ALL-IN-ONE UNIFIED EGOCENTRIC WORLD SIMULATOR

Anonymous authors
Paper under double-blind review

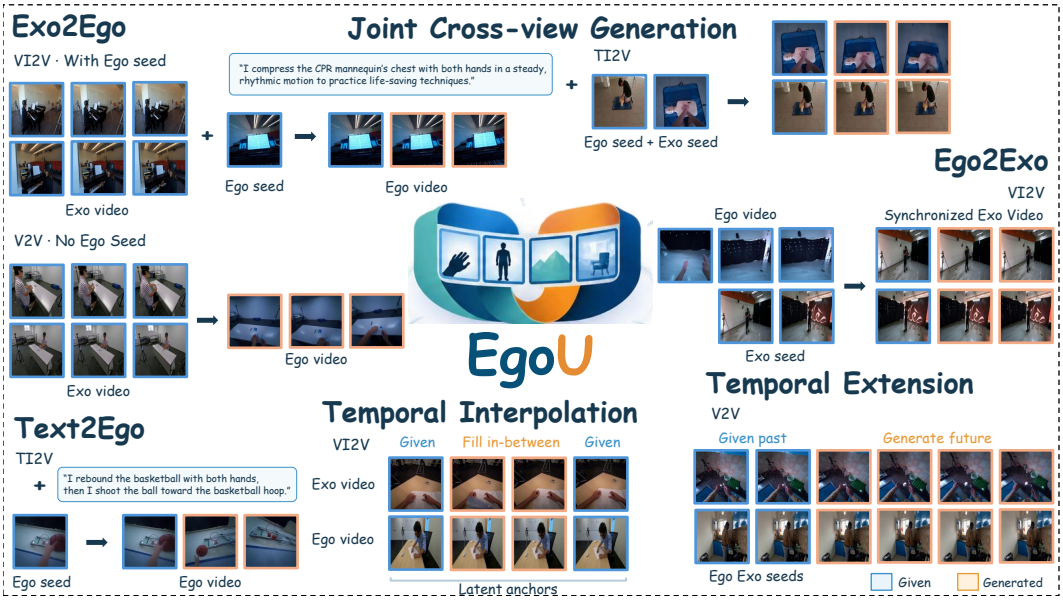


Figure 1: **EgoU: one model, seven ego–exo video tasks.** Each part is a task written as a clamp mask over a shared multi-view canvas and solved by the same weights. Blue marks the given views, frames, or text; orange marks generated frames of the target view over time.

ABSTRACT

Egocentric and exocentric videos are two views of the same physical event: the first-person view captures intent-driven hand–object interaction, while the third-person view reveals the camera wearer's full-body motion and the surrounding scene. Most cross-view generators translate between the two in one predetermined direction, so different directions and temporal completions require separate models. In this paper, we introduce EgoU, a unified generative framework that places one egocentric and up to three exocentric views on a shared video canvas and expresses diverse generation tasks through conditioning masks and optional text prompts. Fine-tuned from a pretrained video diffusion transformer (DiT), EgoU supports exocentric-to-egocentric generation with or without an egocentric seed frame, egocentric-to-exocentric generation, temporal extension, temporal interpolation, text-driven egocentric synthesis, and joint cross-view generation within a single set of parameters. We incorporate view-type embeddings to differentiate perspectives, condition on scene geometry and the wearer's head trajectory through a dedicated cross-attention branch, and apply an auxiliary pose loss to supervise exocentric human motion. To train and evaluate this formulation, we build EgoExoVideo-600K, a corpus of 604,113 synchronized ego–exo clip pairs, and EgoU-Benchmark, which scores all tasks on the same held-out clips. To our knowledge, EgoU is the first single model to support all of these tasks through a shared interface, and it matches or surpasses task-specific baselines. EgoU is open-sourced anonymously at huggingface.co/EgoVideo/EgoU-5B.

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

1 INTRODUCTION

Egocentric and exocentric observations provide complementary perspectives on the same physical event. While a first-person view captures fine-grained, intent-driven interactions and the camera wearer's surroundings, an external third-person view reveals full-body kinematics and broader spatial context. Jointly modeling these viewpoints enables video prediction that remains geometrically and semantically consistent across perspectives (Wang et al., 2026c; Tran et al., 2026), which matters for learning embodied behavior from human video, where demonstrations are recorded in third person but an agent perceives and acts from its own viewpoint (Hao et al., 2026; Li et al., 2026; Chen et al., 2026). A generative model of both views can translate between them and simulate their synchronized evolution over time.

Recent work adapts pretrained video diffusion backbones to cross-view synthesis, but most models serve one predetermined input–output direction. Exo2Ego-V, EgoX, and Exo2EgoSyn (Liu et al., 2024; Kang et al., 2026; Mahdi et al., 2025) translate exocentric into egocentric video, WorldWander (Song et al., 2025b) fine-tunes a separate model for each direction, and Syn2Seq-Forcing (Mahdi et al., 2026) unifies both directions for a single view pair without temporal completion. This raises a question: *can a single generative framework support both conditioning directions, a variable number of viewpoints, and diverse temporal completions under a common interface?* Unified interfaces such as VACE (Jiang et al., 2025), UniVideo (Wei et al., 2026), and MultiShotMaster (Wang et al., 2026b) demonstrate this versatility for single-view generation and editing, but have not been applied to synchronized ego–exo views.

To address this gap, we introduce EgoU, a unified egocentric world simulator that organizes multiple viewpoints of an event as structured panels on a shared video canvas (Fig. 1). By casting conditioning as visual masks that indicate which views and temporal frames are observed, a single diffusion transformer learns to generate any masked target region. This formulation supports *seven* tasks with fixed weights, namely exo-to-ego translation with and without an egocentric seed frame, ego-to-exo translation, temporal extension, temporal interpolation, text-guided egocentric generation, and joint cross-view synthesis. To guide generation across camera dynamics, we introduce learned view-type embeddings to disambiguate perspective tokens, geometry-aware feature conditioning to inject spatial layout from context frames, and an auxiliary human pose loss to penalize anatomically implausible exocentric body motion. Unlike the two-panel canvas of EgoX, which serves a single task, the panel count varies and the mask is chosen independently per panel. Our contributions are threefold:

- We cast ego–exo video generation as masked completion of a multi-view canvas, so that one diffusion transformer with fixed weights covers all seven tasks through per-panel conditioning masks over one to four synchronized views.
- View-type embeddings disambiguate perspective tokens, a geometry-guided cross-attention branch injects scene structure from observed frames and the wearer's optional head trajectory, and an auxiliary pose loss supervises exocentric body motion.
- EgoExoVideo-600K provides 604,113 synchronized ego–exo clip pairs organized into 294,706 multi-view groups with captions and geometry, and EgoU-Benchmark scores all seven tasks on the same 420 unseen multi-view sequences, on which EgoU matches or surpasses specialized, task-specific baselines.

2 RELATED WORK

Egocentric–Exocentric Video Generation. Most cross-view generators are direction-specific. Early exo-to-ego work includes STA-GAN (Liu et al., 2021), with spatial and temporal adversarial generation, and Exo2Ego (Luo et al., 2024b), which lifts hand layouts across views before frame-wise diffusion. Among recent exo-to-ego methods, Exo2Ego-V (Liu et al., 2024) uses calibrated multi-view inputs, EgoExo-Gen (Xu et al., 2025) adds an initial egocentric frame and hand–object cues, and EgoX, Exo2EgoSyn, Grounded-Exo2Ego, and PCSE (Kang et al., 2026; Mahdi et al., 2025; Wang et al., 2026c;d) adapt pretrained video generators with geometric, cross-view, or semantic conditioning; 4Diff and EgoWorld (Cheng et al., 2024; Park et al., 2025) work at the image level, IDE (Luo et al., 2024a) handles the reverse direction, and other egocentric simulators condition on body motion (Bai et al., 2026; Tu et al., 2026), generate embodied interactions (Hao et al., 2026), or turn exo video into ego training data (Tran et al., 2026). WorldWander (Song et al., 2025b) covers both

directions with separately fine-tuned models and Syn2Seq-Forcing (Mahdi et al., 2026) unifies them in a sequence formulation, but neither handles a variable number of views or temporal completion. EgoU replaces these direction-specific models with one set of weights whose task is defined by a conditioning mask over synchronized views.

Unified Conditioning Interfaces. SEINE (Chen et al., 2023) conditions on random masks for transitions and prediction; VACE (Jiang et al., 2025) unifies video creation and editing through text, context frames, and spatiotemporal masks, and UniVideo (Wei et al., 2026) couples multimodal understanding with generation and editing, all on a single view. Diffusion Forcing and History-Guided Video Diffusion (Chen et al., 2024; Song et al., 2025a) condition on flexible frame subsets through per-token noise levels, CAT4D (Wu et al., 2025) and SV4D (Xie et al., 2025) model views and time for 4D reconstruction and novel-view videos of dynamic objects, and MultiShotMaster (Wang et al., 2026b) arranges sequential shots rather than synchronized views. In-context generation places condition and target in one token sequence, as image panels in IC-LoRA (Huang et al., 2024) and as concatenated source and target videos in ReCamMaster (Bai et al., 2025a); SynCamMaster (Bai et al., 2025b) generates synchronized multi-camera videos from text and camera poses without visual observations, and EgoX (Kang et al., 2026) tiles one clean exo latent beside a noisy ego latent for a single task. EgoU extends mask-based conditioning to a canvas of a variable number of synchronized ego and exo panels, where the observed frames of each panel are chosen independently, so that the same weights translate, extend, interpolate, or jointly generate views.

Geometry and Camera Conditioning. Camera-controlled generators encode trajectories as Plücker rays (He et al., 2025a;b) or re-render source dynamics along new paths (Bai et al., 2025a), while TrajectoryCrafter, GEN3C, and E^{3}C (Yu et al., 2025; Ren et al., 2025; Gu et al., 2026) render explicit point clouds or 3D caches. MultiDiff (Müller et al., 2024) combines depth-warped conditioning and structured noise for consistent novel-view synthesis, while Uni3C (Cao et al., 2025) aligns point-cloud and SMPL-X guidance to combine camera and human-motion control. REPA and Geometry Forcing (Yu et al., 2024; Wu et al., 2026) instead align diffusion features with pretrained visual or geometric representations such as VGGT and VGGT-Ω (Wang et al., 2025a; 2026a). EgoU needs no reconstruction or exocentric calibration: it cross-attends to frozen VGGT-Ω tokens of frames, adds Plücker rays for optional head trajectory, and supervises exocentric body pose during training.

3 METHOD: EGOU

As illustrated in Fig. 2, EgoU generates synchronized egocentric and exocentric videos conditioned on observed clips, seed frames, optional text prompts, and an optional egocentric camera (head-pose) trajectory. A notation table in Appendix A summarizes all symbols used in this section.

3.1 MULTI-VIEW REPRESENTATION AND TASK CONDITIONING

Multi-view Canvas. An event comprises one egocentric clip V^{ego} and K exocentric clips $V_1^{\text{exo}}, \dots, V_K^{\text{exo}}$, where $0 \le K \le 3$. All $N = 1 + K$ clips are temporally synchronized and share F frames at a unified spatial resolution. The Wan VAE compresses time by a factor of four while preserving the first frame, so a clip of F frames yields $f = 1 + (F-1)/4$ latent frames. Let c denote an optional text prompt. We arrange these views horizontally on a single canvas, placing the egocentric view first. Each view is encoded independently by a frozen Wan VAE encoder $\mathcal{E}$, and the resulting clean latents are concatenated along the spatial width axis:

$$\mathbf{Z}_0 = [\,\mathcal{E}(V^{\text{ego}}) \,\|\, \mathcal{E}(V_1^{\text{exo}}) \,\|\, \cdots \,\|\, \mathcal{E}(V_K^{\text{exo}})\,] \in \mathbb{R}^{C \times f \times h \times Nw}, \tag{1}$$

where C denotes the number of latent channels, and (f, h, w) is the latent grid size of an individual view. We refer to each contiguous view block as a *panel*. Encoding views independently keeps panel boundaries out of the VAE (Appendix B.1).

Tasks as Conditioning Patterns. Each panel $a \in \{0, \dots, N-1\}$ is assigned a conditioning pattern π_a specifying which latent frames are observed: all frames, only the first, a temporal prefix, an arbitrary subset, or none (Appendix B.2). These patterns are compiled into a binary mask $\mathbf{M} \in \{0,1\}^{f \times h \times Nw}$ (broadcast across latent channels), where 1 marks observed locations and 0 indicates regions to be generated. During training, the transformer receives the noisy canvas latent:

$$\mathbf{Z}_t = \mathbf{M} \odot \mathbf{Z}_0 + (1 - \mathbf{M}) \odot \big[(1 - \sigma_t)\mathbf{Z}_0 + \sigma_t \boldsymbol{\epsilon}\big], \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{2}$$

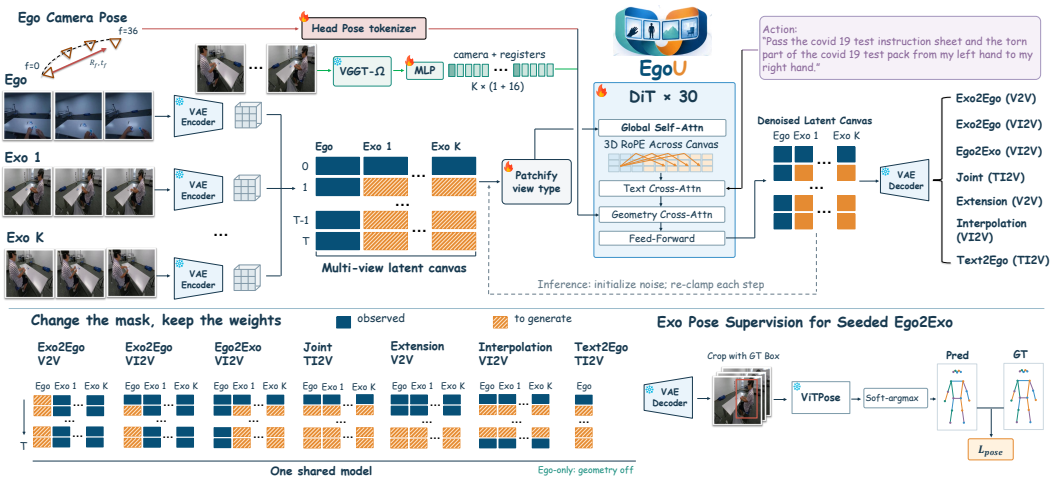


Figure 2: **Overview of the EgoU Framework.** A single diffusion transformer processes the multi-view canvas under task-defining conditioning masks, augmented by view-type embeddings, geometric scene and camera tokens, and auxiliary pose supervision.

where t is the diffusion timestep and $\sigma_t \in [0, 1]$ denotes the noise level. Observed positions remain clamped to their clean latent values during both training and inference (Appendix B.2).

