Grounded-Exo2Ego: Structured Semantic Grounding for Robust Exocentric-to-Egocentric Video Generation

Shengze Wang **Michael Stengel** **Tianye Li** **Seonwook Park**
Amrita Mazumdar **Koki Nagano** **Alex Trevithick** **Shalini De Mello**
NVIDIA
{shengzew, mstengel, tianyel, seonwookp,
amritam, knagano, atrevithick, shalinig}@nvidia.com
https://research.nvidia.com/labs/amri/projects/grounded-exo2ego/

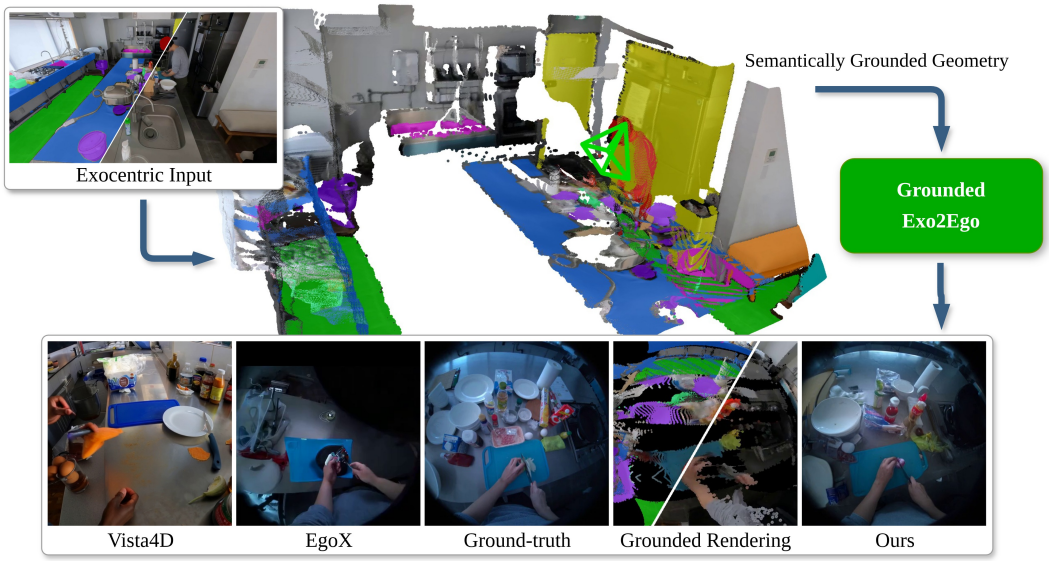
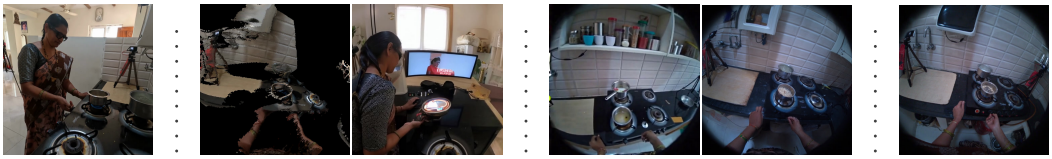


Figure 1: **Grounded-Exo2Ego.** From a single exocentric video (top left), we lift the scene into a semantically grounded 3D reconstruction that anchors a dual-branch video diffusion model. The generated egocentric views (bottom) are substantially more accurate than those from existing methods.

Abstract

Generating egocentric video from a single exocentric video is an emerging and important topic for AR/VR and physical AI. Compared with conventional novel view synthesis, exo-to-ego generation is a significantly harder task because the standard geometric conditioning becomes highly unreliable under extreme view changes and large unobservable regions. We present **Grounded-Exo2Ego**, a principled framework that addresses these challenges at both the architectural and data levels. Architecturally, Grounded-Exo2Ego is a dual-branch video diffusion model that couples a *geometric anchoring* branch, which conditions the generation on the rendering of a 3D reconstruction, with a novel *semantic grounding* branch, which goes beyond the prevailing geometry-based approach and improves quality by synthesizing challenging regions based on object-level context. Additionally, we found that the overlooked issue of camera-reconstruction misalignment severely undermines exo-to-ego learning. We thus introduce a camera re-localization algorithm that resolves this issue and substantially improves quality across all metrics. We further develop a fully automated synthetic data engine that generates and

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



(a) Exo input (b) 3D at ego (c) Vista4D at ego (d) EgoX (e) Ours (f) GT

Figure 2: **Exo-to-Ego is inherently more difficult than general view synthesis.** State-of-the-art methods, *e.g.* Vista4D [20], often render monocular 3D reconstruction of the input view (a) from a novel view to produce the conditioning video. However, such renderings are riddled with holes and corrupted geometry at the ego view (b), leading to bad results (c). Recent exo-to-ego methods [16] achieve better results (d) but fail to recover the exact viewpoint and content.

renders rigged 3D characters in procedurally generated environments. Evaluation on the challenging EgoExo4D dataset shows that our method outperforms recent state-of-the-art approaches by large margins across all metrics. Detailed ablations validate improvements from each of our contributions at both the data and architectural level.

1 Introduction

Imagine if one could turn any third-person video into a first-person view. The vast archive of internet videos would become a source of egocentric training data at unparalleled scale and diversity, providing the first-person supervision that robots and embodied agents urgently need. Important moments from our own lives, currently frozen in third-person recordings, could be re-lived in VR. Every tutorial, surgical demonstration, or sports match could be experienced through the eyes of the performer. This is what exo-to-ego view synthesis aims to achieve. Yet despite rapid progress in general view synthesis, exo-to-ego view synthesis remains practically out of reach for current methods. Closing this gap requires treating exo-to-ego as a problem in its own right and solving it systematically across both data and architecture, rather than transferring machinery built for general view synthesis.

Exo-to-ego is fundamentally harder than conventional novel view synthesis (NVS) for two coupled reasons. **(i) Geometric conditioning is unreliable.** Modern view synthesis methods are often conditioned on renderings of the scene's 3D reconstructions, obtained through monocular depth estimation on the input video. In exo-to-ego, however, these signals break down for two compounding reasons: monocular depth inaccuracies produce stretched surfaces that appear at the wrong positions in the ego view (Figure 2, b), and large portions of the ego view are simply not observed from the exo viewpoint. These two issues leave the ego rendering riddled with distortions and large holes (Figure 2, b, c). Even purpose-built exo-to-ego models fail (e.g. EgoX [16]; Figure 2, d) when relying solely on geometric conditioning. These challenging regions need to be synthesized consistently with the surrounding scene based on semantic context. **(ii) Training data is fragile.** Monocular reconstructions are not aligned with the real-world geometry because of non-rigid distortions and scale/shift ambiguities. As a result, groundtruth cameras calibrated to real-world geometry are also misaligned to monocular reconstructions. Therefore, when existing methods use groundtruth ego camera to render monocular reconstructions during training, much of the rendering (geometric conditioning) is significantly misaligned to the groundtruth during training. This severe misalignment also necessitates heavy filtering, drastically reducing the available training data. For example, EgoX [16] discards most training samples while the ones left are still misaligned to the groundtruth.

We introduce **Grounded-Exo2Ego** (Figure 1), a principled framework designed to tackle these challenges at both the *architectural* and the *data* levels. Architecturally, Grounded-Exo2Ego is a **dual-branch** video diffusion model (Figure 3) built around two complementary signals. The *geometric anchoring* branch helps the model determine the general structure via ego-view renderings of the scene reconstructions. Going beyond the prevailing geometric-conditioned approach, we introduce the *semantic grounding* branch that further improves object-level generation by anchoring object-level semantics to the scene geometry. Ego-view renderings of the reconstruction are often so heavily distorted or occluded that objects become unrecognizable. Object-level semantics addresses this problem by providing the model information about what each region should depict, helping

it restore the correct appearance and location of each object. At the *data* level, we introduce camera re-localization that improves learning by reducing misalignment between rendering and groundtruth video, and we further develop a fully automated synthetic data engine that provides accurately annotated training data. As noted above, monocular reconstructions of exo-view videos are incompatible with the ground-truth ego camera calibrated in the real world. To improve rendering-groundtruth alignment and avoid wasteful filtering, we re-localize the ego camera inside the inaccurate reconstruction so that the renderings and ground-truth videos align well for learning. Beyond real data, our synthetic data generation engine sidesteps the reconstruction quality and alignment issues entirely. Its core is a fully automated data engine that produces animatable 3D humans performing realistic actions in randomly generated environments. Besides providing clean training signals, our data pipeline also fills in the gap for the lack of human-human egocentric interaction data by rendering interactions from all participants' perspectives.

To summarize our work:

- We propose **Grounded-Exo2Ego**, a *dual-branch* video diffusion model that couples a *geometric anchoring* branch with a novel *semantic grounding* branch, which improves generation quality by grounding object-level context onto the 3D reconstruction.
- We find that the training of existing methods is undermined by misalignment between the geometric anchoring (rendering) and ground-truth videos. We thus introduce a camera re-localzation stage that reduces this alignment and notably improves synthesis accuracy.
- Extensive experiments on the challenging EgoExo4D [11] benchmark demonstrate that **Grounded-Exo2Ego** significantly outperforms recent state-of-the-art exo-to-ego generation methods across all metrics. Detailed ablations validate the contributions from our improvements at both the data-level and the architecture-level.

