajectory $\mathbf{T}_{1:n}$, we project and render the augmented point cloud (with both color \mathbf{C} and latent features \mathbf{F}) into each target viewpoint. This produces a sequence of RGB renderings \mathbf{c}^{pc} and rendered feature maps \mathbf{c}^{feat} that are spatiotemporally aligned with the target video frames $\mathbf{y}_{1:n}$. Human pose conditions are also drawn on \mathbf{c}^{pc} , resulting in \mathbf{c}^{ctrl} , as will be detailed in Sec. 3.4. We then encode \mathbf{c}^{ctrl} using the video encoder, yielding latent features $\mathbf{e}^{\text{ctrl}} \in \mathbb{R}^{n' \times H' \times W' \times D}$. In parallel, we encode \mathbf{c}^{feat} using a separate feature encoder to obtain $\mathbf{e}^{\text{feat}} \in \mathbb{R}^{n' \times H' \times W' \times D}$. We fuse the two streams by summation, $\mathbf{e} = \mathbf{e}^{\text{ctrl}} + \mathbf{e}^{\text{feat}}$, and feed the fused features \mathbf{e} through the VACE Context Adapter [28] to inject them into the video DiT blocks as residual conditioning (Fig. 2). Note that E^3C does not take visibility or depth maps associated with \mathbf{e} as explicit inputs, but learns to reason about them implicitly.

3.4 Controllable Video Generation with Ego-Exo Human Poses

Human motion drives egocentric video dynamics, inducing camera motion, self-occlusions, and interactions with the environment. To enable fine-grained control

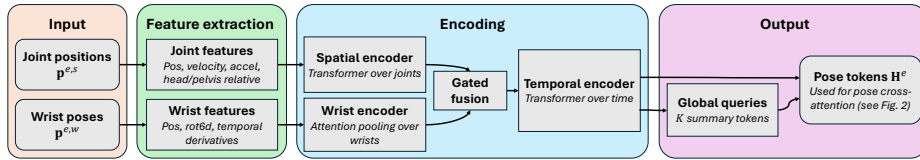


Fig. 3: Ego pose encoder. The encoder maps ego 3D body joints $\mathbf{p}^{e,s}$ and 6DoF wrist poses $\mathbf{p}^{e,w}$ to motion-aware tokens \mathbf{H}^e using spatial and temporal transformers and gated fusion. \mathbf{H}^e is then used for pose cross-attention (Fig. 2), providing persistent ego human motion control when the body is out of view or occluded. See the appendix for network details.

over human motion in the generated videos, $E^{\mathcal{3}C}$ conditions on pose sequences for both the ego human \mathbf{p}^e and the exo humans \mathbf{p}^x over the target n frames and is trained to follow these controls.

Exo human pose control. Exo humans are typically fully visible and appear at a moderate distance from the camera wearer, so coarse articulated motion provides a strong control signal. Following prior work [7, 28, 76], we rasterize their 2D skeletons and overlay them on the point-cloud renderings \mathbf{c}^{PC} to form the RGB conditioning video \mathbf{c}^{ctrl} . See example conditioning frames in Fig. 2.

Ego human pose control. The ego wearer is often only partially observed due to being out of view or close-range self-occlusion, yet their motion is tightly coupled to the camera trajectory and dominates first-person visual dynamics. We condition on target ego 3D body joints $\mathbf{p}^{e,s} \in \mathbb{R}^{n \times J \times 3}$ and 6DoF wrist poses $\mathbf{p}^{e,w} \in \mathbb{R}^{n \times 2 \times 4 \times 4}$, represented as 3D joint positions and $\text{SE}(3)$ transforms, respectively, in the camera coordinate frame. Similar to exo human pose control, we project these 3D signals into each frame using the camera intrinsics and rasterize them into \mathbf{c}^{ctrl} as 2D skeletons and small coordinate axes. However, because the ego body is frequently out of view or occluded, these drawings can be sparse and provide limited control on their own.

Persistent ego pose tokens. To provide ego human pose control even under partial observability, we further encode the full target ego pose sequence into persistent tokens, as illustrated in Fig. 3. Specifically, we augment joint and wrist poses with first-order temporal differences, represent joints by their 3D coordinates and wrists by their positions and a continuous 6D rotation representation [81], and linearly project all features into a shared D -dimensional space. We then embed the joint and wrist sequences, aggregate across joints and wrists with spatial transformer blocks [56] and attention pooling [35], fuse the two streams with a gated module [1, 53], and propagate information temporally. This results in n pose tokens, each summarizing the joint and wrist pose information of one target frame. We additionally introduce K learned summary tokens that cross-attend to the n pose tokens [9], and concatenate them with the pose tokens. The resulting $n + K$ tokens $\mathbf{H}^e \in \mathbb{R}^{(n+K) \times D_{DiT}}$ condition the video DiT via a dedicated pose cross-attention, as shown in Fig. 2, where D_{DiT} denotes the video DiT token dimension. More network details are provided in the appendix.

4 Experiments

We evaluate E^3C on the Nymeria egocentric benchmark [42] and analyze both generation quality and controllability. We compare against 3D-aware methods with explicit spatial representations, conditional video diffusion models that take aligned control videos as input, and egocentric pose-conditioned generators. Beyond standard video-quality metrics, we report camera-motion accuracy, object consistency, and human pose adherence. We analyze each conditioning component via ablations and show video generation results with controllable edits.

4.1 Experiment Setup

We first detail the experimental setup used to evaluate E^3C and baselines.

Dataset. We use the Nymeria dataset [42], which contains over 1,100 egocentric recordings captured with Project Aria glasses [13] across diverse indoor and outdoor environments. Each recording includes dense natural-language captions, time-synchronized 3D body motion from an Xsens inertial motion capture system [46], and 6DoF wrist poses obtained from the miniAria wristbands. We use Project Aria MPS [44] to recover camera trajectories and semi-dense point cloud reconstructions that serve as the 3D memory input. Each recording session includes a *participant* who performs the activity, an *observer*, and additional bystanders. We treat the participant as the ego wearer and use the participant recordings for training and evaluation. Since bystanders lack synchronized 3D pose annotations, we detect exo humans with YOLO-X [19] and extract 2D skeletons using ViTPose [69]; these skeletons are used as exo human pose controls. We adopt the train/validation/test split from [21] with disjoint environments and participant identities across splits.

Baselines. We compare E^3C against baselines that cover complementary points in the design space. Splatfacto [47] is a Gaussian Splatting novel-view synthesis method that reconstructs a scene from context frames and renders target viewpoints. We also evaluate two 3D-aware video generators that use explicit spatial memory, VMem [36] and Gen3C [51]. As a strong conditional video diffusion baseline, we include VACE [28], which takes spatiotemporally aligned conditioning videos as input. We report VACE results for the released pretrained checkpoint and for models fine-tuned on Nymeria under three conditioning settings: exo human pose only, point-cloud renderings only, and point-cloud renderings with exo human pose. Finally, we compare to egocentric video generators trained and evaluated on Nymeria, including PEVA [5], EgoControl [48], and EgoTwin [66].

Implementation Details. We preprocess Nymeria videos by devignetting and rectifying them to a pinhole camera model [20], then subsample to 10 FPS and resize frames to 512×512 . We segment each recording into 280-frame snippets with a stride of 10 seconds. The first $m = 200$ frames form the context $\mathbf{x}_{1:m}$ for constructing the 3D memory, and the remaining $n = 80$ frames define the target clip $\mathbf{y}_{1:n}$. Notably, the final context frame \mathbf{x}_m is also provided as input to E^3C , serving as a visual cue for prediction. We initialize E^3C from

Table 1: Quantitative evaluation on the Nymeria dataset. We indicate whether each model is fine-tuned on the Nymeria training set (“FT”). The **best** and second-best results for each metric are bold and underlined, respectively.