Fig. 1 summarizes the task families: **Exo2Ego** observes the exocentric clips and generates the ego panel, optionally from an ego *seed frame* (the first frame of a target panel given as an observation); **Ego2Exo** observes the ego clip and generates the exocentric panels from their seed frames; **Joint Generation** synthesizes all views from one seed frame per panel; **Extension** and **Interpolation** complete all panels from temporal prefixes or boundary anchors; and **Text2Ego** generates a single ego panel from text and a seed frame. All tasks share identical parameters: the panel count, the conditioning mask, and the optional prompt define the task instance, and both panel count and task family are sampled at every training iteration.

3.2 Shared Video Generator

View-type Embedding. To distinguish ego and exo camera dynamics, we inject a learnable view-type embedding into each patch token:

$$\widetilde{\mathbf{z}}_i = \mathbf{z}_i + \mathbf{e}_{\mathrm{type}(i)}, \qquad \mathbf{e}_{\mathrm{ego}}, \mathbf{e}_{\mathrm{exo}} \in \mathbb{R}^d, \tag{3}$$

where i indexes the token, d denotes the transformer hidden dimension, and $\mathrm{type}(i) \in \{\mathrm{ego}, \mathrm{exo}\}$ is determined by the token's corresponding panel. The embeddings are added after patchification and initialized to zero, so the model starts from the pretrained representation and learns perspective-specific biases (Appendix B.3).

Global Self-attention. The first attention layer of each diffusion transformer (DiT) block is a canvas-wide self-attention over all $L = fh_pNw_p$ tokens of the canvas, where (h_p, w_p) is the per-panel patch grid and tokens are ordered by frame, row, and column. We retain the pretrained 3D rotary positional embeddings (RoPE), sharing temporal and vertical indices across panels while indexing columns across the full canvas width (Appendix B.1).

Flow-matching Objective. Let $v_\theta(\mathbf{Z}_t, t, c; \mathbf{M}, G)$ denote the predicted velocity field, conditioned on the mask $\mathbf{M}$ and the geometric context G of Sec. 3.3. Under the linear interpolation in Eq. 2, the target velocity with respect to noise level σ_t is $\boldsymbol{\epsilon} - \mathbf{Z}_0$ for generated regions. We optimize the network using the masked flow-matching loss:

$$\mathcal{L}_{\mathrm{flow}} = \mathbb{E}\left[\frac{\omega(t)}{D(\mathbf{M})}\left\|(1-\mathbf{M}) \odot \left[v_\theta(\mathbf{Z}_t, t, c; \mathbf{M}, G) - (\boldsymbol{\epsilon} - \mathbf{Z}_0)\right]\right\|_2^2\right], \tag{4}$$

where the expectation is taken over training clips, conditioning patterns, timesteps, and noise samples; $\omega(t)$ is the timestep weighting function; and $D(\mathbf{M})$ is a loss normalization factor (Appendix B.7). The factor $(1 - \mathbf{M})$ explicitly excludes observed regions from the loss.

3.3 GEOMETRY CONDITIONING AND POSE SUPERVISION

We condition on two complementary forms of 3D geometry, static scene structure derived from observed pixels and the 6-DoF egocentric camera trajectory (the wearer's head motion), which governs first-person viewpoint dynamics and is only weakly constrained by third-person observations. We tokenize both modalities onto a unified temporal axis and inject them into each transformer block via a dedicated cross-attention branch; the head trajectory also enters each egocentric token as a viewing ray. For Ego2Exo, an auxiliary pose loss supervises the wearer's body motion in the generated exocentric views.

Scene Geometry Tokens. A frozen VGGT-Ω (Wang et al., 2026a) encodes T conditioning frames sampled from the observed panels; for each frame we keep its camera token and 16 register tokens, and a trainable projection maps them to width d, yielding scene tokens $G^{\mathrm{sc}} = (\mathbf{g}_1, \ldots, \mathbf{g}_{17T})$ (Appendix B.4). Unobserved target frames are never passed to the geometry extractor. A token computed from pixel frame r_n inherits the normalized video timestamp $\tau_n = r_n/(F-1) \in [0,1]$.

Egocentric Camera Tokens. Let $(R_r, \mathbf{o}_r)$ denote the camera-to-world rotation and optical center of the egocentric camera at pixel frame $r \in \{0, \ldots, F-1\}$. For invariance across environments, we express the camera trajectory relative to the initial frame in its local coordinate system:

$$\Delta R_r = R_0^\top R_r, \qquad \Delta \mathbf{o}_r = \tfrac{1}{s} R_0^\top (\mathbf{o}_r - \mathbf{o}_0), \tag{5}$$

where s is a sequence-specific metric scale factor. This parameterization aligns head rotations and translations with the egocentric image axes regardless of the wearer's global position. A lightweight MLP maps a per-frame feature $\boldsymbol{\xi}_r$ from $(\Delta R_r, \Delta \mathbf{o}_r)$ (Appendix B.5) to transformer width, producing F camera tokens $G^{\mathrm{cam}} = (\mathbf{g}_{17T+1}, \ldots, \mathbf{g}_{17T+F})$, where token $n = 17T + 1 + r$ encodes pixel frame r with timestamp $\tau_n = r/(F-1)$. The trajectory is optional: we drop it during training with probability p_{drop} ($\delta = 0$, G^{cam} empty), so the same weights also support trajectory-free inference.

Per-token Camera Rays. The camera tokens above describe overall head translation and orientation, but they do not explicitly specify how rotation displaces individual image pixels. We therefore equip each egocentric token with an explicit viewing ray, following camera-controlled video diffusion principles (He et al., 2025a). For an egocentric token i in latent frame $\ell(i)$ with patch center (x_i, y_i), pinhole intrinsics $\mathbf{K}_{\mathrm{ego}}$, aligned pixel frame $r = 4\ell(i)$, and relative head pose $(\Delta R_r, \Delta \mathbf{o}_r)$ (Eq. 5), the ray in first-frame camera coordinates has unit direction $\mathbf{d}_i \propto \Delta R_r \, \mathbf{K}_{\mathrm{ego}}^{-1} [x_i, y_i, 1]^\top$ and moment $\mathbf{m}_i = \Delta \mathbf{o}_r \times \mathbf{d}_i$. Its Plücker coordinates $\boldsymbol{\zeta}_i = (\mathbf{m}_i, \mathbf{d}_i) \in \mathbb{R}^6$ are encoded by an MLP ψ, and the token embedding from Eq. 3 is extended to:

$$\widetilde{\mathbf{z}}_i = \mathbf{z}_i + \mathbf{e}_{\mathrm{type}(i)} + \delta \, \mathbb{1}[\mathrm{type}(i) = \mathrm{ego}] \; \psi(\boldsymbol{\zeta}_i), \tag{6}$$

where $\delta \in \{0, 1\}$ indicates trajectory availability. This makes egocentric queries pose-aware from the first transformer block: each query encodes its own line of sight, while the cross-attended camera tokens provide context on head motion across neighboring frames.

Unified Temporally Biased Cross-attention. Each transformer block incorporates a dedicated geometric cross-attention branch, positioned after but independent of the text cross-attention. Its context comprises the concatenated tokens $G = [\, G^{\mathrm{sc}} \,\|\, G^{\mathrm{cam}} \,] = (\mathbf{g}_1, \ldots, \mathbf{g}_{|G|})$ with $|G| = 17T + \delta F$. For a given attention head, let $\mathbf{q}_i, \mathbf{k}_n, \mathbf{v}_n \in \mathbb{R}^{d_h}$ represent the query of canvas token i and the key and value of context token n, respectively. Let $\ell(i)$ denote canvas token i's latent frame, and $u_i = \ell(i)/(f-1) \in [0,1]$ its normalized video timestamp (distinct from diffusion timestep t). Let $\mathrm{src}(n) \in \{\mathrm{sc}, \mathrm{cam}\}$ indicate the source domain of context token n. The cross-attention weights are formulated as:

$$A_{in} = \frac{\mathbf{q}_i^\top \mathbf{k}_n}{\sqrt{d_h}} - \beta_{\mathrm{src}(n)} (u_i - \tau_n)^2 + \mu_{in}, \qquad \alpha_{in} = \frac{\exp(A_{in})}{\sum_{n'=1}^{|G|} \exp(A_{in'})}, \tag{7}$$

where $\mu_{in} = -\infty$ if n is a camera token and i an exocentric token, and $\mu_{in} = 0$ otherwise. Scene and camera tokens share queries, normalization, and projection layers within this unified branch, differing only in their encoders and two structural biases. The temporal scales $\beta_{\mathrm{sc}}, \beta_{\mathrm{cam}} > 0$, learned per block, dampen attention to context tokens that are temporally distant from the query, which accommodates sparse scene tokens while letting the dense camera tokens act as frame-aligned guidance. The visibility mask μ_{in} reflects that the exocentric views are captured by other cameras, so the wearer's head trajectory is visible only to egocentric queries, which balance scene layout against

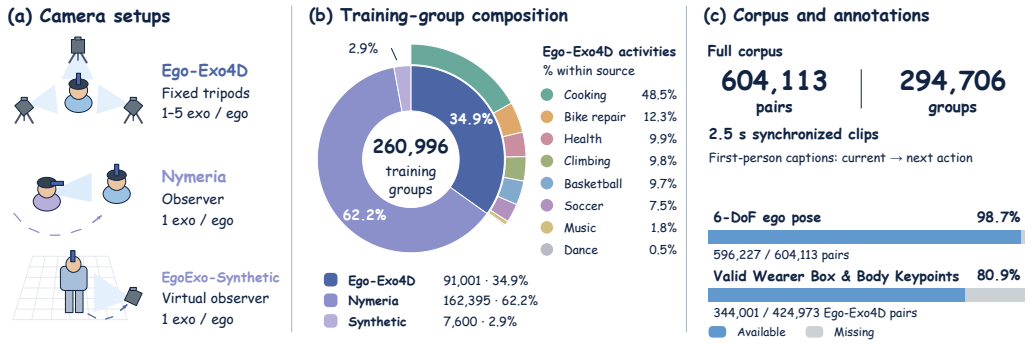


Figure 3: Composition of EgoExoVideo-600K.

camera motion within a single softmax. The branch and the ray MLP ψ are zero-initialized, so that, like the view-type embeddings, they leave the pretrained generator unchanged at the start of training; parameterization and initialization details are given in Appendices B.4 and B.5.

Pose-consistency Supervision. In Ego2Exo tasks, the synthesized exocentric views must accurately depict the camera wearer's full-body pose. Because the latent flow-matching loss does not explicitly penalize anatomical distortion, we introduce an auxiliary pose loss on generated exocentric frames. We form the one-step estimate $\widehat{\mathbf{Z}}_0$ of the clean canvas from the predicted velocity, decode a generated exocentric frame with the frozen VAE decoder $\mathcal{D}$, crop the wearer's bounding box, and run a frozen top-down pose estimator ϕ (ViTPose (Xu et al., 2022)) with a differentiable soft-argmax to obtain $J = 17$ joint $\widehat{\mathbf{p}} \in [0,1]^{J\times 2}$ in the crop-normalized space of targets $\mathbf{p}$. The pose loss is:

$$\ell_{\text{pose}} = \frac{\sum_{j=1}^{J} \nu_j \|\widehat{\mathbf{p}}_j - \mathbf{p}_j\|_2^2}{\max\left(\sum_{j=1}^{J} \nu_j, \varepsilon_{\text{num}}\right)}, \qquad \mathcal{L} = \mathcal{L}_{\text{flow}} + \lambda\, \mathbb{E}[b\, \ell_{\text{pose}}], \tag{8}$$

where ν_j is the visibility of joint j, $\varepsilon_{\text{num}} > 0$ guards the normalization, and $b \in \{0,1\}$ activates the loss on Ego2Exo samples at low noise levels. Gradients flow through the frozen estimator and decoder into the transformer; the module is training-only and adds no inference cost (Appendix B.6).

4 EgoExoVideo-600K Dataset

Training a single model for the tasks in Fig. 1 requires synchronized egocentric and exocentric video at scale, with captions and geometry that cover every conditioning pattern. Existing paired collections are either small, restricted to one camera arrangement, or lack the annotations that several of our tasks need. We therefore assemble **EgoExoVideo-600K**, a corpus of 604,113 synchronized ego–exo clip pairs, and evaluate every task on a held-out portion of it. Fig. 5 in Appendix C gives the construction details, statistics, and samples.

Sources and pairing. The corpus draws on three sources with complementary exocentric setups. Ego-Exo4D contributes skilled activities in 13 categories (basketball, cooking, bouldering, bike repair, COVID test, dance, soccer, rock climbing, CPR, PCR, music, COVID swab, and omelet) captured by Aria glasses and three to five static GoPro cameras, so one egocentric clip typically pairs with three to five exocentric views. Nymeria contributes daily activities in homes, where the exocentric view is a second observer wearing the same glasses and following the participant, so the exocentric camera moves. EgoExo-Synthetic contributes rendered sequences with a virtual observer camera. All clips last 2.5 s. For our canvas, one egocentric clip and up to three of its exocentric clips form a group; the corpus yields 294,706 groups.

Captions and geometry. Every pair contains a first-person caption in the present tense that names the current action and the action that follows, written from the wearer's perspective. The captions are rewritten by Gemini 3.6 Flash because the original atomic-action annotations of Ego-Exo4D and Nymeria are sometimes misaligned with the clip; the resulting caption set is $50\times$ larger than that of EgoX. The corpus also provides the geometry that the tasks in Sec. 3 rely on: a 6-DoF trajectory of the egocentric device for every Ego-Exo4D and Nymeria clip, and, for Ego-Exo4D, a bounding box and 2D body pose of the camera wearer in each exocentric view together with the exocentric camera

calibration. The pose annotations supervise the pose-consistency loss, and the calibration relates each exocentric camera to the egocentric trajectory.