2 Related work

Exo-to-ego video generation. We study single-exo-video to ego-video synthesis: given an exocentric video and a target ego camera trajectory, the model must generate the corresponding first-person video. Camera intrinsics and extrinsics are used when available and estimated otherwise. EgoExo4D [11] opened this line of work by providing synchronized ego-exo recordings. Several follow-ups target related but distinct settings: Exo2Ego-V [22] and Exo2EgoSyn [23] synthesize ego videos from *four* synchronized exo views – Exo2EgoSyn additionally injects pose-aware multi-view conditioning – and EgoWorld [25] addresses exo-to-ego *image* translation , not video. To our knowledge, the only prior work in our exact single-exo-video to ego-video setting is EgoX [16], which combines an ego-view geometric prior and the exo video via width-wise latent concatenation and uses geometry-guided attention to encourage dense cross-view correspondence . EgoX relies primarily on geometry rendered from monocular depth, which becomes unreliable wherever ego-view content is occluded or disoccluded in the exo view. We keep the single-video setting and add two missing ingredients: more reliable ego re-localization for the real-data pipeline, and object-level semantic grounding that tells the generator what corrupted or missing ego-view regions should depict.

Camera-controlled video and dynamic view synthesis. A second line of relevant work uses diffusion models for novel-view or camera-controlled video generation. One thread conditions the generator on *point-cloud renderings* of the scene at the target view: GEN3C [28], Uni3C [6], and Lyra [2] lift the source view to a 3D point cloud and rasterize it at the target camera as the conditioning image, while Vista4D [20] extends this design to dynamic 4D point clouds for re-shooting moving scenes. A second thread conditions the generator on the *camera trajectory alone*, without an explicit 3D scaffold: TrajectoryCrafter [36], ReCamMaster [4], SynCamMaster [3], and Plenoptic Video Generation [10] inject camera intrinsics and extrinsics into the diffusion model, while PRoPE [18] and UCPE [37] reformulate camera control as a relative positional encoding for video tokens. We compare against representatives of both threads as baselines in §4.

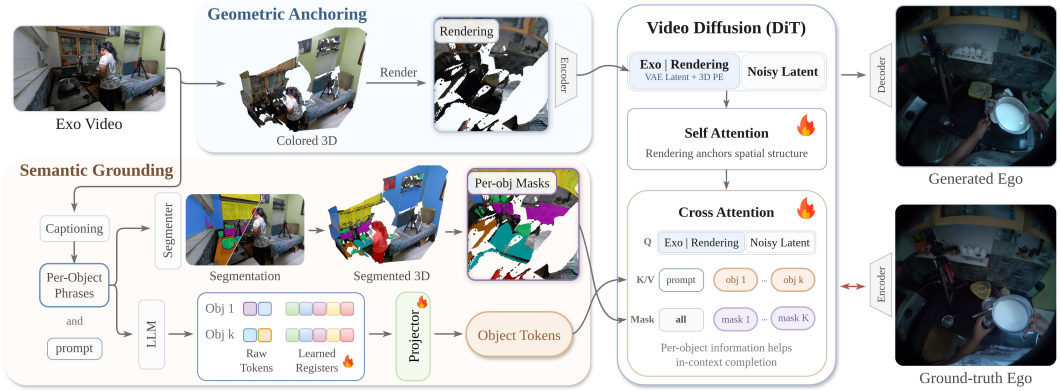


Figure 3: **Method overview.** *Geometric anchoring* (top): Exocentric color and depth are lifted to a 3D reconstruction and rendered at the ego pose. *Semantic grounding* (bottom): Per-object context is extracted as segmentation masks and object tokens. The segmentation masks are reprojected into the ego-view as used as attention masks during cross-attention. The per-object masks encourages noisy tokens to attend to object information relevant to the specific region, providing spatial structure that grounds semantics onto pixels.

3 Method

3.1 Problem Setting

Single-video exocentric-to-egocentric (exo-to-ego) view synthesis takes as input an exocentric (exo) video $\mathbf{V}^{\text{exo}} \in \mathbb{R}^{F \times H_e \times W_e \times 3}$, the exo camera intrinsics $\mathbf{K}^{\text{exo}}$, the egocentric (ego) camera intrinsics $\mathbf{K}^{\text{ego}}$, and a target ego camera trajectory $\{\mathbf{T}_f^{\text{ego}}\}_{f=1}^{F}$. The goal is to generate an egocentric video $\hat{\mathbf{V}}^{\text{ego}} \in \mathbb{R}^{F \times H \times W \times 3}$ consistent with the target trajectory.

3.2 Preliminaries

We build upon a pre-trained latent video generative model [12]. Let $\mathcal{E}$ and $\mathcal{D}$ denote the pre-trained VAE encoder and decoder. We first obtain the clean latent of the ground-truth ego video $\mathbf{z}_0 = \mathcal{E}(\bar{\mathbf{V}}^{\text{ego}})$. Following the flow-matching formulation [21], we define a probability path between a standard Gaussian noise latent $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ at $t = 1$ and the data distribution at $t = 0$. For a sampled time $t \in [0, 1]$, the interpolated latent is constructed as $\mathbf{z}_t = t\mathbf{z}_0 + (1-t)\boldsymbol{\epsilon}$. The target velocity field along this path is defined as the derivative with respect to time $\mathbf{u}_t = \mathbf{z}_0 - \boldsymbol{\epsilon}$. The pre-trained Diffusion Transformer (DiT) [12] backbone, f_θ, learns to predict this velocity field conditioned on signals $\mathbf{c}$ by minimizing the objective $\mathcal{L}_{\text{FM}} = \mathbb{E}_{t,\mathbf{z}_0,\boldsymbol{\epsilon},\mathbf{c}}\left[\|f_\theta(\mathbf{z}_t, t, \mathbf{c}) - \mathbf{u}_t\|_2^2\right]$. During inference, we sample Gaussian noise $\mathbf{z}_{t=0} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. We obtain the clean latent $\hat{\mathbf{z}}_0$ by continuously integrating the model predicted velocity field f_θ forward from $t = 0$ to $t = 1$. This generated latent is then decoded into the output video $\hat{\mathbf{V}}^{\text{ego}} = \mathcal{D}(\hat{\mathbf{z}}_0)$.

3.3 Overview

Figure 3 presents an overview of our dual-branch architecture for exo-to-ego video generation. The *geometric anchoring* branch (§3.4) lifts the exo view into 3D and renders it from the target ego camera, giving the DiT a spatial scaffold for layout and camera motion. The *semantic grounding* branch (§3.5) extracts object phrases and masks from the exo view, reprojects the masks into the ego view, and uses them to route object-specific text tokens to relevant video tokens during cross-attention. Both branches are made reliable by our real and synthetic data pipelines (§3.6).

3.4 Geometric anchoring

Geometric anchoring is the primary conditioning mechanism in state-of-the-art view-synthesis [20, 16, 28, 6]. We adopt this strategy, reconstructing the scene in 3D, rendering it from the ego view, and

injecting it into the DiT. Given the exo-video depth and ego camera pose from our real-data pipeline (§3.6), we lift each exo pixel $\mathbf{p} = (u, v, 1)^\top$ at frame f into a 3D where $\mathbf{X}_f^{\text{exo}}$:

$$\mathbf{X}_f^{\text{exo}}(u, v) = \mathbf{D}_f^{\text{exo}}(u, v)\,(\mathbf{K}^{\text{exo}})^{-1}\mathbf{p}. \tag{1}$$

$\mathbf{D}_f^{\text{exo}}$ is the depth at pixel (u, v) in the f-th exo frame. We then rasterize the 3D geometry at the egocentric camera using `nvdiffrast` [17], yielding the egocentric rendering $\tilde{\mathbf{V}}_f^{\text{ego}}$. The pre-trained VAE, $\mathcal{E}$, encodes both videos into latent token sequences for the DiT:

$$\mathbf{z}^{\text{exo}} = \mathcal{E}(\mathbf{V}^{\text{exo}}), \qquad \tilde{\mathbf{z}}^{\text{ego}} = \mathcal{E}(\tilde{\mathbf{V}}^{\text{ego}}). \tag{2}$$

Here, $\tilde{\mathbf{z}}^{\text{ego}}$ serves as the geometric anchor for the overall image structure. To inject this condition, prior work [16] channel-concatenates the render with the noisy latent $\mathbf{z}_t$. This sometimes leads to large blank regions where the renderings are empty because invalid patches can become distracting signals despite the presence of an occupancy mask. Instead, following IC-LoRA [13], we append $\tilde{\mathbf{z}}^{\text{ego}}$ as separate tokens and we mask out tokens from near-blank regions. $\tilde{\mathbf{z}}^{\text{ego}}_{\text{masked}}$ uses the same (f, x, y) RoPE [30] coordinates as the $\mathbf{z}_t$, thus preserving the geometric layout while safely deactivating unwanted near-blank tokens. Exocentric video tokens $\mathbf{z}^{\text{exo}}$ are similarly concatenated, although their RoPE coordinates are shifted by the width of the ego video. The model then generates the output ego video $\hat{\mathbf{V}}^{\text{ego}}$ by iteratively denoising the noisy latent $\mathbf{z}_t$ as described in §3.2.

3.5 Semantic grounding

Exo-to-ego view translation often requires extreme viewpoint changes, and geometric conditioning alone tends to produce errors when there are large distortions and holes. To synthesize plausible scenes from corrupted geometric signals, the model must know what kinds of objects to render and where to render them, *i.e.* semantic information must be enriched with spatial structure. We achieve structured semantic grounding by extracting rich semantic information from the exo-view and then use it to spatially guide text-to-video cross-attention.

Extracting Rich Semantic Information. As shown in the bottom branch of Figure 3, we deploy a VLM [35] to caption the scene into a descriptive scene prompt, including general caption, object description phrases, and spatial relationships. Then, the VLM deduplicates object candidates (for example, "bike wheels" and "bicycle wheels" become one candidate), and the surviving candidates are ranked by relevance to what the camera wearer is looking at and interacting with. In practice, we select the top $K{=}10$ object candidates. This cap is far less restrictive than it appears, for two reasons. First, each description is passed to the segmentation model as a text prompt, which returns all instances matching that phrase, so a single slot can cover a whole row of bike wheels rather than one wheel. The cap bounds the number of distinct object descriptions, not the number of segmented regions. Second, ten descriptions empirically cover most of the image area.