Method	Metric	Visual Quality				Camera Control		Object		Exo Human Control		Ego Human Control	
	FT	FVD↓	LPIPS↓	PSNR↑	SSIM↑	TErr(cm)↓	RErr(°)↓	Obj-F1↑	Obj-mIoU↑	Exo-F1↑	PCK@10%↑	Hand-F1↑	Hand-mIoU↑
Splatfacto [47]	–	1040	0.494	<u>17.8</u>	<u>0.620</u>	6.60	21.5	41.05	57.42	15.28	63.13	21.28	<u>19.28</u>
Gen3C [51]	×	974	0.565	15.8	0.541	21.37	88.5	25.66	42.00	12.39	42.08	25.54	14.50
VMem [36]	×	465	0.618	12.7	0.470	16.82	113.0	14.74	32.45	8.67	2.55	26.31	17.68
VACE [28] (exo)	×	1180	0.644	11.7	0.401	19.00	125.0	12.43	27.37	11.14	34.69	31.92	11.28
VACE [28] (pts)	×	2460	0.675	11.5	0.357	14.22	72.2	6.30	18.30	3.40	85.75	17.58	15.47
VACE [28] (pts+exo)	×	2440	0.675	11.6	0.359	13.67	65.3	6.37	18.49	4.36	<u>71.26</u>	17.82	15.46
VACE [28] (exo)	✓	496	0.584	13.0	0.454	8.77	101.0	15.83	34.19	36.02	64.95	32.74	13.20
VACE [28] (pts)	✓	329	<u>0.440</u>	17.1	0.604	<u>3.25</u>	24.2	46.95	71.14	19.87	37.51	34.18	18.82
VACE [28] (points+exo)	✓	<u>328</u>	<u>0.440</u>	17.3	0.604	5.41	24.5	<u>47.95</u>	<u>71.57</u>	<u>36.26</u>	64.08	<u>34.24</u>	18.17
<i>E³C</i> (Ours)	✓	249	0.415	18.6	0.629	2.40	<u>23.3</u>	52.85	74.11	39.59	72.88	41.44	24.76

Wan2.1-VACE-1.3B [28, 58], but the proposed framework is compatible with any video latent diffusion model. We LoRA fine-tune the main video DiT and the context adapter with rank 256 [25]. The ego pose encoder, feature encoder, and pose cross-attention layers are trained from scratch, as shown in Fig. 2. *E³C* is trained using the Adam optimizer [31] with a learning rate of 1×10^{-4} for 50 epochs on 64 NVIDIA H100 GPUs, with a per-device batch size of 1. At inference, we use a fixed 50-step FlowMatch sampling schedule [38] with shift parameter $\sigma_{\text{shift}} = 5.0$ and classifier-free guidance scale $\omega = 5.0$.

4.2 Evaluation Metrics

Visual Quality and Fidelity. We report standard video generation metrics: Fréchet Video Distance (FVD) [55] for overall visual and temporal fidelity. Since generation is conditioned on the 3D spatial memory and camera poses, the predicted frames are expected to be pixel-aligned with the ground-truth frames. Therefore, we also evaluate LPIPS [78] for perceptual similarity, and PSNR/SSIM [61] for pixel-level reconstruction and structural consistency.

Object Consistency. To measure semantic object consistency, we run the open-vocabulary detector OWLv2 [45] on both ground-truth and generated videos. We match predicted boxes to ground-truth boxes with the Hungarian algorithm [33] and report Obj-F1 and Obj-mIoU.

Camera Motion Control. We evaluate camera-motion adherence by treating Project Aria MPS poses [44] as ground-truth trajectories. We estimate trajectories from generated videos with VIPE [26], align them to ground truth using a Sim(3) transform, and report translation error (TErr) and rotation error (RErr) following prior work [22, 48].

Ego and Exo Human Pose Control. We evaluate pose adherence for both the ego wearer and exo people. For ego human control, we follow [48, 66] and report Hand-F1 and Hand-mIoU. We segment ego hands in both ground truth and generated videos using SAM3 [8]. Hand-F1 measures hand presence per frame, and Hand-mIoU measures mask overlap over frames where hands are present in the ground truth. For exo human control, we extract 2D bounding boxes and keypoints from both ground truth and generated videos using YOLO-X and ViTPose [19, 69]. We match people using Hungarian matching [33] and report

Table 2: Comparison with egocentric video-generation baselines on Nymeria. Left: we follow the data split and evaluation protocol used in PEVA [5] and EgoControl [48] for a fair comparison. Right: EgoTwin [66] results are reported from the original paper, since the split and implementation details are not available; the protocols may differ.

Model	LPIPS ↓ DreamSim ↓ FID ↓		
PEVA XXL [5]	29.8	18.6	61.10
EgoControl [48]	24.3	11.3	50.68
E^3C (Ours)	17.3	10.5	50.82

Model	I-FID ↓	FVD ↓	TErr(cm) ↓	RErr(*) ↓
EgoTwin [66]	98.2	1033.5	67.0	26.4
E^3C (Ours)	31.0	249.1	2.40	23.3

Table 3: Ablations on conditioning inputs and proposed components. Starting from the *Base* model [28] conditioned only on the initial image and text, we incrementally add rendered SLAM points (*Point*), rendered per-point VAE appearance features (*Feat*), exo human pose control (*Exo*), ego pose drawn on the conditioning frames (*Drawn Ego*), and our encoded ego pose tokens (*Encoded Ego*). Each variant is trained and evaluated independently.

Method	Metric	Visual Quality				Camera Control		Object		Exo Human Control		Ego Human Control	
		FVD ↓	LPIPS ↓	PSNR ↑	SSIM ↑	TErr(cm) ↓	RErr(*) ↓	Obj-F1 ↑	Obj-mIoU ↑	Exo-F1 ↑	PCK@10% ↑	Hand-F1 ↑	Hand-mIoU ↑
Base (Image+Text)		1380	0.729	10.2	0.323	14.56	124.0	4.19	18.99	0.99	1.90	27.72	9.56
+ Point		329	0.440	17.1	0.604	3.25	24.2	46.95	71.14	19.87	37.51	34.18	18.82
+ Feat		310	0.424	17.9	0.617	2.67	23.5	50.83	73.10	20.92	45.07	34.26	17.72
+ Exo		266	0.420	18.3	0.622	4.30	25.9	52.76	73.32	38.40	69.45	33.04	17.35
+ Drawn Ego		256	0.417	18.5	0.627	2.50	21.3	53.02	74.21	38.91	70.77	36.03	19.36
+ Encoded Ego (E^3C)		249	0.415	18.6	0.629	2.40	23.3	52.85	74.11	39.59	72.88	41.44	24.76

Exo-F1 and PCK@10%, where PCK (percentage of correct keypoints) uses a threshold of 10% of the square root of the matched bounding box area. Note that PCK@10% is computed only over matched bounding boxes, and therefore can remain high even when Exo-F1 is low. As such, Exo-F1 and PCK@10% should be interpreted jointly for a complete assessment of exo pose control.

4.3 Results

Quantitative comparison with general-domain video generators. We evaluate all methods on 100 snippets from the Nymeria test set and report results in Tab. 1. Note that the results for object consistency and exo and ego human control metrics are scaled by 100. For Splatfacto [47], we reconstruct a 3D Gaussian Splatting scene for each snippet from the context frames and render the target viewpoints. For Gen3C [51] and VMem [36], we use their official implementations and pretrained weights, feeding the first 200 frames as context. For VACE [28], we evaluate both the released pretrained checkpoint and models fine-tuned on Nymeria using three conditioning settings: exo pose only, point cloud only, and point cloud plus exo pose. Each VACE fine-tune run uses the same training setup as E^3C .

Nymeria poses a challenging distribution shift for general-domain video generators due to rapid egocentric motion, frequent self-occlusion, and complex indoor scenes. As shown in Tab. 1, zero-shot models and some 3D-aware baselines degrade substantially. Fine-tuned VACE models recover strong performance when provided with point cloud conditioning, and adding exo skeletons further improves exo-control metrics. Compared to the best fine-tuned VACE baseline (points+exo), E^3C augments points with appearance features, adds the feature