Processing and splits. Because the egocentric streams are fisheye while the pretrained generator and the geometry encoder expect pinhole imagery, we re-project all egocentric streams to 480×480 rectilinear video and undistort the exocentric streams at 854×480. The model consumes 37 frames at 15 fps per panel, decimated from the 75-frame 30 fps clips, with exocentric frames center-cropped to the square panel. Train and test are separated at the take level, with no recording shared across splits. The evaluation set used in Sec. 5, EgoU-Benchmark, selects 420 Ego-Exo4D test groups from distinct takes with a proportional quota over the 13 activity categories and complete geometry, so that every task family can be scored on the same clips. No recording of these takes appears in training, so every benchmark clip is unseen by the model.

5 Experiments

5.1 Experimental Setup

Implementation Details. EgoU is fine-tuned from Wan2.2-TI2V-5B (Wan et al., 2025) on canvases of 37 frames at 480×480 and 15 fps per panel, updating all transformer weights with AdamW (learning rate 10^{-5}, bf16) on 4 RTX PRO 6000 GPUs. The task family and the number of panels are sampled at every iteration, and the head trajectory is dropped with probability 0.4 so that the same weights serve trajectory-free inference. Inference uses classifier-free guidance 5.0 (2.0 for Ego2Exo). Full training and sampling details are given in Appendix B.7.

Evaluation Protocol. We evaluate on **EgoU-Benchmark**. All task families in Table 1 are instantiated on the same clips by changing only the per-panel reference, mask, and text conditions. Each baseline is initialized from its publicly released checkpoint, where available, and fine-tuned for one epoch on the training set of EgoExoVideo-600K. Video quality is measured by FVD, PSNR, SSIM, LPIPS, CLIP-I, and DINO on all 420 groups; for families that generate both views, the reported value is the mean over the egocentric and exocentric panels.

Cross-view Alignment (Align). Pixel and feature metrics do not test whether the generated view shows the *same event* as the given one, so we add a VLM-judged score in the spirit of question-answering evaluation (Hu et al., 2023; Cho et al., 2024). For each clip, Gemini 3.8 Flash (Google Deepmind, 2025) writes four-way multiple-choice questions about the action, the active object, hand–object contact, the end state of the clip, and the wearer's head or body motion, from the Ego-Exo4D narrations and frames of the ground-truth egocentric and exocentric views, with distractors drawn from other moments of the same recording. A question is kept only if the judge, Qwen3.8-Flash-Next (Qwen Team, 2026), answers it correctly on the ground-truth target view and incorrectly without any video, so that a retained question can be answered only by watching; Align is the judge's accuracy on the retained questions for the generated view. It is computed on an activity-stratified random subset of 20 benchmark groups (47 egocentric and 34 exocentric questions) with the same questions for every method, and averaged over the two panels for families that generate both views.

User Study. Sixteen raters scored the same 20-group subset for every method. Each trial shows the conditioning input next to the generated clip with the method hidden, and is rated on 1–5 scales for consistency with the input and description (Prompt), temporal coherence (Temp.), visual quality (Quality), and overall impression (Overall). Every clip is rated by up to three raters (1,084 ratings in total); ground-truth clips inserted as attention checks received a mean rating of at least 4.25 from every rater. Inter-rater agreement on Overall is Krippendorff's $\alpha = 0.41$ and ICC(2,k) $= 0.71$, and Align agrees with the human ratings: across the 25 method rows of Table 1 it correlates with MOS Overall at Spearman $\rho = 0.70$ (FVD: $\rho = -0.79$). Design and statistics are given in Appendix E.

Baselines. As no all-in-one ego–exo simulator is available, we compare our single model against task-specific methods in each family, and in every family except unseeded Exo2Ego and Interpolation we also report the Wan2.2 (Wan et al., 2025) TI2V-5B and I2V-A14B image-to-video models, fine-tuned as above. For *Exo2Ego*, we compare against Exo2Ego-V (Liu et al., 2024), EgoX (Kang et al., 2026), and EgoExo-Gen (Xu et al., 2025), as well as the bidirectional WorldWander (Song et al., 2025b). EgoX renders an ego prior along the ground-truth ego camera trajectory and Exo2Ego-V reads the ground-truth ego pose; our model is given the same trajectory. For *Ego2Exo*, WorldWander is the only

Table 1: Quantitative comparison between different methods on EgoU-Benchmark.

Type	Method	Video Quality & Video Consistency							User Study (MOS)			
		FVD↓	PSNR↑	SSIM↑	LPIPS↓	CLIP-I↑	DINO↑	Align↑	Prompt↑	Temp.↑	Quality↑	Overall↑
Exo2Ego (V2V)	Exo2Ego-V	390.0	**16.22**	**0.592**	0.583	0.728	0.394	0.32	2.36	1.74	1.88	1.67
	WorldWander	309.9	13.63	0.474	0.681	0.733	0.356	0.43	2.45	2.39	2.29	2.00
	EgoX	403.0	14.15	0.492	0.619	0.779	0.564	0.47	3.07	3.16	2.95	2.95
	Ours (EgoU)	**220.5**	15.35	0.528	**0.564**	**0.834**	**0.652**	**0.70**	**4.48**	**4.14**	**3.89**	**4.14**
Exo2Ego (VI2V)	Wan2.2-5B TI2V	764.2	15.05	0.462	0.572	0.834	0.692	0.55	2.27	2.20	2.33	2.18
	Wan2.2-A14B I2V	187.2	16.86	0.518	0.511	**0.882**	**0.779**	0.68	3.64	3.43	3.45	3.38
	EgoExo-Gen	224.7	15.11	0.500	0.569	0.782	0.533	0.34	2.91	2.69	2.51	2.44
	Ours (EgoU)	**118.3**	**18.38**	**0.606**	**0.392**	0.873	0.774	**0.77**	**4.35**	**4.26**	**3.89**	**4.09**
Ego2Exo (VI2V)	Wan2.2-5B TI2V	381.9	19.53	0.639	0.191	0.891	0.892	0.53	2.88	2.85	2.98	2.63
	Wan2.2-A14B I2V	144.3	22.62	0.821	0.105	0.916	0.943	0.68	4.02	**3.85**	**3.76**	**3.78**
	WorldWander	129.7	20.80	0.690	0.138	0.902	0.927	0.68	3.59	3.62	3.67	3.59
	Ours (EgoU)	**101.5**	**25.04**	**0.861**	**0.074**	**0.926**	**0.947**	**0.85**	**4.08**	3.79	3.56	3.71
Extension (V2V)	Wan2.2-5B TI2V	353.5	17.89	0.626	0.367	0.865	0.778	0.73	3.03	1.81	2.61	2.13
	Wan2.2-A14B I2V	203.8	18.55	0.665	0.342	0.890	0.823	0.71	3.81	3.48	3.70	3.63
	Ours (EgoU)	**131.1**	**21.36**	**0.733**	**0.243**	**0.910**	**0.879**	**0.77**	**4.54**	**4.20**	**4.29**	**4.29**
Interpolation (VI2V)	Wan2.1-14B FLF2V	208.4	19.38	0.674	0.306	0.879	0.819	0.67	3.03	2.53	2.88	2.56
	Framer	211.1	**23.29**	0.723	0.269	0.899	0.841	0.73	3.00	2.48	2.85	2.59
	SEINE	625.1	22.94	0.714	0.272	0.888	0.821	0.68	2.58	1.97	2.16	2.00
	Ours (EgoU)	**104.2**	22.66	**0.755**	**0.203**	**0.907**	**0.879**	**0.80**	**4.47**	**3.97**	**4.03**	**4.20**
Text2Ego (TI2V)	Wan2.2-5B TI2V	409.7	13.10	0.441	0.700	0.807	0.652	0.47	2.24	2.17	2.47	2.02
	Wan2.2-A14B I2V	254.2	14.21	0.465	**0.645**	0.844	0.686	0.64	3.77	3.53	**3.90**	3.66
	Ours (EgoU)	**129.3**	**14.51**	**0.524**	0.649	**0.848**	**0.744**	**0.72**	**4.11**	**3.61**	3.61	**3.68**
Joint Cross-View (TI2V)	Wan2.2-5B TI2V	294.6	17.57	0.590	0.378	0.853	0.753	0.53	2.91	2.93	3.09	2.68
	Wan2.2-A14B I2V	208.1	18.76	0.652	0.337	0.882	0.811	0.69	3.84	3.61	3.52	3.50
	Ours (EgoU)	**151.2**	**20.50**	**0.712**	**0.262**	**0.887**	**0.832**	**0.76**	**4.09**	**3.89**	**3.59**	**3.76**

Table 2: **Ablation on the cross-view families.** Each row is a separately trained model, evaluated on the 420 groups of EgoU-Benchmark. All checkpoints are trained with 10,000 steps in the training set.

Configuration	Ego2Exo (VI2V)					Exo2Ego (V2V)				
	PSNR↑	SSIM↑	LPIPS↓	CLIP-I↑	DINO↑	PSNR↑	SSIM↑	LPIPS↓	CLIP-I↑	DINO↑
Wan2.2-5B TI2V, Direct Training	19.53	0.639	0.191	0.891	0.892	12.74	0.358	0.739	0.763	0.461
+ Canvas Training	23.59	0.836	0.095	0.919	0.935	14.29	0.491	0.614	0.794	0.525
+ Cross-Attention (Exo Geometry)	24.05	0.843	0.088	0.920	0.938	14.74	0.508	0.594	0.810	0.573
+ Exo Pose Supervision	24.48	0.849	0.082	0.923	0.944	15.31	0.529	0.575	0.833	0.649
+ Cross-Attention (Head Pose)	24.50	0.850	0.081	0.922	0.943	15.35	0.528	0.567	0.834	0.651

specialized open-source baseline. For *Extension* and *Text2Ego*, no specialized open-source baseline exists, so we compare against the Wan2.2 models only. For *Interpolation*, we compare against the first–last-frame model Wan2.1-FLF2V-14B (Wan et al., 2025) and the diffusion-based models Framer (Wang et al., 2025b) and SEINE (Chen et al., 2023). For *Joint* generation, no specialized open-source baseline exists, so we run Wan2.2 separately on the egocentric and exocentric views. Grounded-Exo2Ego (Wang et al., 2026c), Exo2EgoSyn (Mahdi et al., 2025), EgoExo-WM (Tran et al., 2026), PEVA (Bai et al., 2026), and PlayerOne (Tu et al., 2026) are not compared because their weights are not publicly available.

5.2 Main Results

Quantitative Evaluation. Table 1 reports a single EgoU checkpoint on all seven task rows of EgoU-Benchmark; every row is produced by the same weights and differs only in the per-panel conditioning pattern. The checkpoint is the full configuration of Sec. 3, with the ground-truth egocentric trajectory supplied at test time, as it is to the pose-conditioned baselines. EgoU attains the lowest FVD in all seven families and the highest Align in all seven. Against the fine-tuned Wan2.2-5B TI2V under the same conditioning, it improves Ego2Exo by 5.5 dB PSNR, seeded Exo2Ego by 3.3 dB, and Text2Ego by 1.4 dB. Unseeded Exo2Ego, the only row in which no target-view pixel is given, remains the hardest setting: our FVD of 221 compares with 310 for WorldWander, 390 for Exo2Ego-V, and 403 for EgoX, and the higher PSNR of Exo2Ego-V comes with an Align of 0.32 against our 0.70. On interpolation, Framer and SEINE reach a higher PSNR by blending the two anchors, while EgoU completes the motion and leads on SSIM, LPIPS, and FVD. The contribution of each component and of the trajectory itself is isolated in Table 2.

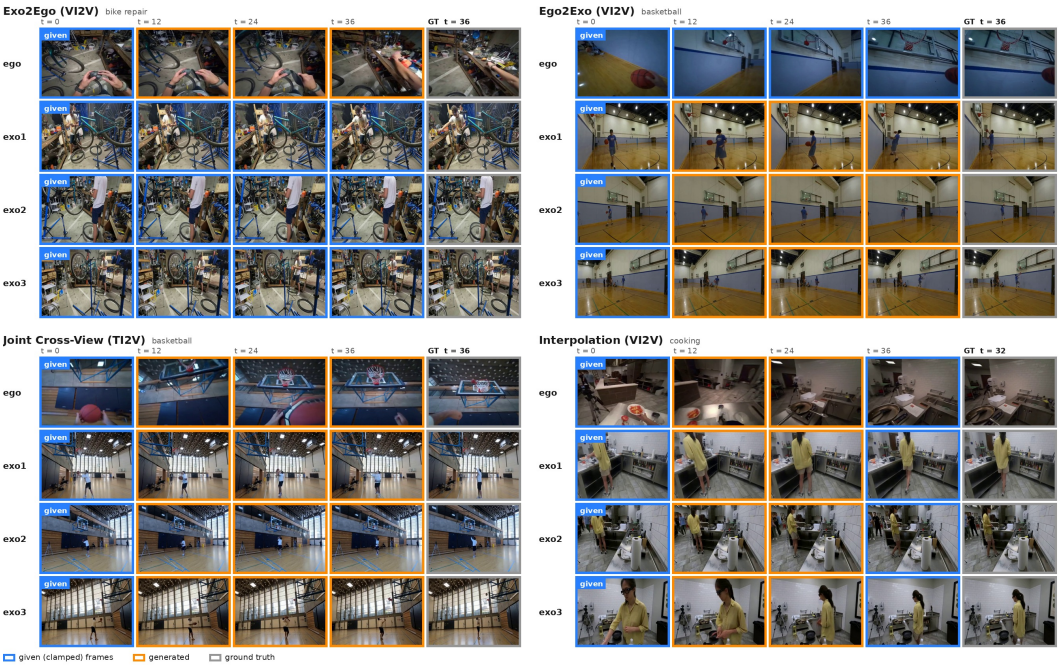


Figure 4: **Qualitative results on EgoU-Benchmark.** Four task families on four held-out takes with large camera or body motion, generated by the model of Table 1; rows are the panels of one canvas (ego, exo1–exo3), columns are frames $t = 0, 12, 24, 36$ of the 37-frame clip, and the last column is the ground truth at the last generated frame. Blue outlines mark given (clamped) frames and orange outlines generated frames.

Qualitative Results. Fig. 4 shows one example per task family, chosen among the groups with the most camera and body motion. The generated views keep the scene layout and the camera-wearer's position consistent with the given views, and the target-view motion follows the given context. A comparison against the baselines on the same inputs is given in Appendix F.