We then use an LLM [31] to encode the scene prompt and each per-object phrase into separate packets of text tokens, which are then projected to the latent space of the video DiT. The K token packets for the K objects are each padded to a fixed length of $L{=}10$ tokens with a set of learned register tokens shared across objects. Each of the K padded token packets is then transformed by a trainable projector into the latent space of the DiT cross-attention, producing the final token packet $\mathbf{O}_k$ for object k. The fixed length and shared learned registers give every object a stable key/value footprint regardless of phrase length.

Semantic Grounding. We reproject the exo segmentation mask $\mathbf{M}_k^{\text{exo}}$ for object k into the ego view to form $\mathbf{M}_k^{\text{ego}}$ for frame f:

$$\mathbf{M}_k^{\text{ego}}(\mathbf{q}) = \begin{cases} \mathbf{M}_k^{\text{exo}}(\mathbf{U}_f(\mathbf{q}))\,, & \text{if } \mathbf{D}_f(\mathbf{q}) > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

In the original text-to-video cross-attention layers, video tokens are used as queries and the scene's full caption prompt as key/value pairs. We enrich this mechanism by adding spatially structured attention patterns between per-object token packets $\{\mathbf{O}_k\}_{k=1}^K$ and the relevant video tokens to denoise (Figure 3, right). Let $a_{n,k} \in \{0, 1\}$ be the gating value for visual query token n and object k. For an ego query token, $a_{n,k}$ comes from the reprojected ego mask $\mathbf{M}_k^{\text{ego}}$, and for an exo query token from

the exo mask $\mathbf{M}_k^{\text{exo}}$. The cross-attention score from query token n to object-token (k, m) is gated by $a_{n,k}$:

$$A^{\text{obj}}_{n,(k,m)} = \begin{cases} 0, & a_{n,k} = 1, \\ -\infty, & a_{n,k} = 0. \end{cases} \tag{4}$$

Only visual tokens inside object k's region attends to k's packet of text tokens, and every other token is blocked. In this way, the gating masks facilitate spatial grounding by allowing object semantics to stay in the original text latent space of the pretrained backbone. We also explored other approaches such as injecting DINOv3 features [29]), directly spatializing the object tokens into feature maps, etc. We found these approaches to be far less effective compared to our design.

Due to practical limits on compute and storage, we only compute segmentation masks for the first frame and broadcast the same mask across time, i.e., $f = 0$ in Eq. 3. We observe that this approach already yields strong improvement (Table 1). Full video segmentation might yield further gains at higher cost. Additionally, during this grounding step, cross-attention to the full scene prompt is unmodified, *i.e.* all visual tokens still attend to the scene prompt.

3.6 Real and Synthetic Data Pipelines

Ground-truth ego poses are calibrated to real-world geometry and can therefore yield misaligned renderings of monocular reconstructions. We address this supervision mismatch through camera re-localization and complement real videos with accurately annotated synthetic pairs.

Ego camera re-localization for real data. We reconstruct each exo video using MoGe2 [33] and obtain a metric reference using Depth-Anything-3 (DA3) [19] with calibrated exo cameras. We then calculate a non-rigid alignment from the ground-truth ego poses to MoGe2's frame, refining camera depth while keeping rotations fixed. This step ensures that the ego camera is positioned appropriately in the MoGe2, but it does not repair the inaccurate reconstruction. Notice that DA3 is used only to derive the alignment transformation and is never a model input. Additionally, we use the original MoGe2 geometry because it exposes the generator to the reconstruction errors present at inference. Appendix E.1 details the alignment and displacement statistics; Table 2 evaluates its contribution.

Automated synthetic data engine. Our synthetic engine generates accurately annotated exo/ego pairs, allowing semantic grounding to be learned without reconstruction and alignment errors. We generate, rig, and animate 3D characters using interaction motions from Embody3D [24]. We render them in Infinigen [26] interiors, producing over 20K training clips. Appendix E.2 details asset generation, rigging, camera sampling, and rendering.

4 Experiments

4.1 Experimental setup

Dataset. We follow EgoX [16] to train and evaluate on EgoExo4D [11] using the *Unseen* (never seen during training) and *New Action* splits (renamed from "Seen" in EgoX, new actions but in environments seen during training). EgoExo4D is a standard benchmark with in-the-wild corpus captured across several countries rather than a studio. It spans large and fast body motion, a wide range of tasks (such as bike repair, cooking, soccer, and health procedures), and diverse environments (such as homes, sports fields, parking lots, and conference rooms). This in-the-wild benchmark thus evaluates methods under imperfect geometry, segmentation masks, and camera poses.

Implementation details. We use LTX-2.3 [12] as our base video diffusion model with 256 trainable LoRA ranks on self-attention and cross-attention layers. We train on 64 NVIDIA H100 GPUs with mini-batches of 1 on each GPU, learning rate 5×10^{-5}, and half precision. We use F=33, H=448, W=448, and $K\leq$10 objects per clip, each carrying L=10 tokens.

Evaluation metrics. We report three groups, similar to EgoX: **Image Metrics** (PSNR, SSIM, LPIPS, CLIP-I), **Video Metrics** (FVD [32] and Temporal-LPIPS (T-LPIPS) [9]), and **Object Metrics** (Location Error, IoU, and Contour Chamfer distance). Object Metrics are computed over object pairs matched using EgoX's protocol using SAM2-Large [27] and DINOv3 [29] for matching. EgoX's

Table 1: **Quantitative comparison on EgoExo4D.** Best in **bold**, second-best underlined. Evaluated on the samples using EgoX's evaluation protocol. *Unseen* indicates brand new environments not seen during training. *New Action* (renamed from EgoX's *Seen*) indicate unseen actions but in environment seen during training. Grey rows show the change over EgoX. Our numbers are averaged over five independent generations (standard deviations in Appendix B.1). *Loc Err* and *Contour* are normalized by the output side length. Values in [brackets] are metrics the EgoX paper does not report so we measured them ourselves.

		Image Metrics				Object Metrics			Video Metrics	
	Method	PSNR↑	SSIM↑	LPIPS↓	CLIP-I↑	Loc Err↓	IoU↑	Contour↓	FVD↓	T-LPIPS↓
New Action (new actions, seen envs.)	Exo2Ego-V [22]	14.53	0.384	0.569	0.774	0.326	0.074	–	622.47	–
	TrajectoryCrafter [36]	13.05	0.375	0.606	0.780	0.210	0.128	–	546.09	–
	Wan-Fun-Control [1]	12.25	0.463	0.617	0.810	0.235	0.076	–	595.07	–
	Wan-VACE [15]	12.95	0.413	0.626	0.829	0.228	0.114	–	508.69	–
	Vista4D [20]	10.39	0.301	0.670	0.796	0.139	0.254	0.090	647.51	4.40
	EgoX [16]	16.05	0.556	0.498	0.896	0.129	0.363	[0.071]	184.47	[2.688]
	Ours	**18.64**	**0.560**	**0.346**	**0.910**	**0.031**	**0.576**	**0.027**	**119.9**	**1.454**
	Δ vs EgoX	+2.59 dB	+0.7%	−30.5%	+1.6%	−76.0%	+58.7%	−62.0%	−35.0%	−45.9%
Unseen (new actions, new envs.)	Exo2Ego-V [22]	12.70	0.439	0.597	0.679	0.447	0.003	–	1283.50	–
	TrajectoryCrafter [36]	12.24	0.297	0.619	0.778	0.400	0.039	–	821.71	–
	Wan-Fun-Control [1]	13.59	0.439	0.604	0.799	0.399	0.042	–	968.78	–
	Wan-VACE [15]	12.17	0.345	0.638	0.820	0.400	0.038	–	1045.45	–
	Vista4D [20]	10.68	0.255	0.668	0.811	0.146	0.229	0.087	794.09	2.91
	EgoX [16]	14.38	0.457	0.552	0.877	0.312	0.092	[0.083]	440.64	[2.375]
	Ours	**16.05**	**0.460**	**0.467**	**0.894**	**0.052**	**0.412**	**0.044**	**385.8**	**1.485**
	Δ vs EgoX	+1.67 dB	+0.7%	−15.4%	+1.9%	−83.3%	+347.8%	−47.0%	−12.4%	−37.5%

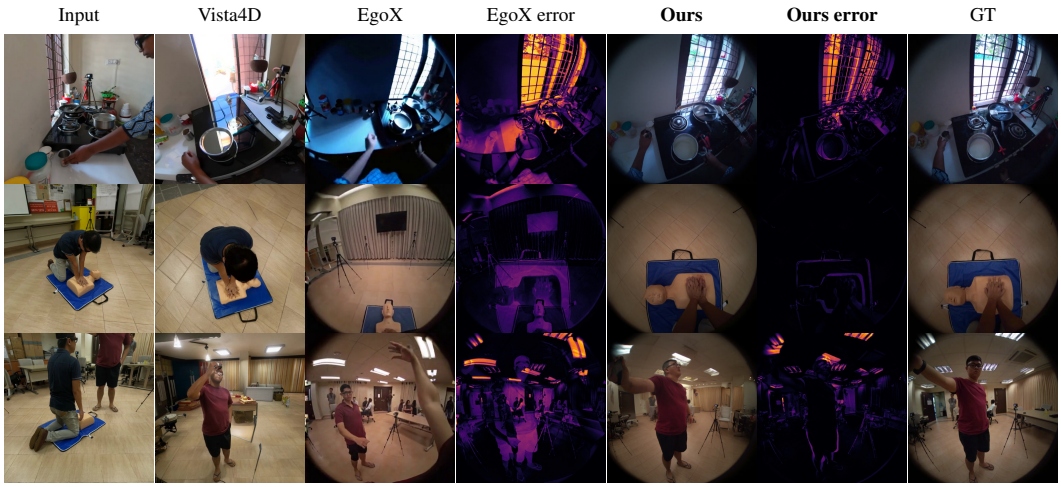


Figure 4: **Qualitative comparison on EgoExo4D (New Activities).**

Contour Accuracy protocol is no longer available, so we choose Chamfer distance normalized by image resolution as an easily reproducible replacement. The three VBench [14] metrics are deferred to Appendix B because these no-reference metrics do not measure the similarity between generated and ground-truth videos. Full definitions and additional metrics are in Appendix B.