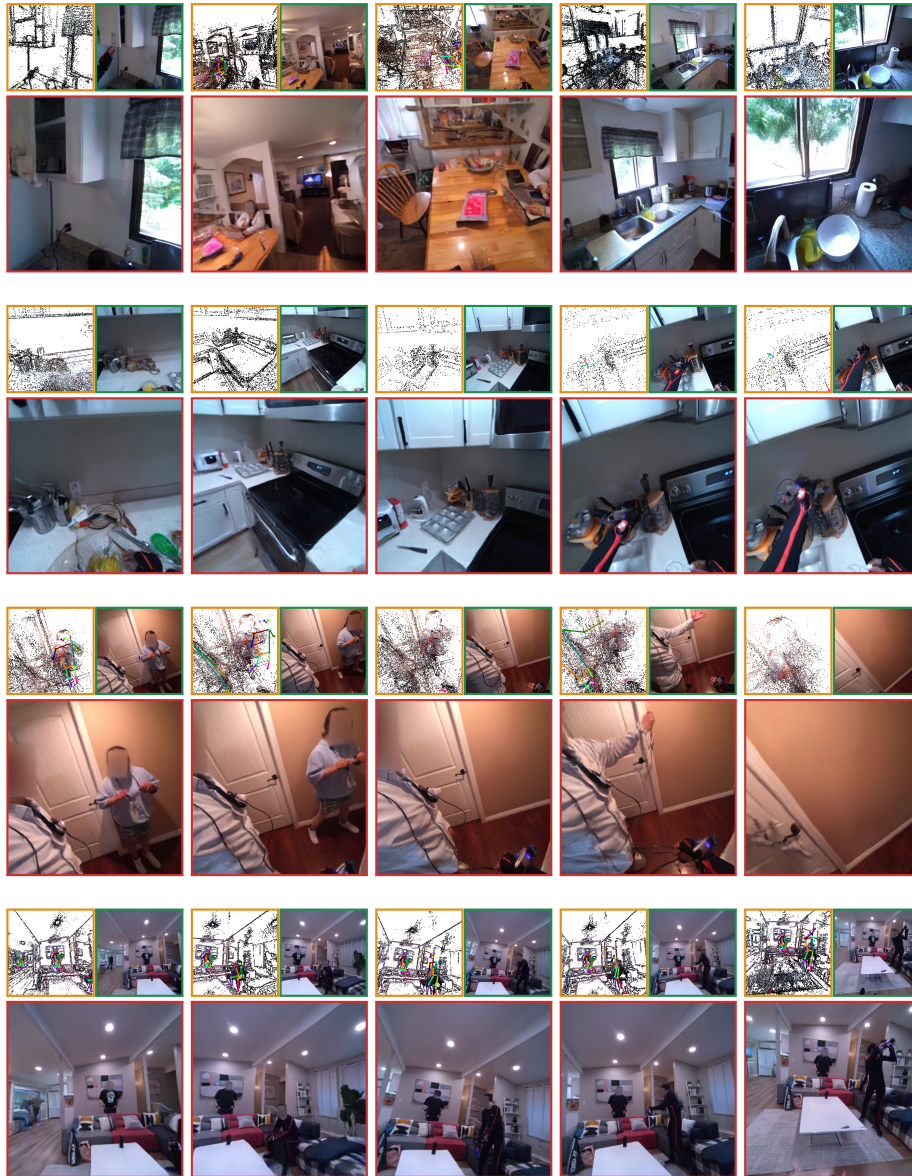


Fig. 4: Qualitative results. Each row corresponds to one video snippet and columns show different timestamps. In each cell, the top row shows the **conditioning visualization** (left) and the **ground-truth** frame (right), and the bottom row shows the **generated** frame from E^3C . E^3C maintains consistent 3D scenes (row 1), follows ego-body motion (row 2), and adheres to exo-human motion (rows 3–4). See full videos on the project page.



Fig. 5: Qualitative ablations. Each row shows the GT frame and the corresponding predictions with and without each type of condition. Top: removing per-point appearance features (*Feat*) increases texture and color drift. Middle: removing exo pose control (*Exo*) harms exo-motion adherence. Bottom: removing encoded ego pose tokens (*Ego*) reduces ego hand/body fidelity under self-occlusion.

encoder branch, and uses the ego pose encoder for persistent ego human control. These help E^3C achieve the strongest overall performance, improving fidelity and object consistency while reducing camera translation error and improving ego hand metrics over the fine-tuned VACE baselines.

Among the non-diffusion baselines, Splatfacto provides accurate viewpoint rendering when the scene is sufficiently reconstructed, but it struggles with dynamic content and regions outside the context views. VMem produces plausible videos but exhibits larger camera-trajectory errors, consistent with the difficulty of transferring to Nymeria’s high-frequency egocentric motion. See qualitative comparisons on the project page.

Quantitative comparison with egocentric video generators. We also compare against egocentric video generators evaluated on Nymeria in Tab. 2. For PEVA [5] and EgoControl [48], we align the data split and follow their evaluation protocol: each method uses a 15-frame context window from the past 3.75 seconds and is evaluated on a single generated frame 2 seconds into the future. We report LPIPS [78], DreamSim [18], and FID [23] in Tab. 2 (left). The results show that E^3C improves LPIPS and DreamSim and achieves comparable FID with the state-of-the-art EgoControl method [48]. For EgoTwin [66], we cannot fully align the evaluation protocol because the original split and implementation details are not available³. We therefore report the numbers from their paper alongside our results as a reference point in Tab. 2 (right), noting that the protocols may differ.

Qualitative Results. Fig. 4 visualizes representative generations by E^3C . The model maintains coherent 3D scene structure across large viewpoint changes

³ EgoTwin authors informed us they no longer have access to the code.



Fig. 6: Editing by manipulating explicit 3D memory and pose conditions. Left four columns: object removal by deleting the object’s points from the 3D memory, which consistently removes it from generated frames. Right four columns: exo-person removal by removing the corresponding pose control. Both **conditioning** and **generated** frames before and after the edits are shown. See full videos on the project page.

while following both ego actions and exo motions, suggesting that the rendered spatial memory and pose controls provide complementary guidance.

4.4 Ablation Study

We ablate the conditioning signals and modules in E^3C and report the results in Tab. 3 and Fig. 5. Starting from a Nymeria-fine-tuned VACE model that is conditioned only on the text prompt and one context image \mathbf{x}_m (Base), we incrementally add point-cloud renderings (Point), per-point VAE appearance features (Feat), exo skeleton controls (Exo), drawn ego pose overlays (Drawn Ego), and our encoded ego pose tokens (Encoded Ego). Each variant is trained and evaluated independently under the same setup as the full model.

Point cloud conditioning provides the most significant improvement in visual fidelity, camera control, and semantic object consistency, showing that explicit geometry is crucial for the video generator to be 3D-consistent and to follow rapid head motion. Adding per-point appearance features further improves LPIPS/PSNR/SSIM, reduces perceptual discrepancy, and enhances texture sharpness (cf., Fig. 5 top row). Exo and drawn ego overlays increase exo and ego human control metrics, respectively. Finally, encoded ego tokens yield the strongest gains for ego human control, improving Hand-F1 from 36.03 to 41.44 and Hand-mIoU from 19.36 to 24.76, indicating more reliable hand presence and segmentation overlap.

4.5 Scene Editing

The explicit conditioning also enables intuitive edits. Fig. 6 shows examples of removing objects by deleting points from the 3D memory and removing exo people by modifying the exo pose inputs. Because the edits are applied at the 3D representation and conditioning level, the resulting changes remain temporally and spatially consistent throughout the generated sequence. This behavior further illustrates the advantages of explicit 3D memory and pose conditioning for video generation. Additional results and videos are provided on the project page.

5 Conclusion

We presented E^3C , a controllable egocentric video generation framework that integrates 3D environmental memory with joint ego-exo human pose control. A semi-dense point cloud-based spatial memory constructed from past views and augmented with per-point appearance features from video latents is rendered into target viewpoints to preserve scene geometry under strong ego motion. Exo humans are controlled via 2D skeleton renderings, while the ego wearer is conditioned by 3D body and wrist poses plus encoded ego pose tokens to robustly handle invisible body parts. Experiments on Nymeria show that E^3C outperforms 3D-aware and pose-conditioned baselines on visual quality, camera-motion accuracy, object consistency, and ego and exo motion adherence.

Limitations: While the spatial memory used in E^3C enables 3D consistency and camera control, it assumes a mostly static environment and models dynamics primarily through human motion, limiting performance when other objects move or the scene changes over long horizons. E^3C provides motion control for ego and exo humans, but it does not explicitly condition on appearance. As a result, a person’s appearance may drift when they leave the field of view and re-enter.

Acknowledgments

The authors thank the Project Aria team for providing open-source support. The authors also thank Armen Avetisyan, Chris Sweeney, Fan Zhang, Jakob Engel, Nan Yang, Pierre Moulon, Ruo Cheng Wang, Tianwei Shen, Zhao Dong, and Zizhang Li for valuable and insightful discussions.