5.3 Ablation Study

Table 2 adds the components of EgoU one at a time on Ego2Exo and unseeded Exo2Ego, the two cross-view families in which the target view's motion must be inferred from the other view; each row is a separately trained model, trained for 10,000 steps and evaluated on all 420 groups at one guidance scale. Training on the canvas is what turns the per-view model into a cross-view generator ($+4.1$ dB on Ego2Exo, $+1.6$ dB on unseeded Exo2Ego). The exocentric geometry branch and the pose loss then add 0.5 and 0.6 dB on Exo2Ego, with LPIPS falling from 0.614 to 0.575, and 0.5 and 0.4 dB on Ego2Exo.

6 Conclusion

We presented EgoU, a unified video diffusion transformer that generates synchronized egocentric and exocentric video by completing a masked multi-view canvas. A single set of weights covers exo-to-ego and ego-to-exo translation, temporal extension and interpolation, text-guided egocentric generation, and joint cross-view synthesis, with the task defined only by which panels and frames are observed. View-type embeddings, a geometry-guided cross-attention over scene tokens and the wearer's head trajectory, and an auxiliary exocentric pose loss keep the generated views consistent across perspectives. Trained on EgoExoVideo-600K and evaluated on EgoU-Benchmark, EgoU attains the lowest FVD in every task family and matches or surpasses task-specific baselines in cross-view alignment and human ratings, while replacing a collection of direction-specific models with one. The current model operates on 2.5 s clips with at most three exocentric views, and its egocentric generation benefits from a head trajectory that is not always available at test time. Longer horizons, more views, and predicting the trajectory itself are natural extensions of the canvas formulation.

AI USE STATEMENT

We used Gemini and Qwen models to assist with dataset curation and evaluation. For dataset construction, Gemini 3.6 Flash rewrote and standardized the first-person action captions, and Gemini 3.8 Flash identified the objects of attention in the head-motion descriptions and captioned the EgoExo-Synthetic sequences (Appendix C.3). For evaluation, Gemini 3.8 Flash wrote the multiple-choice questions of the Align metric, and Qwen3.8-Flash-Next filtered these questions and judged the generated videos (Appendix D). These models provide captions and evaluation judgments; the camera trajectories, calibration, wearer boxes, and body-pose annotations come from the source recordings and non-generative estimators (Appendix C.4). We also used generative AI assistants (Claude Code, OpenAI Codex) to refine manuscript text and software code, check mathematical definitions against the implementation, search for related work, and provide feedback on experimental methodology. The authors manually reviewed the AI-assisted work and take responsibility for the final text, mathematical statements, annotations, evaluation protocol, experimental claims, and artifacts, including those developed with AI assistance.

ETHICS STATEMENT

EgoExoVideo-600K is derived from Ego-Exo4D (Grauman et al., 2024), Nymeria (Ma et al., 2024), and the EgoExo-Synthetic sequences of Song et al. (2025b). Access to and use of the source data remain governed by their licenses, and the more restrictive term governs where they conflict. The corpus is released for research use, requires citation of the upstream sources, prohibits attempts to identify recorded people, and keeps provenance identifiers intact (Appendix C.8). Captions, wearer boxes, and body poses are produced by automated tools with targeted manual review (Appendix C) and can retain errors and biases of the source data and of these tools. The user study was approved by the institutional review board (IRB) of our institution, and every rater gave informed consent; raters are members of our laboratory who did not work on the generators and saw no identifying information about the methods (Appendix E).

EgoU generates realistic video of people from other viewpoints. It can support research on embodied perception, robot learning, and assistive systems, but it can also be misused to fabricate footage of real people or to infer what a person saw from third-person cameras without their consent. Generated views are predictions and do not record what actually happened. Applications should respect informed consent, privacy, and the data-use restrictions of the source datasets, and generated content should be disclosed as synthetic.

REPRODUCIBILITY STATEMENT

Section 3 and Appendix B describe the canvas representation, the conditioning patterns and task sampling, the geometry cross-attention, the camera conditioning, the pose-consistency loss, and the optimization and loss normalization; Table 3 collects the notation. Section 4 and Appendix C document the sources, clip extraction, rectification, captions, geometry annotations, splits, training mixture, and the construction of EgoU-Benchmark. Section 5 and Appendices D and E specify the evaluation protocol, the Align question generation, filtering, and scoring, and the user-study design and statistics. The trained model weights are publicly available in an anonymous repository.

We will release the EgoExoVideo-600K annotations and EgoU-Benchmark together with the Align question sets and judge outputs, subject to the licenses and access requirements of the source datasets.

REFERENCES

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. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14834–14844. IEEE, 2025a.

Jianhong Bai, Menghan Xia, Xintao Wang, Ziyang Yuan, Zuozhu Liu, Haoji Hu, Pengfei Wan, and Di Zhang. Syncammaster: Synchronizing multi-camera video generation from diverse viewpoints. In *International Conference on Learning Representations*, volume 2025, pp. 58038–58060, 2025b.

Yutong Bai, Danny Tran, Amir Bar, Yann LeCun, Trevor Darrell, and Jitendra Malik. Whole-body conditioned egocentric video prediction. *Advances in Neural Information Processing Systems*, 38: 164375–164418, 2026.

Chenjie Cao, Jingkai Zhou, Shikai Li, Jingyun Liang, Chaohui Yu, Fan Wang, Xiangyang Xue, and Yanwei Fu. Uni3c: Unifying precisely 3d-enhanced camera and human motion controls for video generation. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–12, 2025.

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

Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Seine: Short-to-long video diffusion model for generative transition and prediction. In *The Twelfth International Conference on Learning Representations*, 2023.

Yang Chen, Tianqi Wang, Xiaorui Jiang, Yilei Man, Yihua Shao, Mengyuan Liu, Zhi Chen, Xiaofeng Cao, Qibin Zhao, Chi Harold Liu, et al. Human-centric intelligence in the era of foundation models: A survey. *arXiv preprint arXiv:2608.18184*, 2026.

Feng Cheng, Mi Luo, Huiyu Wang, Alex Dimakis, Lorenzo Torresani, Gedas Bertasius, and Kristen Grauman. 4diff: 3d-aware diffusion model for third-to-first viewpoint translation. In *European Conference on Computer Vision*, pp. 409–427. Springer, 2024.

Jaemin Cho, Yushi Hu, Jason Baldridge, Roopal Garg, Peter Anderson, Ranjay Krishna, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. In *International conference on learning representations*, volume 2024, pp. 15625–15645, 2024.

Google Deepmind. Gemini 3 flash model card. `https://deepmind.google/models/model-cards/gemini-3-flash/`, dec 2025. Official system card.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19383–19400, 2024.

Qiao Gu, Lingni Ma, Adam W Harley, Richard Newcombe, Florian Shkurti, and Julian Straub. E3c: Video generation with 3d environmental memory and ego-exo human pose control. *arXiv preprint arXiv:2605.26316*, 2026.

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

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for video diffusion models. In *International Conference on Learning Representations*, volume 2025, pp. 100433–100464, 2025a.

Hao He, Ceyuan Yang, Shanchuan Lin, Yinghao Xu, Meng Wei, Liangke Gui, Qi Zhao, Gordon Wetzstein, Lu Jiang, and Hongsheng Li. Cameractrl ii: Dynamic scene exploration via camera-controlled video diffusion models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13416–13426. IEEE, 2025b.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *2023 ieee/cvf international conference on computer vision (iccv)*, pp. 20349–20360. IEEE, 2023.

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.

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

Taewoong Kang, Kinam Kim, Dohyeon Kim, Minho Park, Junha Hyung, and Jaegul Choo. Egox: Egocentric video generation from a single exocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11116–11126, 2026.

Rawal Khirodkar, He Wen, Julieta Martinez, Yuan Dong, Zhaoen Su, and Shunsuke Saito. Sapiens2. 2026:58484–58507, 2026.

Zhiyuan Li, Wenyan Yang, Wenshuai Zhao, Yue Ma, Yuanpeng Tu, Pekka Marttinen, and Joni Pajarinen. Bridging the embodiment gap: Disentangled cross-embodiment video editing. *arXiv preprint arXiv:2605.03637*, 2026.

Gaowen Liu, Hao Tang, Hugo M Latapie, Jason J Corso, and Yan Yan. Cross-view exocentric to egocentric video synthesis. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 974–982, 2021.

Jia-Wei Liu, Weijia Mao, Zhongcong Xu, Jussi Keppo, and Mike Z Shou. Exocentric-to-egocentric video generation. *Advances in Neural Information Processing Systems*, 37:136149–136172, 2024.

Hongchen Luo, Kai Zhu, Wei Zhai, and Yang Cao. Intention-driven ego-to-exo video generation. *arXiv preprint arXiv:2403.09194*, 2024a.

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

Lingni Ma, Yuting Ye, Fangzhou Hong, Vladimir Guzov, Yifeng Jiang, Rowan Postyeni, Luis Pesqueira, Alexander Gamino, Vijay Baiyya, Hyo Jin Kim, et al. Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In *European Conference on Computer Vision*, pp. 445–465. Springer, 2024.

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. *arXiv preprint arXiv:2511.20186*, 2025.

Mohammad Mahdi, Nedko Savov, Danda Pani Paudel, and Luc Van Gool. From synchrony to sequence: Exo-to-ego generation via interpolation. In *European Conference on Computer Vision*, pp. 134–150. Springer, 2026.

Norman Müller, Katja Schwarz, Barbara Rössle, Lorenzo Porzi, Samuel Rota Bulò, Matthias Nießner, and Peter Kontschieder. Multidiff: Consistent novel view synthesis from a single image. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10258–10268. IEEE, 2024.

Junho Park, Andrew Sangwoo Ye, and Taein Kwon. Egoworld: Translating exocentric view to egocentric view using rich exocentric observations. *arXiv preprint arXiv:2506.17896*, 2025.

Qwen Team. Qwen3.8-Flash-Next: A new architecture, towards ultimate cost-efficiency, August 2026. URL `https://qwen.ai/blog?id=qwen3.8-flash-next`.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6121–6132. IEEE, 2025.

Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion. *arXiv preprint arXiv:2502.06764*, 2025a.

Quanjian Song, Yiren Song, Kelly Peng, Yuan Gao, and Mike Zheng Shou. Worldwander: Bridging egocentric and exocentric worlds in video generation. *arXiv preprint arXiv:2511.22098*, 2025b.

Danny Tran, Roberto Martín-Martín, and Kristen Grauman. Egoexo-wm: Unlocking exo video for ego world models. *arXiv preprint arXiv:2605.15477*, 2026.

Yuanpeng Tu, Hao Luo, Xi Chen, Xiang Bai, Fan Wang, and Hengshuang Zhao. Playerone: Egocentric world simulator. *Advances in Neural Information Processing Systems*, 38:145235–145261, 2026.

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

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306. IEEE, 2025a.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. *arXiv preprint arXiv:2605.15195*, 2026a.

Qinghe Wang, Xiaoyu Shi, Baolu Li, Weikang Bian, Quande Liu, Huchuan Lu, Xintao Wang, Pengfei Wan, Kun Gai, and Xu Jia. Multishotmaster: A controllable multi-shot video generation framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16268–16278, 2026b.

Shengze Wang, Michael Stengel, Tianye Li, Seonwook Park, Amrita Mazumdar, Koki Nagano, Alex Trevithick, and Shalini De Mello. Grounded-exo2ego: Structured semantic grounding for robust exocentric-to-egocentric video generation. *arXiv preprint arXiv:2608.20534*, 2026c.

Wen Wang, Qiuyu Wang, Kecheng Zheng, Hao Ouyang, Zhekai Chen, Biao Gong, Hao Chen, Yujun Shen, and Chunhua Shen. Framer: Interactive frame interpolation. In *International Conference on Learning Representations*, volume 2025, pp. 51601–51628, 2025b.

Xingyue Wang, Weipeng Hu, Jiun Tian Hoe, Jianhui Li, Ping Hu, and Yap-Peng Tan. Progressive temporal compensation and semantic enhancement for exo-to-ego video generation. *Information Fusion*, pp. 104117, 2026d.

Cong Wei, Quande Liu, Zixuan Ye, Qiulin Wang, Xintao Wang, Pengfei Wan, Kun Gai, and Wenhu Chen. Univideo: Unified understanding, generation, and editing for videos. In *International Conference on Learning Representations*, volume 2026, pp. 113905–113933, 2026.

Haoyu Wu, Diankun Wu, Tianyu He, Junliang Guo, Yang Ye, Yueqi Duan, and Jiang Bian. Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling. In *International Conference on Learning Representations*, volume 2026, pp. 119186–119209, 2026.

Rundi Wu, Ruiqi Gao, Ben Poole, Alex Trevithick, Changxi Zheng, Jonathan T Barron, and Aleksander Holynski. Cat4d: Create anything in 4d with multi-view video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26057–26068, 2025.

Yiming Xie, Chun-Han Yao, Vikram Voleti, Huaizu Jiang, and Varun Jampani. Sv4d: Dynamic 3d content generation with multi-frame and multi-view consistency. In *International Conference on Learning Representations*, volume 2025, pp. 33421–33441, 2025.

Jilan Xu, Yifei Huang, Baoqi Pei, Junlin Hou, Qingqiu Li, Guo Chen, Yuejie Zhang, Rui Feng, and Weidi Xie. Egoexo-gen: Ego-centric video prediction by watching exo-centric videos. *arXiv preprint arXiv:2504.11732*, 2025.

Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. volume 35, pp. 38571–38584, 2022.

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

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5738–5746. IEEE, 2019.

APPENDIX

A Notation **16**

B Model Details **16**

B.1 Canvas and Tokenization 16

B.2 Conditioning Patterns and Task Sampling 16

B.3 View-Type Embeddings 18

B.4 Geometry-Guided Cross-Attention 18

B.5 Egocentric Camera Conditioning 19

B.6 Pose-Consistency Loss 20

B.7 Optimization and Loss Normalization 20

C EgoExoVideo-600K Dataset Details **21**

C.1 Sources and Clip Extraction 21

C.2 Rectification and Integrity Checks 21

C.3 Captions 22

C.4 Geometry Annotations 23

C.5 Statistics 23

C.6 Splits and EgoU-Benchmark 23

C.7 Training Mixture 24

C.8 Release and Licensing 24

D VLM-as-a-Judge Alignment Protocol **24**

E User Study Design **25**

F Additional Qualitative Results **25**

G Limitations **26**

A NOTATION

Table 3 collects the symbols used in Sec. 3 and Appendix B. Bold lower-case symbols are vectors and bold upper-case symbols are tensors or matrices; rotations are written R. Two notions of time appear and are kept apart throughout: the diffusion timestep t with noise level σ_t, and the normalized *video* time $u_i, \tau_n \in [0, 1]$ of canvas and context tokens.