Baselines. We compare against EgoX [16] as well as its baselines (Exo2Ego-V [22], TrajectoryCrafter [36], Wan-Fun-Control [1], Wan-VACE [15]) with numbers taken from EgoX [16]. Additionally, we adapt the recent state-of-the-art method in general view synthesis, Vista4D [20], to our setting to demonstrate the need for dedicated exo-to-ego methods. For a fair comparison, we also perform camera re-localization by aligning its Pi3X [34] reconstruction to ours. Two concurrent works, Exo2EgoSyn [23] (multiview-to-ego) and EgoWorld [25] (image-only), are not directly comparable, and neither has released code at the time of writing, so we exclude them from Table 1.

Table 2: **Ablation studies on EgoExo4D.** Top: component configurations; bottom: results averaged over five generations (standard deviations in Appendix B.1). Loc Err and Contour are normalized by image side length. * denotes our LTX-2.3 re-implementations of EgoX, which may understate the original method's performance. Unmarked EgoX values are reported results, except bracketed metrics measured by us.

Variant	Backbone	Our data	Camera reloc.	Seq. concat	Obj tok.	Obj mask	Synth data
EgoX [16]	Wan	–	–	–	–	–	–
EgoX (LTX, data)*	LTX	Y	–	–	–	–	–
EgoX (LTX, data, reloc)*	LTX	Y	Y	–	–	–	–
Ours (no obj tok, mask, synth)	LTX	Y	Y	Y	–	–	–
Ours (no obj mask, synth)	LTX	Y	Y	Y	Y	–	–
Ours (no synth)	LTX	Y	Y	Y	Y	Y	–
Ours	LTX	Y	Y	Y	Y	Y	Y

		Image Metrics				Object Metrics			Video Metrics	
	Variant	PSNR↑	SSIM↑	LPIPS↓	CLIP-I↑	Loc Err↓	IoU↑	Contour↓	FVD↓	T-LPIPS↓
New Action (new actions, seen envs.)	EgoX [16]	16.05	0.556	0.498	0.896	0.129	0.363	[0.071]	184.47	[2.688]
	EgoX (LTX, data)*	17.05	0.506	0.473	0.875	0.065	0.336	0.061	233.4	3.115
	EgoX (LTX, data, reloc)*	17.95	0.532	0.403	0.889	0.045	0.456	0.041	169.6	2.354
	Ours (no obj tok, mask, synth)	17.69	0.529	0.400	0.894	0.040	0.501	0.036	163.9	1.593
	Ours (no obj mask, synth)	18.06	0.541	0.377	0.900	0.035	0.540	0.031	142.8	1.540
	Ours (no synth)	18.39	0.552	0.355	0.908	0.032	0.569	0.028	122.9	**1.417**
	Ours	**18.64**	**0.560**	**0.346**	**0.910**	**0.031**	**0.576**	**0.027**	**119.9**	1.454
	Given ego attn mask	*18.94*	*0.568*	*0.330*	*0.913*	*0.026*	*0.605*	*0.022*	*113.8*	*1.437*
Unseen (new actions, new envs.)	EgoX [16]	14.38	0.457	0.552	0.877	0.312	0.092	[0.083]	440.64	[2.375]
	EgoX (LTX, data)*	15.50	0.439	0.533	0.853	0.085	0.243	0.078	534.6	2.693
	EgoX (LTX, data, reloc)*	15.92	0.449	0.497	0.868	0.070	0.316	0.063	467.4	2.063
	Ours (no obj tok, mask, synth)	15.67	0.445	0.491	0.882	0.062	0.374	0.054	417.9	1.598
	Ours (no obj mask, synth)	15.76	0.451	0.481	0.883	0.058	0.385	0.050	399.9	1.596
	Ours (no synth)	16.00	**0.460**	**0.462**	0.892	**0.052**	**0.418**	0.046	**372.4**	**1.416**
	Ours	**16.05**	**0.460**	0.467	**0.894**	**0.052**	0.412	**0.044**	385.8	1.485
	Given ego attn mask	*16.30*	*0.467*	*0.453*	*0.897*	*0.046*	*0.451*	*0.038*	*373.6*	*1.462*

4.2 Comparisons

Quantitative comparisons. As shown in Table 1, our method substantially outperforms EgoX and recent state-of-the-art methods by a wide margin. The significant improvements in metrics such as PSNR ($+1.67$ dB$/+2.59$ dB on *Unseen/New Action*) and LPIPS ($-15.4\%/-30.5\%$) indicate substantial improvements in reconstruction quality and alignment accuracy. The large improvement in T-LPIPS ($-37.5\%/-45.9\%$) indicates that our model reconstructs temporal dynamics (content dynamics and camera motion) much better than recent state-of-the-art methods. The drastic improvements in Object Metrics (Loc Err: $-83.3\%/-76.0\%$; IoU: $+347.8\%/+58.7\%$; Contour: $-47.0\%/-62.0\%$) indicate that not only are the video's general structure improved significantly, but object-level generation and localization are also substantially improved. The large improvements in generative metrics such as FVD ($-12.4\%/-35.0\%$) prove that our method generates videos much more similar to real egocentric videos in visual characteristics than existing methods.

Qualitative comparisons. As seen from Fig. 4, Vista4D consistently fails to generate ego views correctly. While EgoX performs better, its results show significant misalignment against the ground-truth. Our results consistently reconstruct the scene with accurate spatial alignment as well as the content of the scene. For example, the red boxes in Fig. 5 and 6 show that EgoX completely fails to reconstruct the persons, whereas our method succeeds.

The significant improvement validates our hypothesis that (1) dedicated exo-to-ego is needed as shown by the gap between Vista4D and EgoX, and (2) exo-to-ego benefits from a systematic solution spanning the data pipeline and the architecture. Notably, by adding semantics on top of the geometric scaffold, our method moves past the geometry-centric design that state-of-the-art methods share and achieves further improvements. Additionally, despite the significant improvements, the absolute performance still has much room to improve, for example, the average PSNR is still under 20dB.

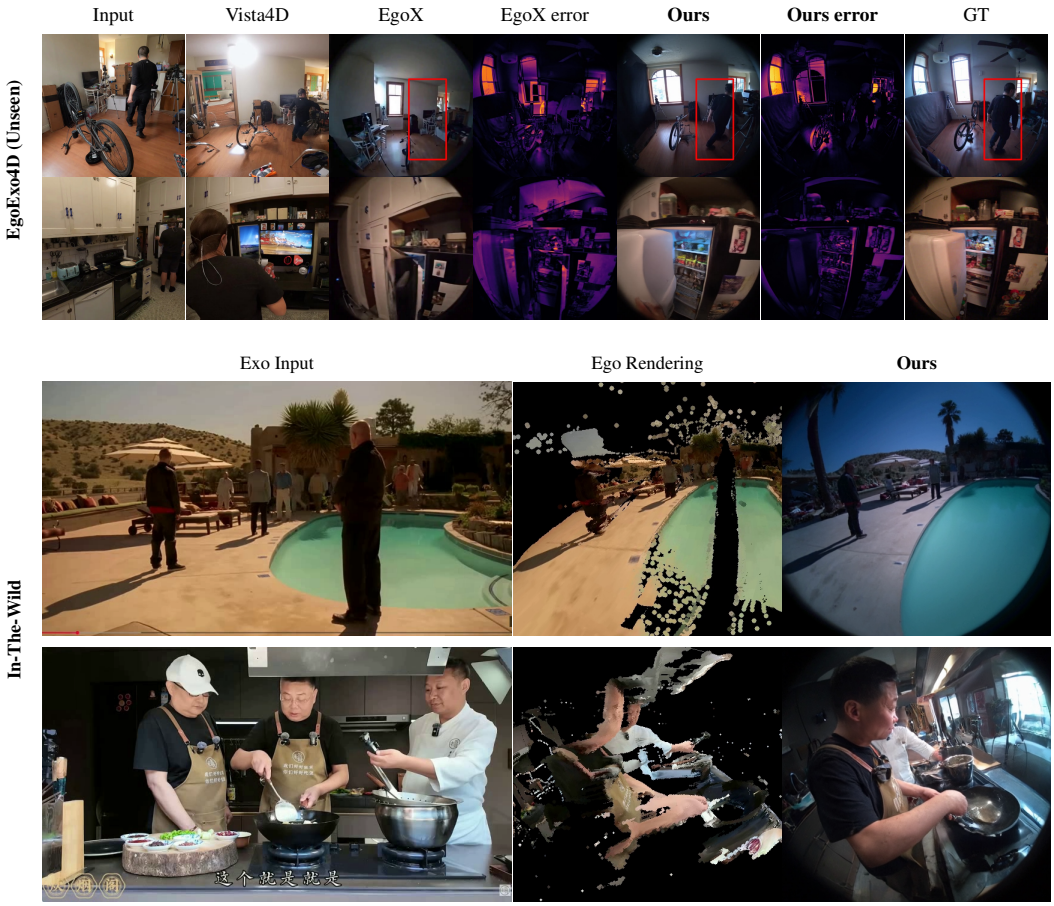


Figure 5: Qualitative comparisons on EgoExo4D (top) and in-the-wild samples (bottom).