References

1. Arevalo, J., Solorio, T., Montes-y Gómez, M., González, F.A.: Gated multimodal units for information fusion. arXiv preprint arXiv:1702.01992 (2017) 8, 24
2. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016) 23
3. Bahmani, S., Skorokhodov, I., Qian, G., Siarohin, A., Menapace, W., Tagliasacchi, A., Lindell, D.B., Tulyakov, S.: Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22875–22889 (2025) 4
4. Bahmani, S., Skorokhodov, I., Siarohin, A., Menapace, W., Qian, G., Vasilkovsky, M., Lee, H.Y., Wang, C., Zou, J., Tagliasacchi, A., Lindell, D.B., Tulyakov, S.: Vd3d: Taming large video diffusion transformers for 3d camera control. Proc. ICLR (2025) 4
5. Bai, Y., Tran, D., Bar, A., LeCun, Y., Darrell, T., Malik, J.: Whole-body conditioned egocentric video prediction. arXiv preprint arXiv:2506.21552 (2025) 2, 4, 9, 11, 13, 26
6. Bar, A., Zhou, G., Tran, D., Darrell, T., LeCun, Y.: Navigation world models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15791–15801 (2025) 26

7. Cao, C., Zhou, J., Li, S., Liang, J., Yu, C., Wang, F., Xue, X., Fu, Y.: Uni3c: Unifying precisely 3d-enhanced camera and human motion controls for video generation. In: Proceedings of the SIGGRAPH Asia 2025 Conference Papers. pp. 1–12 (2025) [2](#), [4](#), [7](#), [8](#)
8. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025) [10](#), [27](#)
9. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020) [8](#), [24](#)
10. Chou, G., Zhang, K., Bi, S., Tan, H., Xu, Z., Luan, F., Hariharan, B., Snavely, N.: Generating 3d-consistent videos from unposed internet photos. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 27934–27945 (2025) [4](#)
11. Danier, D., Gao, G., McDonagh, S., Li, C., Bilen, H., Mac Aodha, O.: View-consistent diffusion representations for 3d-consistent video generation. arXiv preprint arXiv:2511.18991 (2025) [4](#)
12. Du, H., Ye, J., Cong, X., Li, R., Ni, J., Agarwal, A., Zhou, Z., Li, Z., Balestrieri, R., Wang, Y.: Videogpa: Distilling geometry priors for 3d-consistent video generation. arXiv preprint arXiv:2601.23286 (2026) [4](#)
13. Engel, J., Somasundaram, K., Goesele, M., Sun, A., Gamino, A., Turner, A., Tatlatof, A., Yuan, A., Souti, B., Meredith, B., Peng, C., Sweeney, C., Wilson, C., Barnes, D., DeTone, D., Caruso, D., Valleroy, D., Ginjaipalli, D., Frost, D., Miller, E., Mueggler, E., Oleinik, E., Zhang, F., Somasundaram, G., Solaira, G., Lanaras, H., Howard-Jenkins, H., Tang, H., Kim, H.J., Rivera, J., Luo, J., Dong, J., Straub, J., Bailey, K., Eckenhoff, K., Ma, L., Pesqueira, L., Schwesinger, M., Monge, M., Yang, N., Charron, N., Raina, N., Parkhi, O., Borschowa, P., Moulon, P., Gupta, P., Mur-Artal, R., Pennington, R., Kulkarni, S., Miglani, S., Gondi, S., Solanki, S., Diener, S., Cheng, S., Green, S., Saarinen, S., Patra, S., Mourikis, T., Whelan, T., Singh, T., Balntas, V., Baiyya, V., Dreewes, W., Pan, X., Lou, Y., Zhao, Y., Mansour, Y., Zou, Y., Lv, Z., Wang, Z., Yan, M., Ren, C., Nardi, R.D., Newcombe, R.: Project aria: A new tool for egocentric multi-modal ai research. arXiv preprint arXiv:2308.13561 (2023) [7](#), [9](#)
14. Escobar, M., Puentes, J., Forigua, C., Pont-Tuset, J., Maninis, K.K., Arbelaez, P.: Egocast: Forecasting egocentric human pose in the wild. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 5831–5841. IEEE (2025) [2](#)
15. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024) [5](#)
16. Fang, I., Chen, Y., Wang, Y., Zhang, J., Zhang, Q., Xu, J., He, X., Gao, W., Su, H., Li, Y., et al.: Egopat3dv2: Predicting 3d action target from 2d egocentric vision for human-robot interaction. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 3036–3043. IEEE (2024) [2](#)
17. Finn, C., Levine, S.: Deep visual foresight for planning robot motion. In: 2017 IEEE international conference on robotics and automation (ICRA). pp. 2786–2793. IEEE (2017) [2](#)
18. Fu, S., Tamir, N., Sundaram, S., Chai, L., Zhang, R., Dekel, T., Isola, P.: Dreamsim: Learning new dimensions of human visual similarity using synthetic data.

- In: *Advances in Neural Information Processing Systems*. vol. 36, pp. 50742–50768 (2023) [13](#)
19. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430 (2021) [9](#), [10](#)
 20. Gu, Q., Lv, Z., Frost, D., Green, S., Straub, J., Sweeney, C.: Egolifter: Open-world 3d segmentation for egocentric perception. In: *European Conference on Computer Vision*. pp. 382–400. Springer (2024) [9](#)
 21. Guzov, V., Jiang, Y., Hong, F., Pons-Moll, G., Newcombe, R., Liu, C.K., Ye, Y., Ma, L.: Hmd 2: Environment-aware motion generation from single egocentric head-mounted device. In: *2025 International Conference on 3D Vision (3DV)*. pp. 1394–1405. IEEE (2025) [9](#)
 22. He, H., Xu, Y., Guo, Y., Wetzstein, G., Dai, B., Li, H., Yang, C.: Cameractrl: Enabling camera control for text-to-video generation. arXiv preprint arXiv:2404.02101 (2024) [2](#), [4](#), [10](#)
 23. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* **30** (2017) [13](#)
 24. Hou, C., Chen, Z.: Training-free camera control for video generation. arXiv preprint arXiv:2406.10126 (2024) [2](#), [4](#)
 25. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. *ICLR* **1**(2), 3 (2022) [10](#)
 26. Huang, J., Zhou, Q., Rabeti, H., Korovko, A., Ling, H., Ren, X., Shen, T., Gao, J., Slepichev, D., Lin, C.H., Ren, J., Xie, K., Biswas, J., Leal-Taixe, L., Fidler, S.: Vipe: Video pose engine for 3d geometric perception. In: *NVIDIA Research Whitepapers* arXiv:2508.10934 (2025) [10](#), [24](#), [27](#)
 27. Huang, T., Zheng, W., Wang, T., Liu, Y., Wang, Z., Wu, J., Jiang, J., Li, H., Lau, R., Zuo, W., et al.: Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation. *ACM Transactions on Graphics (TOG)* **44**(6), 1–15 (2025) [2](#), [4](#)
 28. Jiang, Z., Han, Z., Mao, C., Zhang, J., Pan, Y., Liu, Y.: Vace: All-in-one video creation and editing. arXiv preprint arXiv:2503.07598 (2025) [3](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#), [21](#), [22](#)
 29. Kang, T., Kim, K., Kim, D., Park, M., Hyung, J., Choo, J.: Egox: Egocentric video generation from a single exocentric video. arXiv preprint arXiv:2512.08269 (2025) [5](#)
 30. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (ToG)* **42**(4), 1–14 (2023) [26](#)
 31. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: *International Conference on Learning Representations* (2015) [10](#)
 32. Koh, J.Y., Lee, H., Yang, Y., Baldrige, J., Anderson, P.: Pathdreamer: A world model for indoor navigation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 14738–14748 (2021) [2](#)
 33. Kuhn, H.W.: The hungarian method for the assignment problem. *Naval research logistics quarterly* **2**(1-2), 83–97 (1955) [10](#)
 34. Lee, J., Jung, J., Han, J., Narihira, T., Fukuda, K., Seo, J., Hong, S., Mitsufuji, Y., Kim, S.: 3d scene prompting for scene-consistent camera-controllable video generation. arXiv preprint arXiv:2510.14945 (2025) [4](#)