B MODEL DETAILS

This appendix specifies the representation, conditioning, auxiliary components, and optimization used in Sec. 3.

B.1 CANVAS AND TOKENIZATION

Each panel contains $F = 37$ frames at 480×480 pixels and 15 fps. The frozen Wan 2.2 VAE downsamples spatially by a factor of 16 and temporally by a factor of 4, retaining a separate initial frame. Thus, each panel has $f = 1 + (F - 1)/4 = 10$ latent frames and spatial dimensions $h = w = 30$. In the temporal indexing convention, latent frame 0 corresponds to pixel frame 0, and latent index $\ell \geq 1$ is associated with pixel indices $4\ell - 3, \ldots, 4\ell$. This grouping specifies temporal resolution, rather than disjoint receptive fields: the causal VAE can incorporate earlier frames.

Views are encoded independently and concatenated along latent width. Before decoding, the canvas is split into its constituent panels, which are also decoded independently. The VAE therefore never processes a boundary between views. Panels share temporal indices and row coordinates. A patch size of $(1, 2, 2)$ gives $h_p = w_p = 15$, with $L = 2250N$ tokens per canvas. Tokens are ordered by frame, row, and column. For zero-based token index i, the latent-frame and panel indices are

$$\ell(i) = \left\lfloor \frac{i}{h_p N w_p} \right\rfloor, \qquad a(i) = \left\lfloor \frac{i \bmod (N w_p)}{w_p} \right\rfloor. \tag{9}$$

Panel 0 is egocentric. The Wan 2.2 TI2V-5B transformer has 30 blocks, width $d = 3072$, and 24 attention heads of width $d_h = 128$. Its rotary encoding allocates 44, 42, and 42 head dimensions to time, row, and column, respectively. We retain this allocation and use absolute column positions across the canvas.

B.2 CONDITIONING PATTERNS AND TASK SAMPLING

The conditioning pattern π_a of a panel (Sec. 3.1) takes one of five forms, all defined on latent frames: `all` observes the complete clip; `first` observes only latent frame 0; $\texttt{prefix}_k$ observes the initial k latent frames; $\texttt{frames}_S$ observes an arbitrary subset S of latent frames; and `none` leaves the panel entirely unobserved. Exo2Ego uses `none` or `first` on the ego panel and `all` on the exocentric panels; Ego2Exo uses `all` and `first`; joint generation and Text2Ego use `first` on every panel; extension and interpolation use $\texttt{prefix}_k$ and $\texttt{frames}_S$ on every panel. For $\texttt{prefix}_k$, observed indices are $\{0, \ldots, k - 1\}$; for $\texttt{frames}_S$, they are the explicitly specified indices in $S \subseteq \{0, \ldots, f - 1\}$. Observed latent values are held fixed throughout denoising and excluded from the flow loss. Following the base model's image-conditioning interface, diffusion timesteps are assigned per token: tokens whose latent cells are all observed receive timestep 0, and the remaining tokens receive t. At inference, unobserved positions are initialized from standard Gaussian noise, and observed positions are re-clamped to their clean observed values after every denoising step. A single initial image can be encoded as latent frame 0. A contiguous pixel prefix of length $1 + 4(k - 1)$ provides the first k latent frames under the causal encoding convention.

Interpolation anchors are latent frames of the encoded clip, each summarizing a four-frame pixel block. During training, anchor latents are sliced from the fully encoded target clip, and at inference the pipeline encodes the pixel clip from frame 0 through the last anchor's pixel block ($4 \max S + 1$ frames) and clamps only the latent frames in S.

At each training step, the requested panel count is sampled from $N \in \{1, 2, 3, 4\}$ with probabilities $(0.10, 0.25, 0.25, 0.40)$ and capped by the number of available views. A single-panel example always uses Text2Ego with the initial ego frame observed. For $N \geq 2$, the task probabilities are 0.35 for

Table 3: **Notation.** Values in parentheses are those of the reported configuration.

Symbol	Meaning
Views, canvas, and tokens	
V^{ego}, $V_1^{\text{exo}}, \dots, V_K^{\text{exo}}$	egocentric clip and $K \in \{0, \dots, 3\}$ synchronized exocentric clips
$N = 1 + K$	number of panels on the canvas; panel 0 is egocentric
F; $f = 1 + (F-1)/4$	pixel frames per clip (37); latent frames per clip (10)
C; (h, w)	latent channels; latent height and width of one panel (30×30)
$\mathcal{E}$	frozen Wan VAE encoder, applied to each view independently
$\mathbf{Z}_0$, $\mathbf{Z}_t$, $\widehat{\mathbf{Z}}_0$	clean canvas latent in $\mathbb{R}^{C \times f \times h \times Nw}$; noisy canvas; one-step clean estimate (Eq. 12)
$(1, p_h, p_w)$; (h_p, w_p)	patch size $(1, 2, 2)$; token grid of one panel (15×15)
$L = f\, h_p\, N w_p$	number of canvas tokens ($2250N$)
i; $\ell(i)$; $a(i)$	canvas token index; its latent frame; its panel (Eq. 9)
$\text{type}(i)$	ego if $a(i) = 0$, otherwise exo
$\mathbf{z}_i$, $\widetilde{\mathbf{z}}_i$	patch embedding of token i before and after the additive terms of Eq. 6
d; d_h	transformer width (3072); attention head width (128)
Conditioning and objectives	
c	optional text prompt
π_a	conditioning pattern of panel a: `all`, `first`, $\texttt{prefix}_k$, $\texttt{frames}_S$, or `none`
$\mathbf{M} \in \{0,1\}^{f \times h \times Nw}$	conditioning mask; 1 marks observed latent positions
t; $\sigma_t \in [0,1]$; $\boldsymbol{\epsilon}$	diffusion timestep; its noise level; Gaussian noise
$v_\theta(\mathbf{Z}_t, t, c; \mathbf{M}, G)$	predicted velocity; its target is $\boldsymbol{\epsilon} - \mathbf{Z}_0$
$\omega(t)$; $D(\mathbf{M})$	timestep weight; loss normalizer (Appendix B.7)
$\mathcal{L}_{\text{flow}}$, ℓ_{pose}, $\mathcal{L}$	masked flow-matching loss; per-example pose loss; total loss
ϕ; J	frozen top-down pose estimator; number of body joints (17)
$\widehat{\mathbf{p}}$, $\mathbf{p} \in [0,1]^{J \times 2}$	predicted and annotated joints in crop-normalized coordinates
ν_j; b; λ; ε_{num}	joint visibility; pose-loss indicator; pose-loss weight (0.5); numerical floor
Geometry and camera conditioning	
T	number of frames passed to the scene-geometry encoder (2)
G^{sc}, G^{cam}	$17T$ scene tokens; F egocentric camera tokens
$G = [\, G^{\text{sc}} \| G^{\text{cam}} \,]$	geometry context $(\mathbf{g}_1, \dots, \mathbf{g}_{\lvert G \rvert})$, $\lvert G \rvert = 17T + \delta F$, indexed by n; $\text{src}(n) \in \{\text{sc}, \text{cam}\}$
$u_i = \ell(i)/(f-1)$; τ_n	video time of canvas token i; video time of context token n (source pixel frame r_n divided by $F-1$)
$\mathbf{q}_i, \mathbf{k}_n, \mathbf{v}_n$; A_{in}, α_{in}; W_o	query, key, and value of one head; attention logits and weights (Eq. 7); output projection
β_{sc}, β_{cam}	temporal scales $\beta = \exp(a_0 + s_\beta \theta)$, one per source and per block, shared across heads (Appendix B.4)
μ_{in}	visibility term: $-\infty$ for a camera token and an exocentric query, else 0 (Eq. 10)
$r \in \{0, \dots, F-1\}$	pixel-frame index; latent frame ℓ is aligned with $r = 4\ell$
$(R_r, \mathbf{o}_r)$	camera-to-world rotation and optical center of the egocentric camera at frame r
ΔR_r, $\Delta \mathbf{o}_r$; s	pose relative to frame 0 (Eq. 5); per-sequence metric scale
R_{dc}	fixed rotation from RGB-camera to device coordinates (Appendix B.5)
$\boldsymbol{\xi}_r \in \mathbb{R}^{11}$; $\rho(\cdot)$	per-frame camera feature; 6D rotation representation
δ_r; $\delta = \delta_0$	pose validity of frame r; trajectory supplied
p_{drop}	probability of withholding the trajectory in training (0.4)
$\mathbf{K}_{\text{ego}}$; (x_i, y_i); P_{rgb}	pinhole intrinsics of the egocentric view; patch center of token i in pixels; fixed signed axis permutation from image to RGB-camera coordinates
$\mathbf{d}_i$, $\mathbf{m}_i$; $\boldsymbol{\zeta}_i = (\mathbf{m}_i, \mathbf{d}_i)$	unit direction and moment of the viewing ray (Eq. 11); its Plücker coordinates
ψ; $\mathbf{e}_{\text{ego}}, \mathbf{e}_{\text{exo}}$	ray-embedding MLP $\mathbb{R}^6 \to \mathbb{R}^d$; view-type embeddings

Ego2Exo, 0.35 for Exo2Ego, 0.12 for joint generation, 0.10 for extension, and 0.08 for interpolation. Extension samples the prefix length k uniformly from $\{1, \ldots, f-1\}$ and interpolation samples an anchor set S of $|S| \in \{1, 2, 3\}$ latent frames uniformly without replacement from $\{0, \ldots, f-1\}$; both apply the same pattern to every panel. On EgoU-Benchmark, Extension observes the first three latent frames of every panel (the first nine pixel frames) and Interpolation observes the first and last latent frames, $S = \{0, f-1\}$. For Exo2Ego, an ego seed is supplied with probability 0.5, producing the V2V and VI2V variants; Ego2Exo always observes the exocentric seed frames. Ego2Exo is not sampled for groups whose exocentric panels include a moving observer camera (Nymeria), and its probability is renormalized over the remaining families. The action caption is dropped with probability 0.3 during training for every family except Text2Ego; a fixed canvas system prompt is always kept, so the text input is never empty.

B.3 View-Type Embeddings

The embeddings $\mathbf{e}_{\text{ego}}, \mathbf{e}_{\text{exo}} \in \mathbb{R}^{3072}$ are added after patchification and before the first transformer block. The panel index in Eq. 9 determines the embedding: panel 0 receives the ego embedding and all other panels receive the exo embedding. Both vectors are initialized to zero and optimized jointly with the transformer. They are used at training and inference, including on single-panel inputs. The zero initialization preserves the token features of the starting checkpoint.

B.4 Geometry-Guided Cross-Attention

Encoder and projection. The frozen VGGT-Ω aggregator (Wang et al., 2026a) processes sampled conditioning frames resized to 512×512. We retain the final-layer camera token and 16 register tokens per frame. Concatenating the frame and global streams gives a 2048-dimensional feature for each token. Layer normalization and a two-layer MLP with a GELU activation map these features to $d = 3072$. A learned slot embedding identifies each sampled-frame position and token slot. The projection's final linear layer and the slot embeddings are initialized to zero.

The reported runs instantiate the encoder for $T = 2$ frames, giving $17T = 34$ tokens; the slot table is sized to T at construction, and the implementation supports up to eight frames. Frames are selected in order from the available conditioning observations (for $T = 2$, the first and the last observed entry in panel order).

Attention and temporal indexing. Each block places geometry cross-attention after text cross-attention, using RMS-normalized queries and keys and 24 heads of width 128. For source pixel index r_n, the source-time coordinate is $\tau_n = r_n/(F-1)$. The canvas coordinate is $u_i = \ell(i)/(f-1)$. The temporal bias in Eq. 7 is shared across heads. When this branch is used, geometry tokens are supplied separately from the text context. The conditioning unit passes the source time τ_n of every geometry token to the transformer.

The geometry output projection and its bias are initialized to zero. The bias strength is parameterized as $\beta = \exp(a_0 + s_\beta \theta)$ with $a_0 = \log 0.1$, $s_\beta = 100$, and a zero-initialized θ kept in fp32, so that β starts at 0.1, stays positive, and receives a nonzero gradient as soon as the output projection leaves zero, since $\partial\beta/\partial\theta = s_\beta \beta \neq 0$. The zero output projection preserves the initial generator output.

Two sources in one branch. With camera tokens present (Appendix B.5), the context holds $17T + F$ tokens and the visibility term of Eq. 7 is

$$\mu_{in} = \begin{cases} -\infty, & \text{src}(n) = \text{cam and type}(i) = \text{exo}, \\ 0, & \text{otherwise}, \end{cases} \tag{10}$$

with $\text{src}(n) = \text{sc}$ for $n \leq 17T$ and cam otherwise. Every query retains the $17T$ scene tokens, so the softmax is always well defined. The temporal scale is one scalar per source and per block, shared across heads. For the scene tokens, which come from a few sparsely sampled frames, a soft penalty keeps every observation reachable while preferring those close in time; for the camera tokens, one per pixel frame, a large β_{cam} turns the same mechanism into nearly frame-aligned guidance. Because the exocentric views are captured by other cameras, the head trajectory does not govern how exocentric images move, which is why exocentric queries are masked from the camera tokens. Egocentric

queries see both sources within a single softmax, so that a block trades scene layout against camera motion per token instead of summing two independently normalized updates.

Without camera tokens, the bias has shape $L \times 17T$. For $N = 4$ and the implementation maximum $T = 8$, it contains 9000×136 entries, requiring approximately 2.45 MB in bf16 before any additional attention buffers. The branch uses standard scaled dot-product attention. It is disabled for $N = 1$.