4.3 Ablation studies

Table 2 isolates the contributions of backbone/training data, camera re-localization, conditioning, semantic grounding, and synthetic data. All deltas below are Unseen / New Action. Figure 6 provides qualitative comparisons.

Backbone and training data. Our EgoX re-implementation with LTX-2.3 and more real training data improves reconstruction fidelity (PSNR $+1.12/+1.00$ dB), but worsens FVD and T-LPIPS. These changes alone therefore do not explain the full model's improvements.

Camera re-localization. Adding re-localization improves all nine metrics, including IoU ($+30.0\%/+35.7\%$) and T-LPIPS ($-23.4\%/-24.4\%$). Aligning the rendered prior with the target video thus improves both spatial and temporal consistency. These gains arise from correcting camera-to-reconstruction alignment while retaining the original monocular geometry used to condition the generator.

Conditioning mechanism. For geometry-only models, sequence concatenation (IC-LoRA [13] style) improves IoU ($+18.4\%/+9.9\%$) and T-LPIPS ($-22.5\%/-32.3\%$) over EgoX's channel concatenation, with a small PSNR trade-off ($-0.25/-0.26$ dB).

Semantic grounding. Adding object tokens without spatial masks yields modest gains (IoU $+2.9\%/+7.8\%$). The full scene prompt is unchanged across these variants, so this comparison tests whether object-specific tokens help beyond the global caption. Adding masks while holding training data fixed further improves IoU ($+8.6\%/+5.4\%$), localization, contour accuracy, and video metrics. Together, object tokens and masks improve IoU by $+11.8\%/+13.6\%$ over the geometry-only variant,

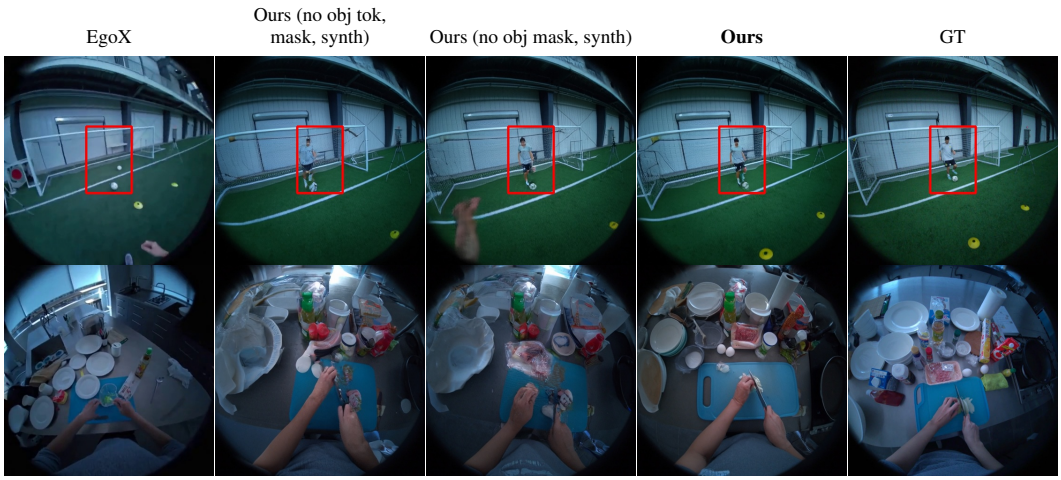


Figure 6: **Ablation study.** Top row: EgoX omits the player entirely, but our variants reconstruct the player progressively closer to the ground truth. Bottom row: EgoX fails to reconstruct the viewpoint and scene correctly, but our variants reconstruct the ego view progressively more accurately.

using no synthetic data in either case. This separates the benefit of spatially grounded semantics from that of synthetic training data.

Synthetic data. Adding synthetic data improves New Action PSNR by 0.25 dB and most other metrics on that split. Gains on Unseen are mixed, indicating that its benefit is strongest within seen environments. Since the architecture is unchanged, this comparison isolates the contribution of the additional synthetic examples.

Mask accuracy. Replacing reprojected masks with masks from the actual first ego frame improves all reported metrics, including IoU ($+9.5\%/ + 5.0\%$), localization, and contour accuracy. This diagnostic uses privileged target-view information unavailable in the intended setting and indicates headroom from more accurate semantic grounding.

4.4 Limitations.

Long video generation. The video length is limited due to VRAM constraints. Extending to longer videos could be done through chunk-by-chunk denoising as explored by recent methods. On our website, we share our on-going work which shows that a short and simple chunk-autoregressive finetuning stage allows the generation of longer videos (e.g. 10 seconds) with stable quality.

Occlusion, out-of-frame content, and dependence on 3D reconstruction. Our model utilizes off-the-shelf 3D reconstruction methods for video generation. While 3D reconstruction methods have been improving rapidly in terms of accuracy and completeness, a persistent challenge is that occlusion and out-of-frame contents are difficult to reconstruct. When the reconstruction is largely blank in the view of the ego camera, the model needs to hallucinate most of the content so accuracy can thus be undermined. When the exo camera is allowed to move across the scene to gather information from different viewpoints, this challenge can be alleviated, though hard to eliminate. This is a fundamental challenge, and we leave it for future work. Appendix C quantifies how far this dependence extends.

5 Conclusion

Grounded-Exo2Ego combines camera re-localization and synthetic training data with a dual-branch architecture that grounds object-level semantics in the reconstructed scene. Experiments on EgoExo4D demonstrate substantial improvements over prior methods, and ablations validate the contributions of both the data pipeline and semantic grounding.

Acknowledgments and Disclosure of Funding

This work was supported solely by NVIDIA. The authors declare no additional competing interests beyond their NVIDIA affiliations.

References

[1] aigc apps. Videox-fun: A video generation pipeline for diffusion transformer, 2026. URL https://github.com/aigc-apps/VideoX-Fun.

[2] 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.

[3] 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. *arXiv preprint arXiv:2412.07760*, 2024.

[4] Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. *arXiv preprint arXiv:2503.11647*, 2025.

[5] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. URL http://www.blender.org.

[6] 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. 2025.

[7] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts. 2025. URL https://arxiv.org/abs/2511.16719.

[8] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. 2022.

[9] Mengyu Chu, You Xie, Jonas Mayer, Laura Leal-Taixé, and Nils Thuerey. Learning temporal coherence via self-supervision for gan-based video generation. *ACM Transactions on Graphics*, 39(4), August 2020. ISSN 1557-7368. doi: 10.1145/3386569.3392457. URL http://dx.doi.org/10.1145/3386569.3392457.

[10] Xiao Fu, Shitao Tang, Min Shi, Xian Liu, Jinwei Gu, Ming-Yu Liu, Dahua Lin, and Chen-Hsuan Lin. Plenoptic video generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

[11] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zach Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, Maria Escobar, and Michael Wray. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives. pages 19383–19400, 06 2024. doi: 10.1109/CVPR52733.2024.01834.

[12] Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, Eitan Richardson, Guy Shiran, Itay Chachy, Jonathan Chetboun, Michael Finkelson, Michael Kupchick, Nir Zabari,

Nitzan Guetta, Noa Kotler, Ofir Bibi, Ori Gordon, Poriya Panet, Roi Benita, Shahar Armon, Victor Kulikov, Yaron Inger, Yonatan Shiftan, Zeev Melumian, and Zeev Farbman. Ltx-2: Efficient joint audio-visual foundation model. 2026. URL https://arxiv.org/abs/2601.03233.

[13] Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Huanzhang Dou, Chen Liang, Yutong Feng, Yu Liu, and Jingren Zhou. In-context lora for diffusion transformers. *arXiv preprint arxiv:2410.23775*, 2024.

[14] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[15] Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. pages 17191–17202, 2025.

[16] Taewoong Kang, Kinam Kim, Dohyeon Kim, Minho Park, Junha Hyung, and Jaegul Choo. Egox: Egocentric video generation from a single exocentric video. 2025. URL https://arxiv.org/abs/2512.08269.

[17] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *ACM Transactions on Graphics*, 39(6), 2020.

[18] Ruilong Li, Brent Yi, Junchen Liu, Hang Gao, Yi Ma, and Angjoo Kanazawa. Cameras as relative positional encoding. *Advances in Neural Information Processing Systems*, 2025.

[19] Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

[20] Kuan Heng Lin, Zhizheng Liu, Pablo Salamanca, Yash Kant, Ryan Burgert, Koichi Namekata, Yiwei Zhao, Bolei Zhou, Micah Goldblum, Paul Debevec, and Ning Yu. Vista4D: Video reshooting with 4d point clouds. 2026.

[21] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

[22] Jia-Wei Liu, Weijia Mao, Zhongcong Xu, Jussi Keppo, and Mike Zheng Shou. Exocentric-to-egocentric video generation. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 136149–136172. Curran Associates, Inc., 2024. doi: 10.52202/079017-4325. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/f5a8b5e5d007e66c929b971c2bc21d76-Paper-Conference.pdf.

[23] 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. URL https://arxiv.org/abs/2511.20186.

[24] Claire McLean, Makenzie Meendering, Tristan Swartz, Orri Gabbay, Alexandra Olsen, Rachel Jacobs, Nicholas Rosen, Philippe de Bree, Tony Garcia, Gadsden Merrill, Jake Sandakly, Julia Buffalini, Neham Jain, Steven Krenn, Moneish Kumar, Dejan Markovic, Evonne Ng, Fabian Prada, Andrew Saba, Siwei Zhang, Vasu Agrawal, Tim Godisart, Alexander Richard, and Michael Zollhoefer. Embody 3d: A large-scale multimodal motion and behavior dataset, 2025. URL https://arxiv.org/abs/2510.16258.

[25] 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*, 2026.