35. Lee, J., Lee, Y., Kim, J., Kosiorek, A., Choi, S., Teh, Y.W.: Set transformer: A framework for attention-based permutation-invariant neural networks. In: International conference on machine learning. pp. 3744–3753. PMLR (2019) [8](#), [24](#)
36. Li, R., Torr, P., Vedaldi, A., Jakab, T.: Vmem: Consistent interactive video scene generation with surfel-indexed view memory. arXiv preprint arXiv:2506.18903 (2025) [2](#), [4](#), [9](#), [10](#), [11](#), [21](#), [25](#)
37. Li, T., Zheng, G., Jiang, R., Zhan, S., Wu, T., Lu, Y., Lin, Y., Deng, C., Xiong, Y., Chen, M., et al.: Realcam-i2v: Real-world image-to-video generation with interactive complex camera control. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 28785–28796 (2025) [4](#)
38. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022) [5](#), [10](#)
39. Liu, J.W., Mao, W., Xu, Z., Keppo, J., Shou, M.Z.: Exocentric-to-egocentric video generation. Advances in Neural Information Processing Systems **37**, 136149–136172 (2024) [5](#)
40. Luo, H., Zhu, K., Zhai, W., Cao, Y.: Intention-driven ego-to-exo video generation. arXiv preprint arXiv:2403.09194 (2024) [5](#)
41. Luo, M., Xue, Z., Dimakis, A., Grauman, K.: Put myself in your shoes: Lifting the egocentric perspective from exocentric videos. In: European Conference on Computer Vision. pp. 407–425. Springer (2024) [5](#)
42. Ma, L., Ye, Y., Hong, F., Guzov, V., Jiang, Y., Postyeni, R., Pesqueira, L., Gamino, A., Baiyya, V., Kim, H.J., et al.: Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In: European Conference on Computer Vision. pp. 445–465. Springer (2024) [3](#), [9](#)
43. Mahdi, M., Fu, Y., Savov, N., Pan, J., Paudel, D.P., Van Gool, L.: Exo2egosyn: Unlocking foundation video generation models for exocentric-to-egocentric video synthesis. arXiv preprint arXiv:2511.20186 (2025) [5](#)
44. Meta Reality Labs Research: Project aria machine perception services. https://facebookresearch.github.io/projectaria_tools/docs/ARK/mps (2023), accessed: 2026-01-27 [9](#), [10](#), [21](#), [24](#), [26](#)
45. Minderer, M., Gritsenko, A., Houlsby, N.: Scaling open-vocabulary object detection. Advances in Neural Information Processing Systems **36**, 72983–73007 (2023) [10](#), [27](#)
46. Movella: Movella Xsens MVN Link Motion Capture System. <https://www.movella.com/products/motion-capture/xsens-mvn-link> (2023), accessed: 2026-01-27 [9](#)
47. Nerfstudio Team: Splatfacto: Splatting for fast radiance fields. <https://docs.nerf.studio/en/latest/nerfology/methods/splatfacto.html> (2023), part of the Nerfstudio framework [9](#), [10](#), [11](#), [21](#), [26](#)
48. Pallotta, E., Azar, S.M., Doorenbos, L., Ozsoy, S., Iqbal, U., Gall, J.: Egocontrol: Controllable egocentric video generation via 3d full-body poses. arXiv preprint arXiv:2511.18173 (2025) [2](#), [4](#), [9](#), [10](#), [11](#), [13](#), [26](#)
49. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023) [5](#), [6](#)
50. Ren, X., Lu, Y., Cao, T., Gao, R., Huang, S., Sabour, A., Shen, T., Pfaff, T., Wu, J.Z., Chen, R., et al.: Cosmos-drive-dreams: Scalable synthetic driving data generation with world foundation models. arXiv preprint arXiv:2506.09042 (2025) [2](#)

51. Ren, X., Shen, T., Huang, J., Ling, H., Lu, Y., Nimier-David, M., Müller, T., Keller, A., Fidler, S., Gao, J.: Gen3c: 3d-informed world-consistent video generation with precise camera control. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6121–6132 (2025) [2](#), [4](#), [7](#), [9](#), [10](#), [11](#), [21](#), [24](#)
52. Shuai, X., Ding, H., Qin, Z., Luo, H., Ma, X., Tao, D.: Free-form motion control: Controlling the 6d poses of camera and objects in video generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12449–12458 (2025) [4](#)
53. Srivastava, R.K., Greff, K., Schmidhuber, J.: Highway networks. arXiv preprint arXiv:1505.00387 (2015) [8](#), [24](#)
54. Tu, Y., Luo, H., Chen, X., Bai, X., Wang, F., Zhao, H.: Playerone: Egocentric world simulator. arXiv preprint arXiv:2506.09995 (2025) [2](#)
55. Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges. arXiv preprint arXiv:1812.01717 (2018) [10](#)
56. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. *Advances in neural information processing systems* **30** (2017) [8](#), [23](#), [24](#)
57. Voleti, V., Yao, C.H., Boss, M., Letts, A., Pankratz, D., Tochilkin, D., Laforte, C., Rombach, R., Jampani, V.: Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In: European Conference on Computer Vision. pp. 439–457. Springer (2024) [4](#)
58. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025) [3](#), [5](#), [10](#), [22](#)
59. Wang, J., Ye, L., Lu, T., Xiao, J., Zhang, J., Guo, Y., Liu, X., Chellappa, R., Peng, C., Yuille, A., et al.: Evoworld: Evolving panoramic world generation with explicit 3d memory. arXiv preprint arXiv:2510.01183 (2025) [4](#)
60. Wang, X., Zhao, K., Liu, F., Wang, J., Zhao, G., Bao, X., Zhu, Z., Zhang, Y., Wang, X.: Egovid-5m: A large-scale video-action dataset for egocentric video generation. arXiv preprint arXiv:2411.08380 (2024) [4](#)
61. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. *IEEE TIP* **13**(4), 600–612 (2004) [10](#)
62. Wang, Z., Yuan, Z., Wang, X., Li, Y., Chen, T., Xia, M., Luo, P., Shan, Y.: Motionctrl: A unified and flexible motion controller for video generation. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–11 (2024) [2](#), [4](#)
63. Watson, D., Saxena, S., Li, L., Tagliasacchi, A., Fleet, D.J.: Controlling space and time with diffusion models. In: The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24–28, 2025. OpenReview.net (2025), <https://openreview.net/forum?id=d2UrCGtntF> [2](#)
64. Wu, T., Yang, S., Po, R., Xu, Y., Liu, Z., Lin, D., Wetzstein, G.: Video world models with long-term spatial memory. arXiv preprint arXiv:2506.05284 (2025) [2](#), [4](#), [7](#)
65. Xie, Y., Yao, C.H., Voleti, V., Jiang, H., Jampani, V.: Sv4d: Dynamic 3d content generation with multi-frame and multi-view consistency. arXiv preprint arXiv:2407.17470 (2024) [4](#)
66. Xiu, J., Hong, F., Li, Y., Li, M., Wang, W., Han, S., Pan, L., Liu, Z.: Egotwin: Dreaming body and view in first person. arXiv preprint arXiv:2508.13013 (2025) [2](#), [4](#), [9](#), [10](#), [11](#), [13](#)

67. Xu, D., Nie, W., Liu, C., Liu, S., Kautz, J., Wang, Z., Vahdat, A.: Camco: Camera-controllable 3d-consistent image-to-video generation. arXiv preprint arXiv:2406.02509 (2024) [4](#)
68. Xu, J., Huang, Y., Pei, B., Hou, J., Li, Q., Chen, G., Zhang, Y., Feng, R., Xie, W.: Egoexo-gen: Ego-centric video prediction by watching exo-centric videos. arXiv preprint arXiv:2504.11732 (2025) [5](#)
69. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: ViTPose: Simple vision transformer baselines for human pose estimation. In: Advances in Neural Information Processing Systems (2022) [9](#), [10](#)
70. Yang, J., Liu, S., Guo, H., Dong, Y., Zhang, X., Zhang, S., Wang, P., Zhou, Z., Xie, B., Wang, Z., et al.: Egolife: Towards egocentric life assistant. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 28885–28900 (2025) [2](#)
71. Yang, S., Hou, L., Huang, H., Ma, C., Wan, P., Zhang, D., Chen, X., Liao, J.: Direct-a-video: Customized video generation with user-directed camera movement and object motion. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–12 (2024) [2](#), [4](#)
72. Yu, J., Bai, J., Qin, Y., Liu, Q., Wang, X., Wan, P., Zhang, D., Liu, X.: Context as memory: Scene-consistent interactive long video generation with memory retrieval. In: Proceedings of the SIGGRAPH Asia 2025 Conference Papers. pp. 1–11 (2025) [2](#)
73. Yu, W., Yin, S., Easterbrook, S., Garg, A.: Egosim: Egocentric exploration in virtual worlds with multi-modal conditioning. In: The Thirteenth International Conference on Learning Representations (2025) [2](#)
74. Zhang, H., Chen, X., Wang, Y., Liu, X., Wang, Y., Qiao, Y.: 4diffusion: Multi-view video diffusion model for 4d generation. Advances in Neural Information Processing Systems **37**, 15272–15295 (2024) [4](#)
75. Zhang, L., Ye, J., Wang, Y., Zhong, M., Cao, M., Xia, W., Zeng, B., Zhang, Z., Tang, H.: Egoldc: Egocentric video generation with long context diffusion. arXiv preprint arXiv:2512.04515 (2025) [4](#)
76. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023) [8](#)
77. Zhang, Q., Zhai, S., Martin, M.A.B., Miao, K., Toshev, A., Susskind, J., Gu, J.: World-consistent video diffusion with explicit 3d modeling. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21685–21695 (2025) [4](#)
78. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018) [10](#), [13](#)
79. Zhao, J., Wei, F., Liu, Z., Zhang, H., Xu, C., Lu, Y.: Spatia: Video generation with updatable spatial memory. arXiv preprint arXiv:2512.15716 (2025) [4](#)
80. Zheng, G., Li, T., Jiang, R., Lu, Y., Wu, T., Li, X.: Cami2v: Camera-controlled image-to-video diffusion model. arXiv preprint arXiv:2410.15957 (2024) [4](#)
81. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5745–5753 (2019) [8](#), [24](#)