B.5 Egocentric Camera Conditioning

Coordinate frames. The dataset stores the trajectory of the Aria device frame, inside which the RGB camera is pitched by approximately $38.6°$; the optical axis is $(0.08, -0.62, 0.78)$ in device coordinates, and the factory calibration agrees to within about $1°$ across takes. Interpreting the stored rotation as the camera's rotation therefore mixes pan, tilt, and roll. With R_{dc} the fixed rotation taking RGB-camera coordinates to device coordinates, the relative rotation of Eq. 5 is moved into the RGB camera frame by conjugation, $\Delta R_r \leftarrow R_{\mathrm{dc}}^{\top} \Delta R_r \, R_{\mathrm{dc}}$, and the relative translation is expressed in RGB-camera axes, $\Delta \mathbf{o}_r \leftarrow R_{\mathrm{dc}}^{\top} \Delta \mathbf{o}_r$, which requires no change to the stored poses. We verified the convention against dense optical flow of the egocentric video: the pan and tilt rates implied by the converted trajectory correlate with the mean horizontal and vertical flow with Pearson coefficients 0.90 and 0.80. The scale s is the per-sequence normalizer of the camera rig (Appendix C.4); $\log(1+s)$ is supplied as a feature so that rooms of different size remain distinguishable after normalization. A pose is marked valid when a trajectory sample lies within 20 ms of the frame, and a clip whose first frame is invalid is treated as having no trajectory.

Camera tokens. Each frame's camera state is $\boldsymbol{\xi}_r = [\, \rho(\Delta R_r),\, \Delta \mathbf{o}_r,\, \log(1+s),\, \delta_r \,] \in \mathbb{R}^{6+3+1+1}$, where $\rho(\cdot) \in \mathbb{R}^6$ is the continuous 6D rotation representation (the first two columns of the rotation matrix) (Zhou et al., 2019) and $\delta_r \in \{0, 1\}$ indicates whether frame r has a valid pose; $\boldsymbol{\xi}_r = \mathbf{0}$ when $\delta_r = 0$. The indicator in Eq. 6 is $\delta = \delta_0$, since all poses are expressed relative to frame 0, and $\delta = 0$ when the trajectory is absent or dropped. The $F = 37$ feature vectors $\boldsymbol{\xi}_r \in \mathbb{R}^{11}$ are mapped to width d by a two-layer MLP with hidden width 512 and a GELU activation, followed by layer normalization. The first linear layer is evaluated in fp32: the resolution of bf16 near one is about 0.008, which is comparable to the inter-frame change of a rotation component during slow head motion. Camera tokens are appended to the scene tokens of Appendix B.4 and use the same RMS-normalized keys; the mask of Eq. 10 is added to the temporal bias, so the combined bias has shape $L \times (17T + F)$. When the trajectory is withheld, the F camera tokens are removed from the context rather than zeroed.

Ray embedding. For an egocentric token i with patch center (x_i, y_i) in pixels and aligned pixel frame $r = 4\ell(i)$, the viewing ray in first-frame camera coordinates has direction and moment

$$\mathbf{d}_i = \frac{\Delta R_r \, P_{\mathrm{rgb}} \, \mathbf{K}_{\mathrm{ego}}^{-1} [x_i,\, y_i,\, 1]^{\top}}{\left\| \mathbf{K}_{\mathrm{ego}}^{-1} [x_i,\, y_i,\, 1]^{\top} \right\|_2}, \qquad \mathbf{m}_i = \Delta \mathbf{o}_r \times \mathbf{d}_i. \tag{11}$$

The final linear layer of the ray MLP ψ is initialized to zero, so the rays do not alter the pretrained generator at initialization. The rays are a deterministic function of the same trajectory that produces G^{cam}, so they are not listed as a separate argument of v_θ in Eq. 4. Since ΔR_r and P_{rgb} are orthogonal, $\mathbf{d}_i$ has unit norm, and $\mathbf{m}_i$ is independent of the point chosen on the ray. Because $F - 1 = 4(f - 1)$, the video time of pixel frame $4\ell(i)$ equals the query time u_i exactly. Rays are cast through the centres of the $h_p \times w_p = 15 \times 15$ token grid of the egocentric panel using the pinhole model of the rectified egocentric video, focal length 300 px on a 480 px image (Appendix C), with the principal point at the image centre, which defines $\mathbf{K}_{\mathrm{ego}}$ in Eq. 11. The calibrated RGB frame has its x axis pointing down and its y axis pointing left in the image, so the fixed signed permutation $P_{\mathrm{rgb}} = \left[\begin{smallmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{smallmatrix}\right]$, a $90°$ rotation about the optical axis, maps back-projected image directions into that convention before ΔR_r is applied. Latent frame ℓ uses pixel frame 4ℓ, the last pixel frame of its temporal group, which matches the temporal anchors used elsewhere in the model. The embedding ψ is a two-layer MLP $6 \to 256 \to d$ with a GELU activation whose first layer is evaluated in fp32 and whose last layer is initialized to zero.

Training. The trajectory is sampled at pixel-frame rate rather than at the compressed latent rate, so head motion within the four pixel frames spanned by one latent frame is preserved. The trajectory

is withheld with probability $p_{\text{drop}} = 0.4$ and is unavailable for EgoExo-Synthetic, which ships no camera poses. Parameters of the camera encoders use a learning rate ten times that of the transformer, because their zero-initialized outputs otherwise remain a negligible fraction of the residual stream for many thousands of steps.

B.6 Pose-Consistency Loss

Eligibility and sampling. Pose supervision is applied only to Ego2Exo examples with $N \geq 2$ and annotations for the first exocentric panel. Eligible examples must satisfy $\sigma_t \leq 0.6$ and are selected with probability 0.25 to limit the additional decoding and pose-estimation cost. Frames with fewer than four visible joints or a degenerate person crop are skipped. The indicator b in Eq. 8 is zero for all other examples.

Clean estimate and decoded frame. The loss is computed on a one-step estimate of the clean canvas in which observed regions are restored,

$$\widehat{\mathbf{Z}}_0 = \mathbf{M} \odot \mathbf{Z}_0 + (1 - \mathbf{M}) \odot \big[\mathbf{Z}_t - \sigma_t v_\theta(\mathbf{Z}_t, t, c; \mathbf{M}, G)\big], \tag{12}$$

which equals $\mathbf{Z}_0$ when the predicted velocity equals its target $\boldsymbol{\epsilon} - \mathbf{Z}_0$. We retain only the first exocentric panel and its first two latent frames, and decode this segment. The last decoded frame is pixel frame 4 under zero-based indexing. Its corresponding latent frame is generated, whereas latent frame 0 is the observed seed. Pose targets and the person crop are taken from annotations for the same pixel frame 4; because the annotations are on a stride-3 grid of the 30 fps source clip, the target for this frame is linearly interpolated between the two bracketing annotated frames, and a joint counts as visible only if it is visible in both. Restricting the decode to one panel and two latent frames reduces the additional computation and memory relative to decoding all views and frames.

Estimator and coordinates. We use a frozen ViTPose-Base estimator with the simple decoder (Xu et al., 2022), trained on COCO, with input resolution 256×192 (height $\times$ width) and $J = 17$ COCO body joints, matching the joint set of the dataset annotations (Appendix C.4). The annotations themselves were produced with a larger estimator offline on the real exocentric video; the training-time estimator only needs to be differentiable and cheap, since it runs inside the training graph. The targets $\mathbf{p}$ are therefore these annotations rather than the predictions of ϕ itself, and both $\widehat{\mathbf{p}}$ and $\mathbf{p}$ are expressed in the coordinates of the annotated person crop, which also provides the framing a top-down estimator expects. The loss is restricted to Ego2Exo because its exocentric panels are always seeded: the seed fixes the wearer's appearance and the target viewpoint without determining the subsequent motion. Because all task families share one set of weights, the body-motion prior learned from this supervision also benefits Exo2Ego, which gains 0.6 dB PSNR from the pose loss in Table 2.

The crop is the bounding box of visible annotated joints, expanded by 25% of its width and height on each side and clipped to the panel. The generated crop is resized to the estimator's input resolution. For heatmap $H_j \in \mathbb{R}^{H_q \times W_q}$, define

$$P_j(y, x) = \frac{\exp H_j(y, x)}{\sum_{y', x'} \exp H_j(y', x')}, \quad \widehat{\mathbf{p}}_j = \sum_{y=0}^{H_q - 1} \sum_{x=0}^{W_q - 1} P_j(y, x) \begin{bmatrix} x/(W_q - 1) \\ y/(H_q - 1) \end{bmatrix}. \tag{13}$$

Here, (y, x) indexes heatmap rows and columns; coordinates are returned in (x, y) order. The softmax is taken over all spatial locations, and the two coordinates are its spatial expectations. Ground-truth joints are transformed into the same crop-normalized coordinate system, and ν_j in Eq. 8 is the annotated visibility flag. We use $\lambda = 0.5$. Gradients propagate through the soft-argmax, frozen estimator, and frozen VAE decoder to the trainable generator. Neither auxiliary network is optimized, and the pose estimator is omitted at inference.

B.7 Optimization and Loss Normalization

All configurations update all transformer weights of Wan 2.2 TI2V-5B with AdamW (learning rate 10^{-5}, weight decay 0.01) on four RTX PRO 6000 GPUs, with warm-up linear from 10^{-6} to the peak rate over the first 500 data-loader steps. Training uses one canvas per GPU per microstep,

gradient accumulation over two microsteps, bf16 activations, activation checkpointing, and sharded fp32 master copies of the weights, giving eight canvases per optimizer update. Timesteps are drawn uniformly over the 1000-point Wan schedule with shift 5, $\sigma = 5s/(1 + 4s)$ for s uniform on $(0, 1]$, so the noise level is not uniform in σ. Step counts in this paper are per-process data-loader steps; one epoch over the training mixture of Appendix C.7 is 65,249 steps (32,625 updates), and every row of Table 2 is trained for 10,000 steps. Inference uses 30 denoising steps, classifier-free guidance scale 5.0 (2.0 for Ego2Exo in Table 1), and a fixed seed, with N set to all available views up to four.

The implementation applies the scheduler's timestep weight $\omega(t)$ to the masked flow error. Its default normalization divides by the number of generated scalar latent elements,

$$D_{\text{gen}}(\mathbf{M}) = \max\left(1,\, C \sum_{\ell,y,x} (1 - M_{\ell yx})\right), \tag{14}$$

where (ℓ, y, x) index the latent canvas. All reported runs use $D(\mathbf{M}) = D_{\text{gen}}(\mathbf{M})$ in Eq. 4, so the loss is an average over generated elements and an example that generates three panels weights each element one third as much as an example that generates one. Equation 4 makes the normalizer explicit because this choice affects task weighting when the number of generated views varies. The weight $\omega(t)$ is the scheduler's bell-shaped function of the timestep, $\exp(-2((t - 500)/1000)^2)$ shifted to zero minimum and rescaled to unit mean over the training grid.

C EgoExoVideo-600K Dataset Details

This appendix expands Sec. 4: how the clips were cut and rectified, how the captions and geometry were produced, the statistics of the release, and the construction of the evaluation split. Fig. 5 shows sample groups from each source.

C.1 Sources and Clip Extraction

Ego-Exo4D. Each take is recorded by Aria glasses worn by the participant and by three to five GoPro cameras on tripods, all time-synchronized by the dataset release (Grauman et al., 2024). We cut one 2.5 s window around every timestamped atomic-action narration whose subject is the camera wearer, requiring consecutive windows from the same take to be at least 1 s apart, and pair the egocentric window with the same window from every exocentric camera of the take. Takes span 13 activity categories, grouped into the eight scenarios of Table 4, recorded at 11 sites. Exocentric cameras that were mounted in portrait orientation (480×854 after processing; 16,576 clips) are excluded so that every exocentric clip has the same 854×480 geometry.

Nymeria. Nymeria records daily activities with Aria glasses worn by the participant and a second pair worn by an observer who follows them (Ma et al., 2024), so the exocentric camera is hand-held in effect and moves through the scene. We cut one 2.5 s window per narrated action segment and pair the two synchronized streams; every group has exactly one exocentric view. Most clips are 75 frames at 30 fps; 3.6% of clips come from 15 fps recordings and have 38 frames.

EgoExo-Synthetic. The rendered sequences of Song et al. (2025b) pair a first-person camera with a virtual observer camera in a game engine. Each sequence is split into consecutive 2.5 s parts at 60 fps (150 frames), and the on-screen interface of the observer view is cropped away. This source contributes 1.3% of the pairs and carries no geometry.

C.2 Rectification and Integrity Checks

Aria egocentric video is a fisheye stream, whereas the pretrained generator and the geometry encoder of Sec. 3 expect pinhole imagery. We re-project it to a pinhole model at 480×480 with a focal length of $1.15\times$ the fisheye focal length, following the manufacturer's rectification convention. Exocentric GoPro streams are undistorted from a Kannala–Brandt model to a pinhole model at 854×480 with focal length $0.9875 \times f_x$; the scale factor was chosen by maximizing sub-pixel feature-match agreement with the released frames over eight camera–take pairs. Wearer boxes and 2D poses are defined in this rectified exocentric frame and overlay the shipped video directly.