[26] Alexander Raistrick, Lahav Lipson, Zeyu Ma, Lingjie Mei, Mingzhe Wang, Yiming Zuo, Karhan Kayan, Hongyu Wen, Beining Han, Yihan Wang, Alejandro Newell, Hei Law, Ankit Goyal, Kaiyu Yang, and Jia Deng. Infinite photorealistic worlds using procedural generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12630–12641, 2023.

[27] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. 2024. URL `https://arxiv.org/abs/2408.00714`.

[28] 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*, 2025.

[29] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3. 2025. URL `https://arxiv.org/abs/2508.10104`.

[30] Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. 2021.

[31] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel,

Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report. 2025. URL https://arxiv.org/abs/2503.19786.

[32] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *CoRR*, abs/1812.01717, 2018. URL http://arxiv.org/abs/1812.01717.

[33] 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. 2025. URL https://arxiv.org/abs/2507.02546.

[34] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347*, 2025.

[35] Long Xing, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Jianze Liang, Qidong Huang, Jiaqi Wang, Feng Wu, and Dahua Lin. Caprl: Stimulating dense image caption capabilities via reinforcement learning. 2025. URL https://arxiv.org/abs/2509.22647.

[36] 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)*, pages 100–111, October 2025.

[37] 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. 2026. URL https://arxiv.org/abs/2512.07237.

Supplementary Material

A Additional qualitative results

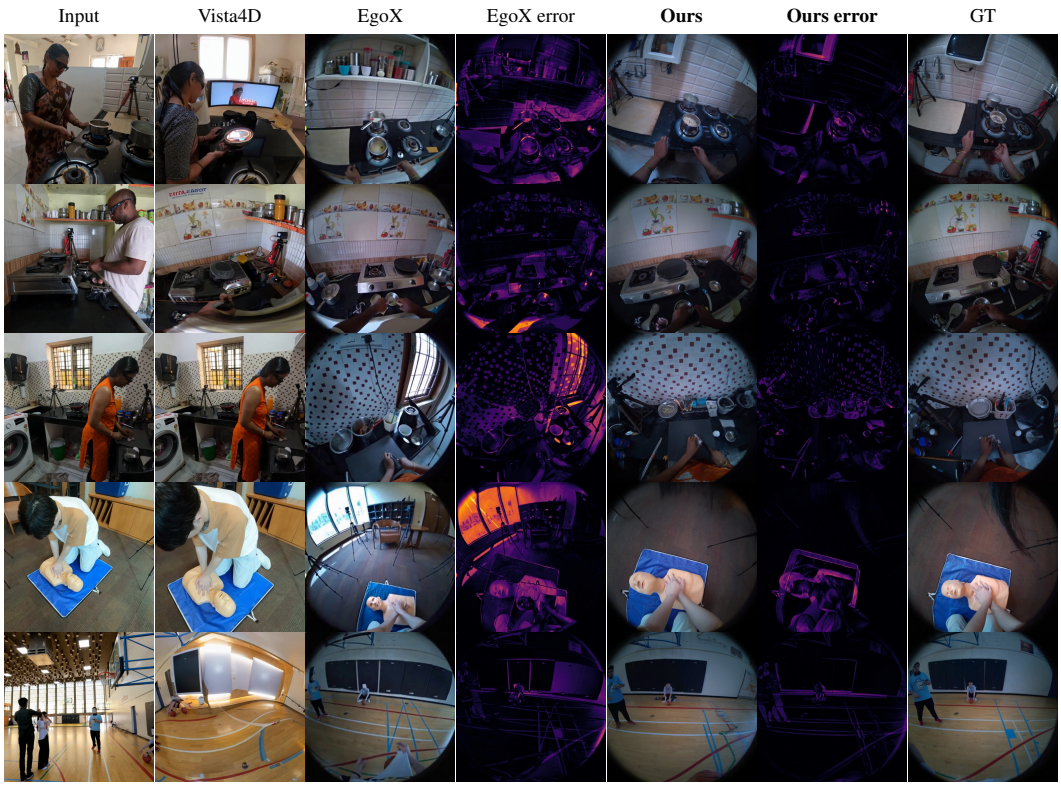


Figure 7: **Additional New-Activity scenes.** Columns as in Figure 4.

B Evaluation metrics

This section gives the definitions for the Chamfer contour distance and describes why we deferred the VBench metrics.

Chamfer Contour distance. Following EgoX's convention, we run SAM2-Large [27] to automatically generate segmentation masks for each of the generated and GT frames. We embed each detected object using DINOv3 ViT-L/16 [29]. We then match objects based on their cosine similarities of DINOv3 features and Hungarian matching (this is shared with the Location Error and IoU metrics as done in EgoX). For each pair of masks we extract the boundary pixel set $\mathcal{B}^{\text{gen}}$ and $\mathcal{B}^{\text{gt}}$ and compute the two-sided Chamfer distance

$$\text{CD} = \frac{1}{l}\left[\frac{1}{|\mathcal{B}^{\text{gen}}|}\sum_{p\in\mathcal{B}^{\text{gen}}}\min_{q\in\mathcal{B}^{\text{gt}}}\|p-q\|_2 \;+\; \frac{1}{|\mathcal{B}^{\text{gt}}|}\sum_{q\in\mathcal{B}^{\text{gt}}}\min_{p\in\mathcal{B}^{\text{gen}}}\|p-q\|_2\right],$$

where the per-pixel boundary distances are normalized by the square image's side length ($l = 448$ px) into a dimension-agnostic score for the ease of reproducibility. This metric is the lower the better.

ΔVBench (Temporal Flickering, Motion Smoothness, Dynamic Degree). All three VBench [14] metrics (temporal flickering, motion smoothness, dynamic degree) used in EgoX [16] are **no-reference metrics that scores a video without comparing to the ground-truth**, and real ego videos do not maximize these scores. The ground-truth videos actually score worse than several of the baselines, which indicates flawed measurement. Therefore, scoring higher on these metrics

do not indicate better fidelity. In this section, we report the metrics as delta deviation from GT, *i.e.* |score − GT score|. Comparing to VBench, T-LPIPS natively compares against GT, thus we use it for the main table.

	New Action			Unseen		
Method	$\Delta_{\text{TF}}\downarrow$	$\Delta_{\text{MS}}\downarrow$	$\Delta_{\text{DD}}\downarrow$	$\Delta_{\text{TF}}\downarrow$	$\Delta_{\text{MS}}\downarrow$	$\Delta_{\text{DD}}\downarrow$
Vista4D [20]	0.013	0.004	0.072	0.010	0.003	0.160
EgoX [16] (paper)	0.008	0.002	0.137	0.013	0.004	0.209
Ours	**0.004**	**0.001**	**0.003**	**0.005**	**0.001**	**0.010**

Table 3: **ΔVBench on the New-Activity (seen) and Unseen splits.** Lower magnitude is better; 0 indicates a perfect match to GT on the corresponding signal.

B.1 Variance across generations

Numbers in the main paper are averaged over five independent generations. Table 4 reports the standard deviation across those runs.

	Variant	Loc Err	IoU	Contour	PSNR	SSIM	LPIPS	FVD	T-LPIPS
New Action	EgoX (LTX, data)*	±0.001	±0.003	±0.001	±0.03	±0.001	±0.002	±7.6	±0.023
	EgoX (LTX, data, reloc)*	±0.002	±0.006	±0.002	±0.03	±0.001	±0.002	±8.2	±0.038
	Ours (no obj tok, mask, synth)	±0.001	±0.005	±0.001	±0.03	±0.001	±0.001	±4.3	±0.022
	Ours (no obj mask, synth)	±0.000	±0.004	±0.001	±0.03	±0.001	±0.001	±8.6	±0.031
	Ours (no synth)	±0.000	±0.005	±0.001	±0.04	±0.001	±0.002	±4.1	±0.022
	Ours	±0.001	±0.005	±0.001	±0.05	±0.001	±0.001	±2.0	±0.016
Unseen	EgoX (LTX, data)*	±0.001	±0.018	±0.002	±0.05	±0.003	±0.001	±21.4	±0.056
	EgoX (LTX, data, reloc)*	±0.001	±0.008	±0.000	±0.06	±0.003	±0.002	±10.0	±0.066
	Ours (no obj tok, mask, synth)	±0.004	±0.007	±0.005	±0.05	±0.004	±0.003	±5.6	±0.067
	Ours (no obj mask, synth)	±0.002	±0.010	±0.002	±0.05	±0.003	±0.003	±14.6	±0.061
	Ours (no synth)	±0.003	±0.004	±0.003	±0.05	±0.002	±0.003	±10.2	±0.060
	Ours	±0.002	±0.008	±0.003	±0.05	±0.002	±0.002	±15.4	±0.027

Table 4: **Variance across generations.** Standard deviation over five independent generations. Means are reported in Tables 1 and 2.

C Robustness to imperfect geometry, masks, and ego pose

Our conditioning is assembled from three estimated quantities: the monocular depth behind the ego prior, the object segmentation masks that gate the semantic tokens, and the target ego camera pose. This section measures how the generation quality degrades as each of them is corrupted.

Every corruption is applied at evaluation, so each row differs from its baseline by exactly one perturbation. Depth and pose are perturbed before rendering, so a single knob degrades the whole geometric and semantic pathway together: rasterization, the reprojected UV field, the 3D positional encoding, occupancy, and the mask reprojection that rides the same field. Perturbing them further downstream would corrupt the render while the gate still read clean geometry, which is not a failure mode that can occur in practice. Depth error is injected either as spatially smooth (low-frequency) error or as independent per-pixel noise of the same magnitude.

Mask perturbations are done after reprojection, so erosion and dilation radii are in gate pixels.

The uncorrupted baselines are those of Table 1.

Imperfect geometry. In Table 5 we show robustness analysis under different levels of geometric noise. We compare to the strongest EgoX variant, i.e. our implementation of EgoX (LTX, data, reloc).