Appendix

A Qualitative Video Results

On the [project webpage](#), we present multiple qualitative results. These include:

- Video results from our full method, E^3C , including visualizations of the input conditions, generated videos, and ground-truth (GT) videos.
- Video results from ablation studies on per-point appearance features, exo human motion control, and ego human motion control.
- Video results from comparisons with baselines, including zero-shot and fine-tuned VACE [28], Splatfacto [47], Gen3C [51], and VMem [36].
- Video results demonstrating scene editing, including object removal, exo people removal, target camera changes, and motion and scene composition.

Please refer to the project webpage for qualitative results and visualizations. In the remainder of this appendix, we provide additional details on the method and experiments.

B Method Details

In the following, we elaborate on implementation details.

B.1 3D Spatial Memory Construction and Rendering

In our implementation, we build the RGB point cloud and the feature point cloud in the spatial memory separately. The two memories use different voxel sizes in voxel pooling, and they are rendered into conditioning videos of different resolutions. Their construction and rendering processes are described below.

RGB point cloud memory. As described in Sec. 3.3, we build the persistent 3D memory from the context frames $\mathbf{x}_{1:m}$ and render it onto the target camera views $\mathbf{T}_{1:n}$. To build the 3D memory, we use the semi-dense points and observations (`semidense_points.csv.gz` and `semidense_observations.csv.gz`) produced by Project Aria MPS [44]. These files jointly provide a 3D point cloud in the world coordinate system and the corresponding observations at each capture timestamp. From the MPS results, we extract the 3D points observed in the context frames $\mathbf{P}^{raw} \in \mathbb{R}^{m \times N_{raw} \times 3}$ (the point cloud from each frame is padded to the same size N_{raw} with NaN values). Note that MPS runs on the grayscale SLAM cameras on Aria glasses, which have a wider field of view than the RGB camera. In our experiments, to ensure fair comparison between baselines that only use RGB inputs, we only keep the points that fall in the RGB camera frustum.

From \mathbf{P}^{raw} , we first aggregate the points over time, remove all NaN entries, and merge duplicate 3D locations. Each point is assigned a color from the context

RGB images; if the same 3D point is observed multiple times, the implementation pools its colors across observations before aggregation. After aggregation, the point set is voxel-downsampled using a voxel size of 0.01 m to obtain a compact memory, which corresponds to \mathbf{P} and \mathbf{C} in Sec. 3.3. This memory is then rendered into each target view by transforming the world points into the target camera frame, projecting them to the image plane, discarding invalid or out-of-bounds projections, and splatting them as pixels or small disks on the rendered image. The renderer uses the pooled point colors.

Feature point cloud memory. In parallel with the RGB point memory, we build the feature point cloud \mathbf{F} in the 3D spatial memory. The context video is first converted to dense per-frame feature maps using the Wan2.1 video encoder [58]. Note that in this encoding process, every context frame is treated as the first frame in a snippet, so there is no temporal downsampling, and the output feature maps have the same temporal resolution as the input video. The video encoder encodes each context frame to a 16-channel latent map at one-eighth spatial resolution.

The MPS point cloud \mathbf{P}^{raw} is then projected into these feature maps, and point features are sampled by bilinear interpolation at the projected locations. The sampled descriptors are pooled in world space with a coarser voxel size of 0.02 m, which is larger than the RGB point-memory voxel size because feature descriptors are smoother and do not require the same geometric fidelity. The voxel feature memory is rendered back to the target camera by projecting voxel centers to the target feature-map grid and splatting the closest visible voxel at each pixel with a simple z-buffer rule.

B.2 Feature Encoder

The rendered feature maps are not passed through the video VAE encoder again. Instead, they are processed by the feature encoder, a separate lightweight 3D convolutional encoder. The feature encoder follows a similar architecture to the Wan2.1 video encoder, but since its input is already at one-eighth spatial resolution, it does not need to further downsample the input. Specifically, the feature encoder uses hidden width 1024, one residual block per stage, channel multipliers [0.25, 0.125, 0.125], two temporal downsampling stages, and no additional spatial downsampling. Its output is a 96-channel tensor with the same shape as the VACE encoder [28], so the rendered feature context can be added directly to the RGB VACE context before entering the diffusion transformer, as mentioned in Sec. 3.3.

The RGB conditioning video and the conditioning feature video play different roles. The RGB control stream starts from full-resolution video, so it uses the standard VACE encoder path [28]. The feature stream starts from already-compressed, low-resolution multi-channel maps, so it uses a separate encoder with a different composition. This separation keeps the RGB pathway close to the pretrained video model while still allowing the rendered 3D feature memory to inject view-consistent appearance cues.

B.3 Ego Pose Encoder

This subsection provides the network details for the persistent ego pose tokens introduced in Sec. 3.4. The encoder maps the target ego body joints $\mathbf{p}_{1:n}^{e,s}$ and wrist trajectories $\mathbf{p}_{1:n}^{e,w}$ to a compact sequence of motion-aware tokens that condition the video DiT through dedicated pose cross-attention.

Input and output. For a target clip of length n with J body joints, the inputs are

$$\mathbf{p}_{1:n}^{e,s} \in \mathbb{R}^{n \times J \times 3}, \quad (5)$$

$$\mathbf{p}_{1:n}^{e,w} \in \mathbb{R}^{n \times 2 \times 4 \times 4}. \quad (6)$$

The encoder outputs a pose-token sequence

$$\mathbf{H}^e \in \mathbb{R}^{(n+K) \times D_{\text{DiT}}}, \quad (7)$$

where K is the number of learned global summary tokens.

Per-joint motion features. Following the main text, the body branch starts from camera-frame 3D joints. For each joint trajectory, we compute first- and second-order temporal differences,

$$\Delta \mathbf{p}_{t,j}^{e,s} = \mathbf{p}_{t,j}^{e,s} - \mathbf{p}_{t-1,j}^{e,s}, \quad (8)$$

$$\Delta^2 \mathbf{p}_{t,j}^{e,s} = \Delta \mathbf{p}_{t,j}^{e,s} - \Delta \mathbf{p}_{t-1,j}^{e,s}, \quad (9)$$

with first-frame replication at the sequence boundary. Let h and p denote the head and pelvis joint indices. For frame t and joint j , the encoder forms

$$\mathbf{u}_{t,j}^e = [\mathbf{p}_{t,j}^{e,s}, \Delta \mathbf{p}_{t,j}^{e,s}, \Delta^2 \mathbf{p}_{t,j}^{e,s}, \mathbf{p}_{t,j}^{e,s} - \mathbf{p}_{t,h}^{e,s}, \mathbf{p}_{t,j}^{e,s} - \mathbf{p}_{t,p}^{e,s}] \in \mathbb{R}^{15}. \quad (10)$$

The first three terms preserve motion in the default camera coordinate frame, while the last two terms express each joint relative to the head and pelvis joints, providing complementary head-centered and pelvis-centered body context. Each $\mathbf{u}_{t,j}^e$ is then linearly projected to a D_{model} -dimensional embedding, summed with a learnable joint-identity embedding, and passed through dropout.