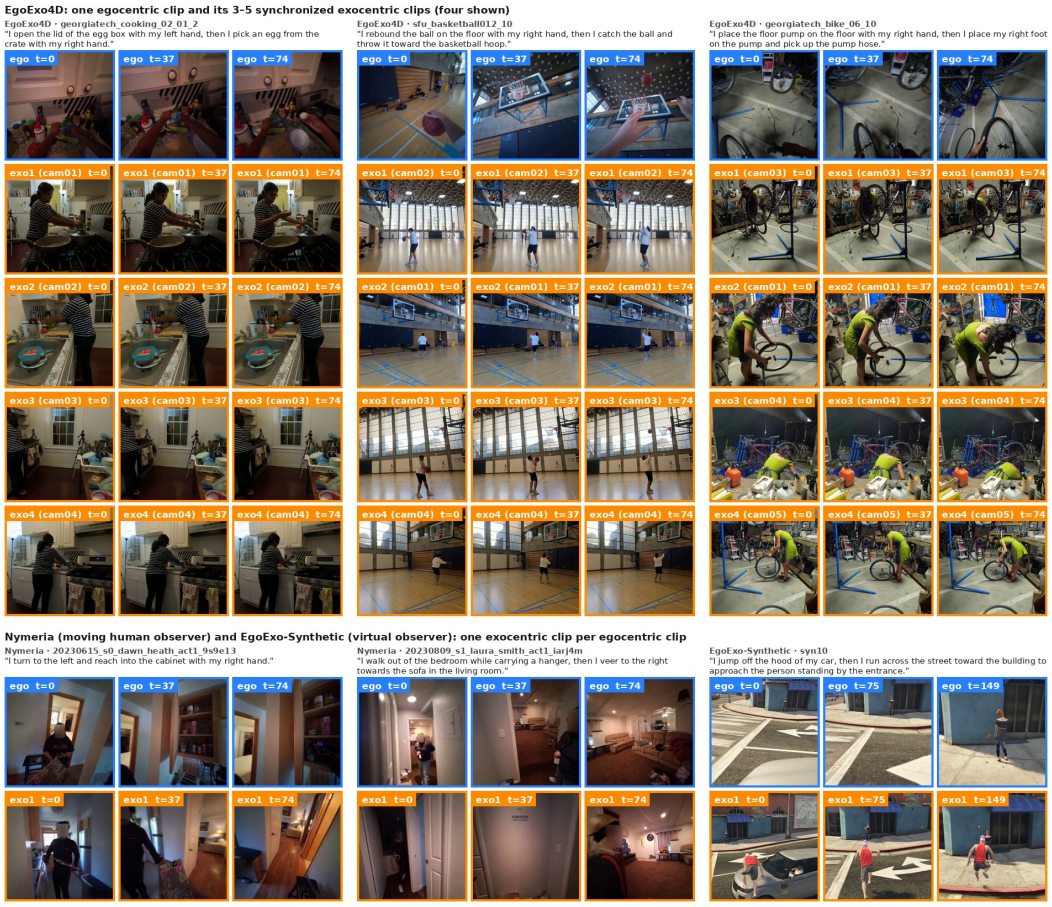


Figure 5: **Samples from EgoExoVideo-600K.** Each block is one group: rows are views, columns are the first, middle, and last frame of the 2.5 s clip, and the caption is the shared first-person text. Top: three Ego-Exo4D groups (cooking, basketball, bike repair) with four synchronized exocentric tripod cameras. Bottom: two Nymeria groups, where the exocentric view is a second person following the wearer, and one EgoExo-Synthetic group with a virtual observer. Exocentric frames are shown center-cropped to a square, as the model sees them.

Every one of the 898,819 video files was decoded end to end rather than trusted by header. This census identified 33 clips whose container advertised 75 frames while the H.264 stream was broken; all were rebuilt from the untransformed originals and re-verified. Egocentric Ego-Exo4D clips decode to 75 or 76 frames against 75 for their partner; the extra frame is trailing, so frames 0–74 correspond one to one. Measured frame counts are stored per clip.

C.3 Captions

Captions are written in the first person and present tense from the wearer's perspective, and each names the current action followed by the next one. For Ego-Exo4D, the source is the timestamped atomic-action narration; Gemini 3.6 Flash rewrites it into this form and appends the following narrated action of the same take, and the original single-action narration is kept alongside. A separate head-motion sentence is produced for every Ego-Exo4D clip by combining the direction of the head turn, derived from the Aria trajectory, with the object of attention identified by Gemini 3.8 Flash from the video ("I turn my head right and look up, looking at bike tire."). Nymeria narrations, which are written in the third person, are rewritten by Gemini 3.6 Flash into the same first-person form, and the original narration is kept. EgoExo-Synthetic captions are generated by Gemini 3.8 Flash from the exocentric frames sampled at 5 fps, which we found more reliable than captioning the rendered

first-person view. Captions average 23.6 words for Ego-Exo4D, 15.9 for Nymeria, and 22.3 for EgoExo-Synthetic; the caption repeats across the exocentric views of one egocentric clip.

C.4 Geometry Annotations

Egocentric device pose. For every Ego-Exo4D and Nymeria clip we provide the 6-DoF pose of the Aria device at each of the 75 frames, resampled from the device trajectory in the take's world frame, together with a per-frame validity flag (a trajectory sample within 20 ms). Poses are comparable within a take, not across takes. This product covers 596,227 of the 604,113 pairs; only EgoExo-Synthetic lacks it.

Camera-wearer box in the exocentric view. Exocentric frames of Ego-Exo4D usually contain several people. We track the person wearing the egocentric camera and store a per-frame box for the clip, with a per-clip validity verdict: a clip is accepted only if the tracklet covers all 15 keyframes used for verification, and rejected rather than extrapolated otherwise. Of 424,973 pairs, 344,001 (80.9%) have a valid box; the rejections are dominated by gating failures and partial coverage. A frame-by-frame manual review of 300 accepted clips found no errors, bounding the error rate at about 1% at the 95% level.

2D body pose of the wearer. For every valid box we run Sapiens 2 (Khirodkar et al., 2026) top-down on the crop at 25 of the 75 frames (stride 3) and keep the 17 COCO body joints with their heatmap peak values, gathered from the estimator's 308-keypoint output through an explicit index map. Using a peak threshold of 0.3 as the visibility test, shoulders are available in 99.6% of measured frames, wrists in 97.6%, knees in 88%, and ankles in about 68%, the last reflecting counters and workbenches that hide the lower body. These annotations supervise the pose-consistency loss of Sec. 3.

Exocentric calibration and derived conditions. Intrinsics and extrinsics of every Ego-Exo4D tripod camera are provided per take, so that the relative pose between any exocentric camera and the egocentric trajectory is available for every group. For training we also pack, per group, the egocentric pose at each of the 37 model frames and the static exocentric poses, expressed relative to the egocentric camera at frame 0 and normalized by a per-sequence scale.

Coverage by source. Wearer boxes and 2D poses are provided for Ego-Exo4D, whose static exocentric cameras share the time anchor of the egocentric stream. Nymeria contributes the egocentric trajectory, and EgoExo-Synthetic contributes video and captions.

C.5 Statistics

Ego-Exo4D egocentric clips have one to five synchronized exocentric views (4.0%, 9.0%, 12.6%, 64.8%, and 9.7% of training groups for one to five views), of which a canvas uses up to three; Nymeria and EgoExo-Synthetic groups have exactly one. Because every exocentric camera of a take is paired with the egocentric clip, the egocentric-to-exocentric direction is one-to-many for Ego-Exo4D (4.16 exocentric clips per egocentric clip on average), and the exocentric seed frame identifies which camera an Ego2Exo target renders. Table 4 gives the activity distribution of the Ego-Exo4D groups.

C.6 Splits and EgoU-Benchmark

Splits are made at the recording level. Ego-Exo4D follows the official take assignment; Nymeria is split by participant session (14 of the sessions are held out, 5% of clips), and EgoExo-Synthetic by source sequence. No take appears in both splits of any source.

EgoU-Benchmark is drawn from the Ego-Exo4D test split, whose 455 takes contain 53,664 videos; no take of this split is used in training, so every benchmark clip is unseen. We keep one group per take. A group is eligible if its egocentric pose is valid at every frame, its wearer box is valid, its 2D pose is available with at least 85% of joints above the visibility threshold and a box diagonal of at least 200 pixels, and both clips decoded to their full length. From the eligible groups we select 420

Table 4: **Ego-Exo4D activities** in EgoExoVideo-600K, as fractions of the training groups, with take-name labels grouped into the Ego-Exo4D scenarios.

Scenario (take labels)	Groups	Share
Cooking (cooking, omelet, salad, sushi)	44,123	48.5%
Bike repair (bike)	11,202	12.3%
Health (covid, covidtest, pcr, cpr)	8,988	9.9%
Climbing (bouldering, rockclimbing)	8,931	9.8%
Basketball	8,828	9.7%
Soccer	6,870	7.5%
Music (music, piano, guitar)	1,593	1.8%
Dance	466	0.5%

groups from 420 distinct takes with a largest-remainder quota proportional to each activity category's share of the test takes and a floor of one group per category, which gives 108 basketball, 75 cooking, 55 bouldering, 45 bike repair, 35 COVID test, 33 dance, 17 soccer, 16 rock climbing, 14 CPR, 10 PCR, 7 music, 4 COVID swab, and 1 omelet group. In 413 of the 420 groups the scored exocentric view is the one marked as most informative for the take. All task families in Sec. 5 are instantiated on these same 420 groups by changing only the per-panel conditioning. A further 60 EgoExo-Synthetic test groups are held out; they carry no geometry and are not part of EgoU-Benchmark.

C.7 Training Mixture

Training samples groups by source with weights 0.70 (Ego-Exo4D), 0.25 (Nymeria), and 0.05 (EgoExo-Synthetic); one epoch holds as many canvas slots as there are training groups (260,996), so it visits every Ego-Exo4D group about twice, a strided 40% of the Nymeria groups, and every synthetic group about 1.7 times. The weights favor Ego-Exo4D, whose multi-camera groups supply canvases with up to four synchronized views for every task family. Each panel receives 37 frames at 15 fps, decimated by exact $2\times$ subsampling from the 75-frame 30 fps clips, resized to 480×480 with a center crop for the exocentric views.

C.8 Release and Licensing

The corpus is released for research use with per-pair captions, the geometry products above, named subsets, and a browser-based viewer. It requires citation of the upstream sources, prohibits attempts to identify recorded people, and keeps provenance identifiers intact; the Ego-Exo4D, Nymeria, and EgoExo-Synthetic portions remain under their own licenses, and the more restrictive term governs where they conflict.

D VLM-as-a-Judge Alignment Protocol

Question Generation. For each benchmark group Gemini 3.8 Flash writes ten candidate four-way multiple-choice questions, two for each of five slots: the action, the active object or its state, the hand–object contact, the state at the end of the clip, and the wearer's head or body motion. It is given the wearer's narration of the clip, the head-motion description, five frames each of the egocentric and of the vetted exocentric clip ($t = 0, 9, 18, 27, 36$), and the narrations of other moments of the same recording, from which the distractors must be drawn, so that every option is an event that does occur in that recording and none is more typical of the activity than the true one. Every question is required to be answerable from either view alone, to name objects by type and location rather than colour, to keep the correct option no longer or more specific than the distractors, and to carry the frame interval in which its evidence is visible; for Extension and Interpolation only questions whose interval overlaps the generated frames count.

Filtering and Scoring. The judge, Qwen3.8-Flash-Next, receives one single-view video at the native 15 fps and high media resolution and one question per call, so that the mutual consistency of the correct options across questions cannot serve as a clue. A question is retained only if the judge answers it correctly on the ground-truth target view and incorrectly when given the question without

any video; the first retained candidate of each slot is kept, at most five questions per group. Align is the judge's accuracy on the retained questions over the 20-group subset. Exocentric target views are judged with a third-person variant of the prompt that identifies the camera wearer. For Text2Ego the generator sees only the seed frame and the description, so its Align measures agreement with the described event. The judge is run once at temperature 0.

Sanity Checks. Across the 25 method rows of Table 1, Align correlates with the MOS Overall of the user study at Spearman $\rho = 0.70$ ($p = 10^{-4}$; Appendix E). Two references bound the measure. The ground truth of another group of the same activity, scored against a group's questions, reaches 0.32 on the retained questions, close to the 0.25 of a four-way guess, so the questions test the event rather than the activity. With distractors drawn from other narrations of the same recording and the no-video filter, the no-video accuracy of the retained set is zero by construction. Reversing the ground truth in time lowers the end-state slot to 0.33 while leaving the contact slot near 0.9. Align is reported on an activity-stratified random subset of 20 benchmark groups (47 egocentric and 34 exocentric questions after filtering); the question sets and the judge outputs are released with the benchmark.

E User Study Design

Subset and trials. The study uses the activity-stratified random subset of 20 benchmark groups of the Align metric, so that the two measures are computed on the same clips. Every method of Table 1 is rated on every group of its family: Exo2Ego without a seed (Exo2Ego-V, WorldWander, EgoX, ours), seeded Exo2Ego (Wan2.2-5B, Wan2.2-A14B, EgoExo-Gen, ours), Ego2Exo (Wan2.2-5B, Wan2.2-A14B, WorldWander, ours), Joint generation and Text2Ego (Wan2.2-5B, Wan2.2-A14B, ours), and, on ten of the groups, Extension (Wan2.2-5B, Wan2.2-A14B, ours) and Interpolation (Wan2.1-FLF2V-14B, Framer, SEINE, ours). A trial is one 37-frame clip at 15 fps that shows, side by side, exactly what the method was given on the left (the exocentric video, a seed frame as a thumbnail, the first frames of both views, or the first and last frames with the frames to be generated darkened) and the generated panel on the right, together with the narration. Ground-truth clips are inserted as attention checks (two per family, 14 in total; 444 trials overall).

Raters and assignment. Raters are members of our laboratory who did not work on the generators. Each rater receives a fixed list in which trials of one family are shown as a block in random order, preceded by two practice trials; the method is never shown. Trials are assigned so that each is rated by up to three different raters, and the interface stores every rating locally and posts it to a shared sheet. A rater would be excluded if the mean Overall rating of their attention checks fell below 3; every rater gave the ground-truth clips a mean rating of at least 4.25, so no rater was excluded. The rating scales are 1–5 for Prompt (the generated video shows the described event: the same action, objects, and order), Temp. (temporal coherence and motion plausibility), Quality (visual quality), and Overall.

Statistics. Table 1 reports per-method means over all ratings of a family (16 raters, 1,084 ratings; 2.4 ratings per trial on average). Agreement on Overall is Krippendorff's $\alpha = 0.41$ (interval) and ICC(2,k) $= 0.71$ on the trials with at least two ratings. Across the 25 method rows of Table 1, MOS Overall correlates with Align at Spearman $\rho = 0.70$ ($p = 10^{-4}$) and with FVD at $\rho = -0.79$, while the pixel metrics are less predictive of the human judgement. Our model receives the highest mean Overall rating in six of the seven families and exceeds the fine-tuned Wan2.2-5B in every family where it is compared ($+1.1$ to $+2.2$ points).

F Additional Qualitative Results

Figs. 6–16 compare our model with every baseline of Table 1 on all task families and on eight activities (cooking, CPR, bouldering, soccer, basketball, health care, dance, and rock climbing). Each block shows one held-out group: the ground truth, our model, and every baseline of that family on the same inputs at frames $t = 0, 6, 12, 18$ of the generated panel(s), i.e., the first 20 frames of the clip. The groups are those of the user-study subset on which our model receives its highest ratings relative to the baselines, so they illustrate the behaviours behind the MOS gaps of Table 1. Without a seed frame (Exo2Ego V2V), EgoX places the scene from a wrong viewpoint and WorldWander keeps a static frame, while our egocentric view follows the hands on the plate and on the manikin. With a seed frame, the fine-tuned Wan2.2-5B either freezes the seed or drifts into unrelated content (soccer),

EgoExo-Gen loses the scene after the first frames, and the fine-tuned A14B model changes viewpoint; in Ego2Exo the target camera is fixed and the seed frame gives the scene, so the comparison centers on the wearer, whose motion our model reproduces most faithfully. For Extension and Interpolation the Wan2.2 baselines and the frame interpolators either hold the anchors or hallucinate a new camera, whereas our model completes the motion in both views; for Text2Ego, given only the seed frame and the caption, both base models leave the climbing wall and the 5B model the court, while ours stays in the scene described.