Noisy ego pose. In Table 6 we show robustness analysis under different levels of pose noise. We compare to the strongest EgoX variant, i.e. our implementation of EgoX (LTX, data, reloc).

		Object Metrics			Image Metrics			Video Metrics	
	Depth error	Loc Err↓	IoU↑	Contour↓	PSNR↑	SSIM↑	LPIPS↓	FVD↓	T-LPIPS↓
Ours	none (deployed)	0.0518	0.4121	0.0437	16.05	0.4600	0.4673	385.8	1.485
	2.5 cm smooth	0.0555	0.3980	0.0480	15.96	0.4579	0.4752	400.2	1.530
	2.5 cm per-pixel	0.0552	0.3911	0.0464	15.91	0.4571	0.4782	400.6	1.562
	5 cm smooth	0.0581	0.3827	0.0502	15.83	0.4540	0.4848	410.1	1.580
	5 cm per-pixel	0.0611	0.3627	0.0525	15.75	0.4504	0.4905	410.0	1.582
	10 cm smooth	0.0630	0.3533	0.0534	15.62	0.4472	0.4978	429.6	1.654
	10 cm per-pixel	0.0681	0.3151	0.0595	15.40	0.4376	0.5089	429.9	1.665
	20 cm smooth	0.0721	0.3036	0.0626	15.32	0.4367	0.5133	427.5	1.822
	20 cm per-pixel	0.0773	0.2696	0.0684	15.06	0.4242	0.5254	452.8	1.830
EgoX (LTX, data, reloc)*	none	0.0702	0.3156	0.0631	15.92	0.4489	0.4973	467.4	2.063
	2.5 cm smooth	0.0737	0.3083	0.0662	15.94	0.4515	0.5008	470.2	2.160
	2.5 cm per-pixel	0.0735	0.3038	0.0666	15.93	0.4513	0.5026	466.2	2.193
	5 cm smooth	0.0732	0.3015	0.0652	15.92	0.4513	0.5044	472.5	2.233
	5 cm per-pixel	0.0776	0.2887	0.0689	15.88	0.4506	0.5072	473.4	2.279
	10 cm smooth	0.0781	0.2868	0.0718	15.83	0.4493	0.5121	488.4	2.350
	10 cm per-pixel	0.0788	0.2767	0.0719	15.76	0.4474	0.5153	484.1	2.433
	20 cm smooth	0.0815	0.2633	0.0753	15.71	0.4467	0.5227	497.7	2.651
	20 cm per-pixel	0.0856	0.2478	0.0801	15.59	0.4425	0.5302	513.9	2.719

Table 5: **Robustness to imperfect geometry.** Depth is corrupted before rendering, at four magnitudes and in two spatial profiles. *Ours* degrades faster in relative terms but stays ahead in absolute terms on every object and video metric at every level; *EgoX* overtakes only on the two pixel-fidelity metrics, PSNR and SSIM.

		Object Metrics			Image Metrics			Video Metrics	
	Pose error (rot. / trans.)	Loc Err↓	IoU↑	Contour↓	PSNR↑	SSIM↑	LPIPS↓	FVD↓	T-LPIPS↓
Ours	none (deployed)	0.0518	0.4121	0.0437	16.05	0.4600	0.4673	385.8	1.485
	1° / 1% (0.017 m)	0.0541	0.4009	0.0461	15.96	0.4556	0.4717	400.5	1.480
	2° / 2% (0.034 m)	0.0538	0.3887	0.0461	15.85	0.4507	0.4774	398.5	1.493
	5° / 5% (0.085 m)	0.0618	0.3254	0.0539	15.44	0.4360	0.5012	383.3	1.543
EgoX (LTX, data, reloc)*	none	0.0702	0.3156	0.0631	15.92	0.4489	0.4973	467.4	2.063
	1° / 1% (0.017 m)	0.0697	0.3185	0.0644	15.92	0.4501	0.4974	474.5	2.062
	2° / 2% (0.034 m)	0.0724	0.3031	0.0650	15.85	0.4476	0.5010	473.5	2.073
	5° / 5% (0.085 m)	0.0747	0.2742	0.0675	15.51	0.4385	0.5166	464.9	2.070

Table 6: **Robustness to a noisy target ego pose.** One constant SE(3) offset per clip, applied before rendering: a mis-localized ego camera rather than per-frame jitter. Rotation is about a random axis. Translation is a fraction of that clip's median scene depth — the reconstruction is scale-ambiguous, so a fixed offset in metres would be a different-sized error in every clip — with the implied median offset in brackets (median scene depth is 1.69 m over *Unseen*). Both methods are largely insensitive up to 2° / 2%.

Imperfect masks In Table 7 we show robustness analysis under different levels of pose noise.

D Method details

D.1 Geometric Anchoring

Mesh rasterization. Given re-localized ego poses and monocular depth, we build a triangle mesh from the exo depth grid by connecting neighboring pixels, cull triangles at depth discontinuities (relative threshold 5%), transform vertices to the ego frame, and rasterize via nvdiffrast [17]. This produces ego-view RGB renderings, depth maps, and UV correspondence maps that the subsequent stages consume.

	Object Metrics			Image Metrics			Video Metrics	
Mask condition	Loc Err↓	IoU↑	Contour↓	PSNR↑	SSIM↑	LPIPS↓	FVD↓	T-LPIPS↓
Ours	0.0518	0.4121	0.0437	16.05	0.4600	0.4673	385.8	1.485
Given ego attn mask	0.0461	0.4513	0.0381	16.30	0.4674	0.4529	373.6	1.462
erode 2 px	0.0528	0.4089	0.0452	16.01	0.4591	0.4708	401.3	1.492
erode 4 px	0.0557	0.3993	0.0479	15.99	0.4586	0.4714	397.7	1.499
erode 8 px	0.0543	0.4044	0.0467	15.99	0.4584	0.4715	397.0	1.499
dilate 2 px	0.0538	0.4050	0.0454	15.98	0.4568	0.4708	390.4	1.470
dilate 4 px	0.0571	0.3835	0.0486	15.90	0.4533	0.4758	392.2	1.454
dilate 8 px	0.0610	0.3687	0.0526	15.72	0.4469	0.4853	396.8	1.455
object tokens ungated	0.0825	0.2945	0.0750	15.15	0.4224	0.5216	447.7	1.739

Table 7: **Robustness to imperfect object masks** (*Ours* only; EgoX has no object pathway). The bottom block varies how open the attention gate is, from tighter than deployed (erode), through looser (dilate), to removed altogether (ungated). Removing the gate costs more than any perturbation of it.

Render-token injection. The geometric rendering tokens $\mathbf{Z}_{\text{ego}}^{\text{render}}$ (Eq. 2) are appended to the noisy target latent $\mathbf{z}_\sigma$ rather than channel-concatenated. They share the same RoPE spatial coordinates as the noisy latent tokens, but their RoPE coordinates are x-shifted by the width of the ego video width.

Occupancy masking. Rendered ego-view images contain holes (black regions) where the exo camera has no coverage. After VAE encoding and patchification, some tokens cover mostly-empty regions and carry no useful information. Furthermore, they could mislead the model into generating black regions. Therefore, we generate token occupancy masks that filter out tokens that correspond to spatial temporal cells with less than 10% of the underlying pixels occupied.

D.2 Semantic grounding

Caption generation. We caption the first frame of each video split using a CapRL-Qwen3VL-4B [35]. Each caption contains four sections: an overall scene description, up to 15 object noun phrases describing visible objects, object spatial and interaction relations, and a predicted ego-view description. The captioning covers all ego and exo cameras.

Object extraction. For each split we union the noun phrases across cameras with ego phrases prioritized. Gemma-3-12B [31] then returns up to 50 unique objects ranked by relevance, with a single prompt per split, then deduplicates and re-ranks them. The prompt instructs the model to merge phrases denoting the same object (e.g. "bike wheels" and "bicycle wheels" are considered the same), and to rank by relevance to the camera wearer's viewpoint and interactions rather than by image area, so that objects the ego view will actually contain are preferred over objects that merely occupy many exo pixels. We use the top $K{=}10$ candidates. Because each phrase is used as a text prompt for the segmentation model, which returns every instance matching that phrase, the cap limits the number of distinct descriptions rather than the number of segmented regions, for example, a single "bike wheels" slot may cover several wheels.

Obtaining ego segmentation masks from exo. The attention gates used by the cross-attention mechanism (§3.5) are per-object binary masks computed from the exo view. We run SAM3 [7] with the ranked and deduplicated noun phrases as text prompts, producing per-object masks. Exo masks are first KB3-undistorted and then reprojected to the ego view.

Embedding object tokens. Each noun phrase q_k is encoded by Gemma-3 as raw embeddings $G(q_k)$, padded to the fixed per-object length $L{=}10$ tokens with learned register tokens $\mathbf{r}$, and projected into the DiT cross-attention space by the trainable projector C_ϕ:

$$\mathbf{O}_k = C_\phi\big([\, G(q_k), \mathbf{r}_{|G(q_k)|}, \ldots, \mathbf{r}_L \,]\big) \in \mathbb{R}^{L\times d}. \tag{5}$$

The max length of the raw embeddings $G(q_k)$ across the dataset is 6, so they are always padded.

Mixed-source mask training. During training we replace the reprojected exo masks with ground-truth ego segmentation masks with probability p=0.5.

E Data pipelines

E.1 Real data pipeline

The misalignment problem. EgoExo4D provides ground-truth ego camera poses and exo camera poses calibrated with respect to the real world geometry. However, monocular depth estimators produce 3D reconstructions whose scale and geometry do not match the real world. Rendering from the GT ego pose often places the scene at the wrong position. Such renderings provide unreliable signals that are usually misaligned or completely blank. Providing well-aligned renderings only at test time does not meaningfully improve quality, because the model has learned to distrust misaligned rendering conditions during training.