Spatial mixing and frame token pooling. For each frame, the J joint embeddings are processed by n_{spatial} transformer encoder layers [56] that operate across the joint dimension and are shared over time. A learnable pooling query then applies multi-head attention (MHA) [56] over the full joint set, followed by layer normalization (LN) [2], to produce one body token per frame:

$$\mathbf{f}_t^e = \text{LN}\left(\text{MHA}(\mathbf{q}_{\text{pool}}^e, \mathbf{X}_t^e, \mathbf{X}_t^e)\right) \in \mathbb{R}^{D_{\text{model}}}. \quad (11)$$

Wrist 6-DoF branch and gated fusion. For each wrist transform in $\mathbf{p}_{1:n}^{e,w}$, we extract translation $\mathbf{p} \in \mathbb{R}^3$ and a continuous 6D rotation representation $\mathbf{r}_{6D} \in \mathbb{R}^6$ [81] given by the first two columns of the rotation matrix. The per-wrist feature vector is

$$\mathbf{u}_{t,\ell}^w = [\mathbf{p}_{t,\ell}, \Delta\mathbf{p}_{t,\ell}, \Delta^2\mathbf{p}_{t,\ell}, \mathbf{r}_{6D,t,\ell}, \Delta\mathbf{r}_{6D,t,\ell}, \Delta^2\mathbf{r}_{6D,t,\ell}] \in \mathbb{R}^{27}, \quad (12)$$

where $\ell \in \{L, R\}$ indexes the left and right wrists. These two wrist tokens are linearly embedded, augmented with wrist-identity embeddings, and pooled by a second learnable attention query in the style of set attention pooling [35] to produce a wrist token $\mathbf{f}_t^w \in \mathbb{R}^{D_{\text{model}}}$. Body and wrist streams are fused with a learned gate [1, 53]:

$$\mathbf{g}_t = \sigma(W_g[\mathbf{f}_t^e; \mathbf{f}_t^w]), \quad (13)$$

$$\tilde{\mathbf{f}}_t = \mathbf{f}_t^e + \mathbf{g}_t \odot \mathbf{f}_t^w. \quad (14)$$

This gated residual form lets the encoder modulate how much wrist information contributes to each frame token.

Temporal modeling and global summary tokens. The fused frame tokens are augmented with sinusoidal temporal embeddings and processed by an n_{temporal} -layer temporal transformer [56] over $t \in \{1, \dots, n\}$. Next, K learnable global queries attend to the full sequence of frame tokens, following the object-query style used in DETR [9], to extract global summary tokens $\mathbf{G}^e \in \mathbb{R}^{K \times D_{\text{model}}}$. The final output is

$$\bar{\mathbf{H}}^e = [\tilde{\mathbf{f}}_{1:n}; \mathbf{G}^e] \in \mathbb{R}^{(n+K) \times D_{\text{model}}}, \quad (15)$$

$$\mathbf{H}^e = \text{Dropout}(W_{\text{dit}}\bar{\mathbf{H}}^e) + \mathbf{e}_{\text{pose}}. \quad (16)$$

We use $D_{\text{model}} = 512$, $D_{\text{DiT}} = 1536$, $K = 4$, $n_{\text{spatial}} = 2$, and $n_{\text{temporal}} = 2$.

Practical note. Compared with using only rasterized ego pose drawings in \mathbf{c}^{ctrl} , these persistent tokens preserve camera-frame 3D joint kinematics and wrist orientation cues in token space, which helps maintain motion control under self-occlusion, truncation, and out-of-view body motion.

C Experiment Details

C.1 Baselines

Gen3C [51]. We use the official Gen3C codebase and the released pretrained checkpoint. Following the official setup, we use VIPE [26] to obtain dense depth maps for each input context frame, where the camera extrinsics and intrinsics from MPS [44] are used as the initialization for VIPE optimization. Gen3C warps the context frames into the target camera views according to the estimated dense depth maps and uses the warped images as conditioning for video generation.

Algorithm 1 Forward pass of the ego pose encoder

-
- Require:** Body joints $\mathbf{p}_{1:n}^{e,s} \in \mathbb{R}^{n \times J \times 3}$, wrist poses $\mathbf{p}_{1:n}^{e,w} \in \mathbb{R}^{n \times 2 \times 4 \times 4}$
Require: Head index h , pelvis index p
Ensure: Pose tokens $\mathbf{H}^e \in \mathbb{R}^{(n+K) \times D_{\text{DiT}}}$
- 1: Compute $\Delta \mathbf{p}^{e,s}$ and $\Delta^2 \mathbf{p}^{e,s}$ with first-frame replication
 - 2: Build per-joint features $\mathbf{u}^e = [\mathbf{p}^{e,s}, \Delta \mathbf{p}^{e,s}, \Delta^2 \mathbf{p}^{e,s}, \mathbf{p}^{e,s} - \mathbf{p}_{:,h}^{e,s}, \mathbf{p}^{e,s} - \mathbf{p}_{:,p}^{e,s}]$
 - 3: Embed joints: $\mathbf{X}^e \leftarrow \text{Linear}(\mathbf{u}^e) + \text{JointIDEmbed}$
 - 4: Apply spatial transformer stack over the joint dimension
 - 5: Frame pooling: $\mathbf{f}_{1:n}^e \leftarrow \text{AttnPool}(\mathbf{X}^e)$
 - 6: Extract per-wrist 27-D features from translation, 6D rotation, and temporal derivatives
 - 7: Wrist embedding + wrist-ID + wrist attention pooling to get $\mathbf{f}_{1:n}^w$
 - 8: Compute gate $\mathbf{g}_{1:n} \leftarrow \sigma(W_g[\mathbf{f}_{1:n}^e; \mathbf{f}_{1:n}^w])$
 - 9: Fuse: $\tilde{\mathbf{f}}_{1:n} \leftarrow \mathbf{f}_{1:n}^e + \mathbf{g}_{1:n} \odot \mathbf{f}_{1:n}^w$
 - 10: Add sinusoidal temporal embeddings to $\tilde{\mathbf{f}}_{1:n}$
 - 11: Apply temporal transformer stack over time
 - 12: Global extraction: $\mathbf{Q}_{\text{global}}$ cross-attends to $\tilde{\mathbf{f}}_{1:n}$ to produce K global tokens \mathbf{G}^e
 - 13: Concatenate $\bar{\mathbf{H}}^e \leftarrow [\mathbf{f}_{1:n}^e; \mathbf{G}^e]$
 - 14: Project $\bar{\mathbf{H}}^e$ to D_{DiT} , add pose embedding, apply output dropout
 - 15: **return** \mathbf{H}^e
-

The official inference script generates a 121-frame video chunk at a resolution of 1280×704 in each inference pass. To align with this setup, we pad our 80-frame target camera trajectory to 121 frames by repeating the last camera poses. We also resize the input context video and target camera intrinsics to 1280×704 before inference and resize the generated video back to 512×512 for evaluation and visualization. Empirically, we find that this reduces the domain gap between the training data of Gen3C and our evaluation setting, leading to better results for Gen3C.

We adapt the ‘‘video generation from multiview images’’ example script provided by the authors. The script selects only two context frames that jointly have the maximum overlap with the entire target camera trajectory and uses them to condition video generation. We find that this *per-trajectory* selection provides insufficient view coverage for large view changes in our evaluation. Therefore, we modify the script to select two context frames for each target viewpoint independently (*per-frame* selection), which significantly improves the performance of Gen3C in our setting.

VMem [36]. We use the official VMem codebase and the released pretrained checkpoint. The released VMem model takes 4 context frames as input and generates 4 target frames in each inference pass. Therefore, we generate the full 80-frame video autoregressively using 20 inference passes.

In each inference pass, 4 context frames are selected from VMem’s surfel-indexed memory and used to condition the video generator. We conduct experiments ablating whether the generated frames from each inference pass are

inserted back into the memory for future generation, and how the camera poses are normalized at each inference pass. However, we find that these factors minimally affect VMem’s performance in our setting.

The input is resized to 576×576 for model inference, and the generated video is resized back to 512×512 for evaluation and visualization.

Splatfacto [47]. We use the 3D Gaussian splatting-based [30] method Splatfacto provided by NerfStudio [47]. In our experiments, Splatfacto uses the camera poses and intrinsics estimated by MPS [44] and initializes the 3D Gaussians using the MPS semi-dense point cloud. One 3D Gaussian splatting model is trained and evaluated individually on each snippet. We use the default training settings provided by NerfStudio with 30,000 training iterations.

Comparison with PEVA [5] and EgoControl [48]. Since PEVA and EgoControl have not released their code or model weights, we reproduce their evaluation setup to obtain the comparison results reported in Tab. 2. Specifically, we obtain the full list of 300 snippets used in the PEVA evaluation from the authors. As their data split differs from ours, we train a separate E^3C model on sequences from the Nymeria dataset that are not included in the PEVA evaluation set and evaluate it on the PEVA snippets.

To further align the evaluation protocol, for this separately trained model, we disable the devignetting and undistortion preprocessing and use the original fisheye images. As mentioned in Sec. 4.3, this evaluation uses a 15-frame context window from the past 3.75 seconds and evaluates a single generated frame 2 seconds into the future.

Following the suggestion of the PEVA authors, we use the evaluation script from NWM⁴ [6] to compute the LPIPS, DreamSim, and FID metrics reported in Tab. 2.

C.2 Evaluation Metrics

Notation. Let the GT target video be

$$\mathbf{y}_{1:n} \in \mathbb{R}^{n \times H \times W \times 3},$$

and let the generated video be

$$\hat{\mathbf{y}}_{1:n} \in \mathbb{R}^{n \times H \times W \times 3}.$$

When spatial resolutions differ, generated frames are resized to the GT resolution before comparison. Let $\mathbf{T}_{1:n} = \{T_t\}_{t=1}^n$, $T_t = (R_t, \mathbf{t}_t) \in SE(3)$, denote the GT camera trajectory, and let $\mathbf{p}_{1:n}^x$ denote the GT exocentric 2D human poses.

⁴ https://github.com/facebookresearch/nwm/blob/main/isolated_nwm_eval.py

Object Consistency (Obj-F1, Obj-mIoU). We run OWLv2 [45] on each aligned frame of $\mathbf{y}_{1:n}$ and $\hat{\mathbf{y}}_{1:n}$, using the default indoor prompt set and a score threshold of 0.2. This yields detection sets

$$\mathcal{D}_t^{\text{obj}}, \hat{\mathcal{D}}_t^{\text{obj}}.$$

Each detection contains a bounding box, and GT/generated boxes are matched independently in each frame by Hungarian assignment on cost $1 - \text{IoU}$, keeping only matches with IoU at least $\tau_{\text{obj}} = 0.5$. Obj-F1 is then computed from the aggregate TP/FP/FN counts over all aligned frames. Obj-mIoU is the mean of frame-wise matched-box IoU scores,

$$m_t = \begin{cases} \frac{1}{|\mathcal{M}_t|} \sum_{(i,j) \in \mathcal{M}_t} \text{IoU}(\mathbf{b}_i, \hat{\mathbf{b}}_j), & |\mathcal{M}_t| > 0 \\ 0, & |\mathcal{M}_t| = 0, |\mathcal{D}_t^{\text{obj}}| + |\hat{\mathcal{D}}_t^{\text{obj}}| > 0, \\ 1, & |\mathcal{D}_t^{\text{obj}}| = |\hat{\mathcal{D}}_t^{\text{obj}}| = 0 \end{cases}$$

where \mathcal{M}_t is the filtered match set at frame t . The reported score is

$$\text{Obj-mIoU} = \frac{1}{n} \sum_{t=1}^n m_t.$$

Camera Motion Control (TErr, RErr). For camera control, we estimate camera poses from generated frames $\hat{\mathbf{y}}_{1:n}$ using VIPE [26], yielding $\hat{T}_t = (\hat{R}_t, \hat{\mathbf{t}}_t)$. We then align $\hat{T}_{1:n}$ to the GT trajectory $\mathbf{T}_{1:n}$ using a global Sim(3) transform

$$S = (s, R_S, \mathbf{t}_S), \quad (\hat{R}'_t, \hat{\mathbf{t}}'_t) = (R_S \hat{R}_t, s R_S \hat{\mathbf{t}}_t + \mathbf{t}_S),$$

with S fit once over the whole trajectory. Per-frame translation and rotation errors are

$$e_t^{\text{trans}} = \|\hat{\mathbf{t}}'_t - \mathbf{t}_t\|_2, \\ e_t^{\text{rot}} = \frac{180}{\pi} \arccos \left(\text{clip} \left(\frac{\text{tr}((\hat{R}'_t)^\top R_t) - 1}{2}, -1, 1 \right) \right).$$

The reported metrics are frame averages,

$$\text{TErr} = \frac{1}{n} \sum_{t=1}^n e_t^{\text{trans}}, \quad \text{RErr} = \frac{1}{n} \sum_{t=1}^n e_t^{\text{rot}}.$$

In our table, TErr is displayed in centimeters.

Ego Hand Control (Hand-F1, Hand-mIoU). We run SAM3 [8] on each aligned frame of $\mathbf{y}_{1:n}$ and $\hat{\mathbf{y}}_{1:n}$. Since the ego human in the Nymeria dataset always wears a mocap suit that covers part of their hands and wrists, we use the text prompts “hand” and “glove” for SAM3. We use a confidence threshold

of 0.3, and masks smaller than 1536 pixels are filtered out to remove detected hands on exo people. This produces binary hand masks

$$\mathbf{M}_t^{\text{hand}}, \widehat{\mathbf{M}}_t^{\text{hand}} \in \{0, 1\}^{H \times W}$$

for the GT and generated frames, respectively. Hand presence indicators are

$$z_t = \mathbb{1} \left[\sum \mathbf{M}_t^{\text{hand}} > 0 \right], \quad \hat{z}_t = \mathbb{1} \left[\sum \widehat{\mathbf{M}}_t^{\text{hand}} > 0 \right].$$

Hand-F1 is computed from the aggregate TP/FP/FN counts over (z_t, \hat{z}_t) across all aligned frames. For mask overlap, the per-frame hand IoU is

$$\text{IoU}_t^{\text{hand}} = \begin{cases} \frac{|\mathbf{M}_t^{\text{hand}} \cap \widehat{\mathbf{M}}_t^{\text{hand}}|}{|\mathbf{M}_t^{\text{hand}} \cup \widehat{\mathbf{M}}_t^{\text{hand}}|}, & |\mathbf{M}_t^{\text{hand}} \cup \widehat{\mathbf{M}}_t^{\text{hand}}| > 0 \\ 1, & \text{otherwise} \end{cases}.$$

The reported Hand-mIoU is the mean of $\text{IoU}_t^{\text{hand}}$ over all n aligned frames, so empty-empty frames contribute 1.

Exo Human Gesture Control (Exo-F1, PCK@10%). From $\mathbf{y}_{1:n}$ and $\hat{\mathbf{y}}_{1:n}$, we extract per-frame exo person boxes and keypoints using YOLO-X + ViTPose. The frames are resized to 640×640 for inference and the resulting boxes/keypoints are mapped back to the original image resolution. Detections whose average keypoint confidence is below 0.5 are discarded. This yields

$$(\mathcal{B}_t, \mathcal{K}_t), (\widehat{\mathcal{B}}_t, \widehat{\mathcal{K}}_t).$$

For each frame t , we match \mathcal{B}_t and $\widehat{\mathcal{B}}_t$ by Hungarian assignment on cost $1 - \text{IoU}$, retaining matches with $\text{IoU} \geq \tau_{\text{exo}}$, where $\tau_{\text{exo}} = 0.2$. Exo-F1 is computed from the aggregate TP/FP/FN counts over all aligned frames.

Pose accuracy is computed only on matched person pairs. For a matched pair m and keypoint j , let A_m be the GT box area, with GT keypoint $\mathbf{k}_{m,j}$ and generated keypoint $\widehat{\mathbf{k}}_{m,j}$. The normalized error is

$$e_{m,j} = \frac{\|\widehat{\mathbf{k}}_{m,j} - \mathbf{k}_{m,j}\|_2}{\sqrt{A_m}}.$$

The reported normalized MPJPE averages $e_{m,j}$ over all valid matched keypoints. PCK@10% is

$$\text{PCK@10\%} = \frac{1}{|\mathcal{V}|} \sum_{(m,j) \in \mathcal{V}} \mathbb{1}[e_{m,j} < 0.10].$$

If there are no valid matched keypoints, the pose metrics are undefined and are not considered in the normalization. Therefore, as noted in the main text, PCK@10% should be interpreted together with Exo-F1.