Re-localizing within the monocular reconstruction. Because the monocular reconstructions are inaccurate, we need to re-localize the ego camera inside the monocular reconstruction to achieve good alignment between the rendering and the ground-truth videos. Directly estimating the person's head pose with human pose estimation or face tracking is unreliable under occlusion and challenging head motion. Instead, we refine the provided ego trajectory labels within the monocular reconstruction. Concretely, we build a per-clip alignment between MoGe-2 [33] reconstructions and a multi-view reconstruction from Depth-Anything-3 [19]. Notice that the DA3 reconstruction is metric aligned because it uses ground-truth exo camera calibration, and it is only used to calculate the ego camera transformation, and we never use it during training or testing. We then compute a non-rigid alignment from DA3's depth map to MoGe2. The alignment is non-rigid in the sense that the depth alignment function is a depth-dependent alignment function. We then apply this alignment to the GT ego camera, bringing it from the real-world coordinate frame into the MoGe2 reconstruction. Only the depth of the camera is refined (and hence the 3D camera center), and the GT ego rotations are unchanged. The result, used everywhere in training and evaluation, is a set of ego poses that, when rendered from inside the monocular reconstruction, produces ego-view images better aligned to the GT ego view.

Table 8 reports how far the camera relocalization step moves the ego camera, measured over 14.0M EgoExo4D frames. For ego cameras deeper than 6 meters in the exo view, the median displacement is approximately 1.86 meters. Table 2 shows that the correction is worthwhile: adding it alone (*EgoX (LTX, data, reloc)**, our re-implementation) improves all nine metrics.

Ego depth in exo view (m)	Shift (cm)	Shift (%)
0–1	4.5	5.2
1–3	6.9	3.9
3–6	83.0	17.6
> 6	185.5	24.7

Table 8: **Ego-camera displacement from re-localization.** Median over 14.0M frames, binned by the ego camera's depth along the exocentric optical axis. Shift (%) is relative to the ego camera's distance from the exo camera.

Coordinate consistency, not depth repair. It is worth stating explicitly what the camera relocalization step does not attempt to correct the reconstruction and make it match the groundtruth video perfectly. This is because the reconstructions generated by off-the-shelf depth estimators are usually non-rigidly distorted, and no repositioning of the ego camera will make the rendering exactly match the ground-truth ego frame everywhere. Our method does not attempt this. The failure it removes is a coordinate-frame failure. The EgoExo4D ego pose is calibrated against the real scene, whereas MoGe2 produces an independently scaled and non-rigidly distorted proxy, so the calibrated camera is not a valid camera in the proxy's frame at all; applying it directly places the virtual camera in the wrong part of the reconstruction and yields a displaced or nearly blank rendering regardless of how good the local geometry is.

An alternative would be to align the monocular depth itself to the multi-view reconstruction and train on the corrected geometry. We deliberately do not do this. Ground-truth depth is unavailable at

inference time, creating a train/test mismatch. Training on the original MoGe2 reconstruction instead exposes the generator to the same noise and distortion characteristics it will see at test time, so that completion and correction are learned jointly. The cost of this choice is that the learned correction behavior is to some extent tuned to the error characteristics of the reconstruction model, which we list as a limitation in Appendix G.

Why this generalizes to in-the-wild data. Our model only ever observes renderings from the original MoGe2 reconstruction during training or testing. Therefore, as long as the ego camera can be localized within the MoGe2 reconstruction during test time, $e.g.$ via face tracking or manual view selection, our model can generate ego views in the same way as trained.

E.2 Generative synthetic data engine

Our synthetic dataset uses randomized combinations of interior scenes, characters and animations. For each combination we render images and depth maps for ego-centric cameras, an exocentric camera, and save the camera world transform and field of view as meta data. The generation process works as follows:

We first create a set of interior scenes using Infinigen [26] (640 in total). Our customized character pipeline uses *gpt-image-1* to generate character images from text prompts and *Meshy.ai* to generate 3D characters from those images. We then apply a custom algorithm to rig the generated characters, producing an asset pool of 550 rigged and textured characters. We animate them with a wide variety of motion sequences (3,000 in total, 30 fps) drawn from the Embody3D [24] motion capture dataset, which records real human-human social interactions among groups of two to four people.

We then create random selections of pre-generated interior scenes, one or two characters, and assign a random set of character animations. We then select a random 100-frame region in the animation timeline. Ego cameras are then attached to the character's neck bone, therefore following the character animation. We add a temporal smoothing filter to reduce unnatural camera jitter, which is common to these captured sequences. We then search for a valid exo camera through a randomized placement strategy. We consider the exo camera to be in a valid configuration if at least 50% of the joint positions of the characters are free of occlusions at all times in the selected animation range. At the same time the exo camera smoothly follows the average locations of the characters. If no valid configuration could be found after 1000 iterations we discard the configuration. After the exo camera has been placed in a valid location. Using the described strategy, we sample 10,000 sequences from an asset pool. We render the scenes through high-quality path tracing in Blender [5] and write final color and metric depth passes. The ego cameras use fisheye projection while the exo camera uses perspective projection.

F Vista4D coordinate alignment

Vista4D [20] is a conventional novel-view-synthesis video model that targets arbitrary 4D re-shooting from a casually captured monocular input.

Metric depth. Vista4D's reconstruction depth is up-to-scale, which is incompatible with metric GT ego poses. Thus, similar to our own data pipeline, we calculate a non-rigid transform that aligns its reconstruction with same DA-3 depth used in our own pipeline (Appendix E.1). This alignment is then used to locate the ego camera inside Vista4D's reconstruction.

Aria fisheye target. Vista4D is trained with pinhole cameras, but we find that it is generalizable to fisheye cameras. While the fisheye distortion is not always accurate, the viewpoint and contents of the scenes are consistent across fisheye and pinhole conditionings. We observe the same inability to generate ego views correctly.

G Limitations and societal impact

Limitations. Several caveats apply to the deployed model. (i) *Limited resolution and temporal length.* The output egocentric video is generated at 448×448 over $F = 33$ frames. Longer or

higher-resolution synthesis would require chunk-by-chunk generation. **On our website, we show the latest chun-autoregressive results.** (ii) *A target ego camera trajectory needs to be estimated at test time.* The method synthesizes the ego view for a given ego camera path rather than inferring where the camera wearer will look. In our in-the-wild pipeline, this trajectory is produced automatically from human and face pose estimation together with depth estimation, so the ego camera estimation accuracy bounds the accuracy of the view synthesis result. (iii) *First-frame masks broadcast over time.* The deployed model computes the object gate from segmentation of the first frame and reuses that spatial gate for the remainder of the clip. For sequences with substantial camera or object motion the gate becomes progressively stale, and per-frame masks would be the natural remedy, though at a higher processing and storage cost before training can start. (iv) *Dependence on semantic segmentation.* We use off-the-shelf models for semantic segmentation. If the masks are too permissive, object semantics can be routed outside the objects they describe and can potentially undermine generation. (v) *Dependence on 3D reconstruction.* When the exocentric depth is severely incomplete, corrupted, or the ego view observes contents not shown in the exo video, video generation quality can be undermined. (vi) *Dependence on large language models.* The pipeline also leverages off-the-shelf VLM and LLM for captioning. This makes the system capable but heavy, and reproducing it requires access to all of these components. (vii) *Evaluation breadth.* Following EgoX [16] convention, our quantitative evaluation is performed on subsets of the EgoExo4D dataset because the full test set is too large for modern models to finish video generation within a reasonable compute and time budget. The in-the-wild results we show are qualitative and are not a second quantitative benchmark. (viii) *Ego-Exo consistency.* Due to occlusions and the generative nature of our model, the generated ego video is not perfectly consistent with the input exo video. This challenge might be alleviated by generating in the 3D space instead of 2D videos. We leave this for future work.

Future work. One interesting direction is to generate semantic masks as intermediate outputs as part of visual-chain-of-thoughts. Our early experiments show that an additional Mask2Former-style [8] branch can generate rather accurate semantic masks, and we leave this for future work.

Societal impact. The negative societal impacts of this work are scenarios common to generative models, such as potential misrepresentation of one's first-person viewpoint and potential use for fraud. The positive societal impacts are described as the motivation of this work (Section 1).

H Safeguards

Given the fraud and misrepresentation risks noted in Appendix G, exo-to-ego generation systems should ship with provenance safeguards. A standard option is invisible digital watermarking: embedding signatures in the generated pixels that are imperceptible to the human eye but recoverable by specialized detector models, allowing downstream tools to flag synthesized first-person video.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: Our claims are backed by superior quantitative performance and strong ablation studies that justify the claims

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We discuss the limitations in the limitation section

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [N/A]

Justification: the paper presents an empirical method without formal theoretical results.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: We have disclosed the implementation of the model and the data pipeline in the paper to the best of our ability. Additionally, we are currently going through the internal processes to release the code and data.

   Guidelines:

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

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

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

Answer: [Yes]

Justification: We are currently going through the internal processes to release the code and data.

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: We have shared such details in the experiments section and more in the supplementary to the best of our abilities.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [No]

   Justification: While there are inherent randomness in diffusion models, our method reliably produces the reported numbers closely.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: We provide these information in the experiments section

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: We have reviewed and confirmed that the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: In the supplementary, we discuss the negative society impact, involving scenarios common to generative models, such as potential misrepresentation of one's first person viewpoint and potential use for fraud. The positive societal impacts are described as the motivation of this work.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [Yes]

    Justification: In supplementary

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: Yes, we cited the papers.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [Yes]

    Justification: We have documented and anonymized the assets to the best of our abilities. The assets have not been released yet because we are going through internal processes to release them to the public.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [Yes]

    Justification: We described how LLMs are used to caption images and provide input for our video generation model.

    Guidelines:

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