G LIMITATIONS

Backbone scale. EgoU fills the gap of an all-in-one egocentric simulator by covering seven ego–exo tasks with a single set of weights. Within our computational budget of four GPUs, we build it on the 5B Wan 2.2 TI2V backbone and do not train a canvas model on the larger Wan 2.2 A14B. The fine-tuned A14B image-to-video model is already the strongest baseline in several families of Table 1, which indicates headroom from backbone scale. The canvas formulation, the conditioning masks, and the geometry branch are independent of the backbone, so scaling EgoU to larger video generators is a direct extension.

Data scale. EgoExoVideo-600K provides 604,113 synchronized ego–exo clip pairs with captions and geometry, yet it remains small next to the single-view corpora used to pretrain video generators, because synchronized multi-view capture requires dedicated recording setups. Larger ego–exo collections are a natural way to further improve EgoU, and the same training recipe applies to them without modification.

Embodied and robot data. We train and evaluate on human activities. Robot learning offers the same structure, with wrist-mounted or head-mounted cameras paired with third-person cameras, and the per-panel masks of EgoU accommodate such camera arrangements directly. Applying the canvas to embodied robot data, for example to translate human demonstrations into the robot's viewpoint or to simulate future observations for policy learning, is a promising direction for future work.

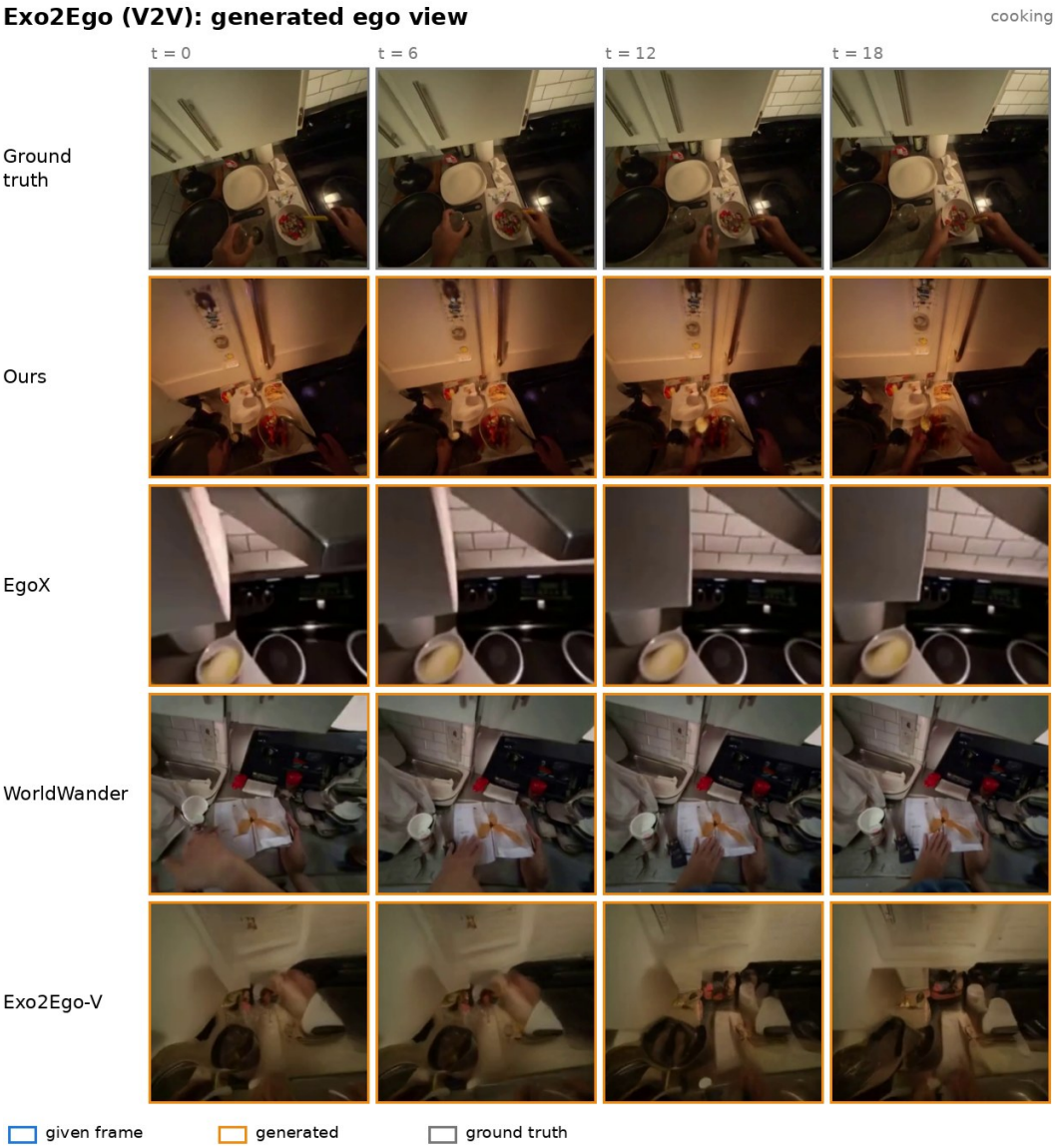


Figure 6: **Exo2Ego without a seed frame, cooking.** The egocentric view is generated from the exocentric video alone. Rows: ground truth, ours, and the baselines of Table 1 on the same inputs; columns: frames $t = 0, 6, 12, 18$ (the first 20 frames) of the generated panel. Blue outlines mark given frames, orange generated frames; the group is drawn from the 20-group user-study subset.

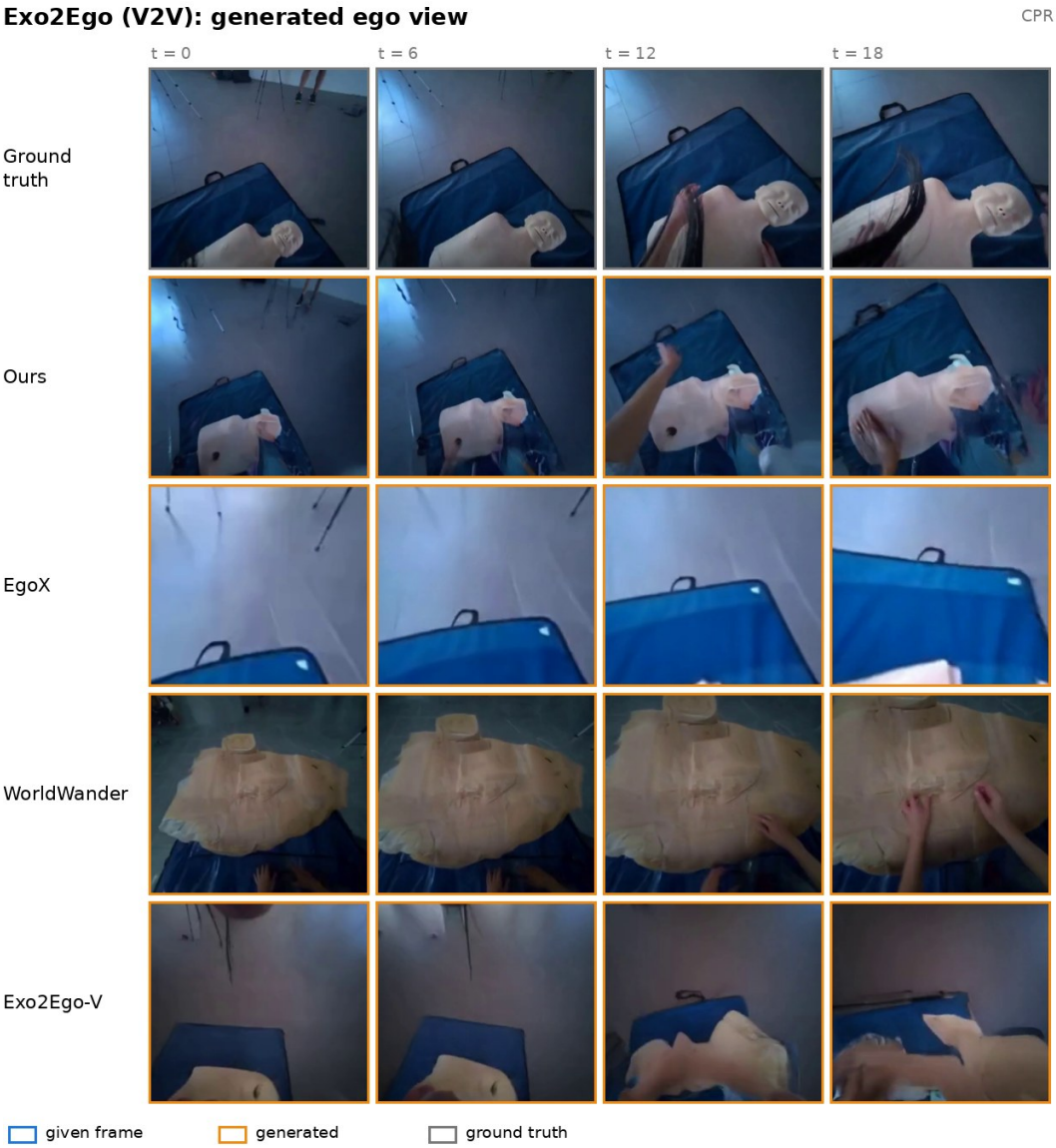


Figure 7: **Exo2Ego without a seed frame, CPR.** Rows: ground truth, ours, and the baselines of Table 1 on the same inputs; columns: frames $t = 0, 6, 12, 18$ (the first 20 frames) of the generated panel. Blue outlines mark given frames, orange generated frames; the group is drawn from the 20-group user-study subset.

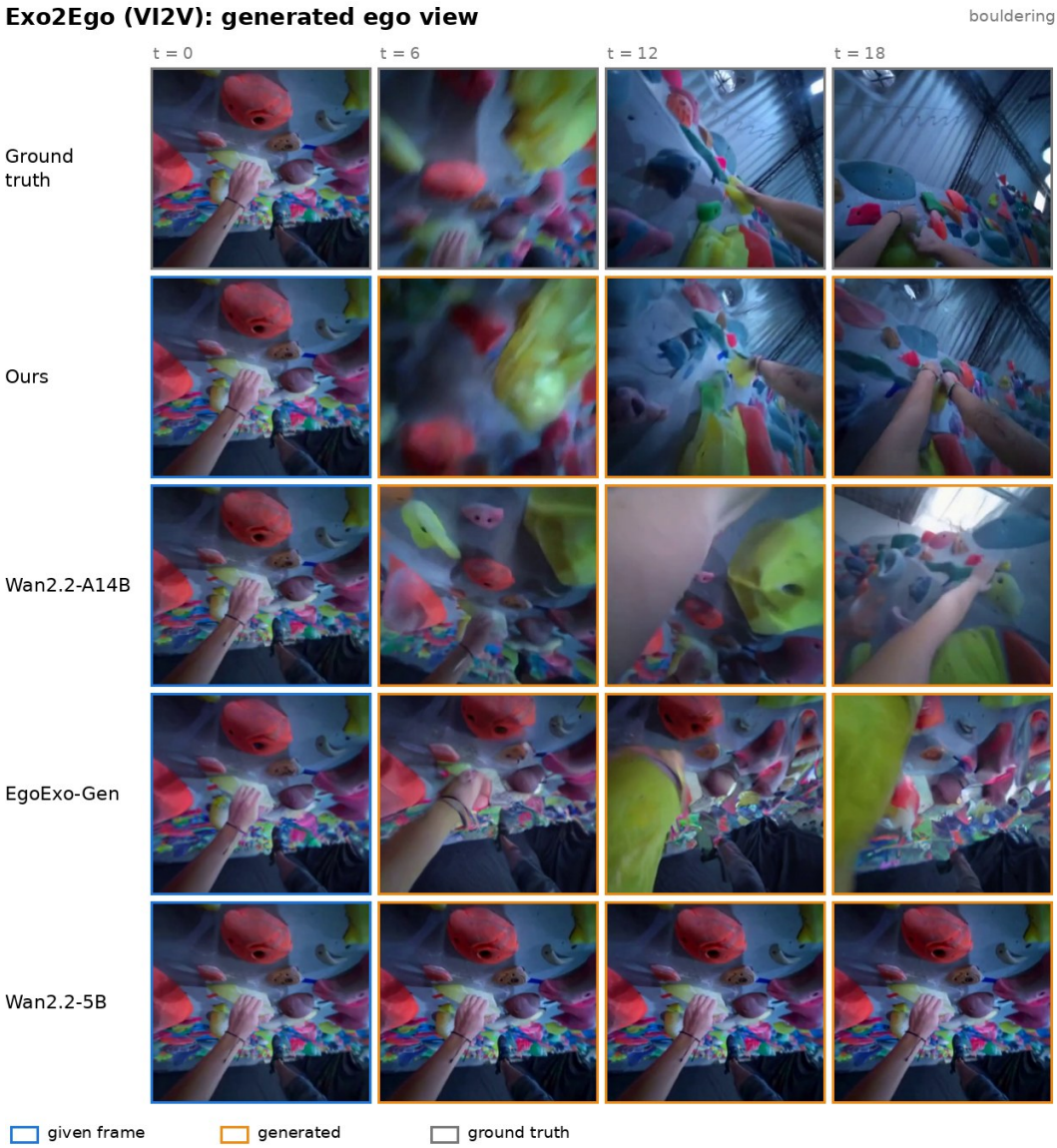


Figure 8: **Exo2Ego with a seed frame, bouldering.** The egocentric first frame is given in addition to the exocentric video. Rows: ground truth, ours, and the baselines of Table 1 on the same inputs; columns: frames $t = 0, 6, 12, 18$ (the first 20 frames) of the generated panel. Blue outlines mark given frames, orange generated frames; the group is drawn from the 20-group user-study subset.

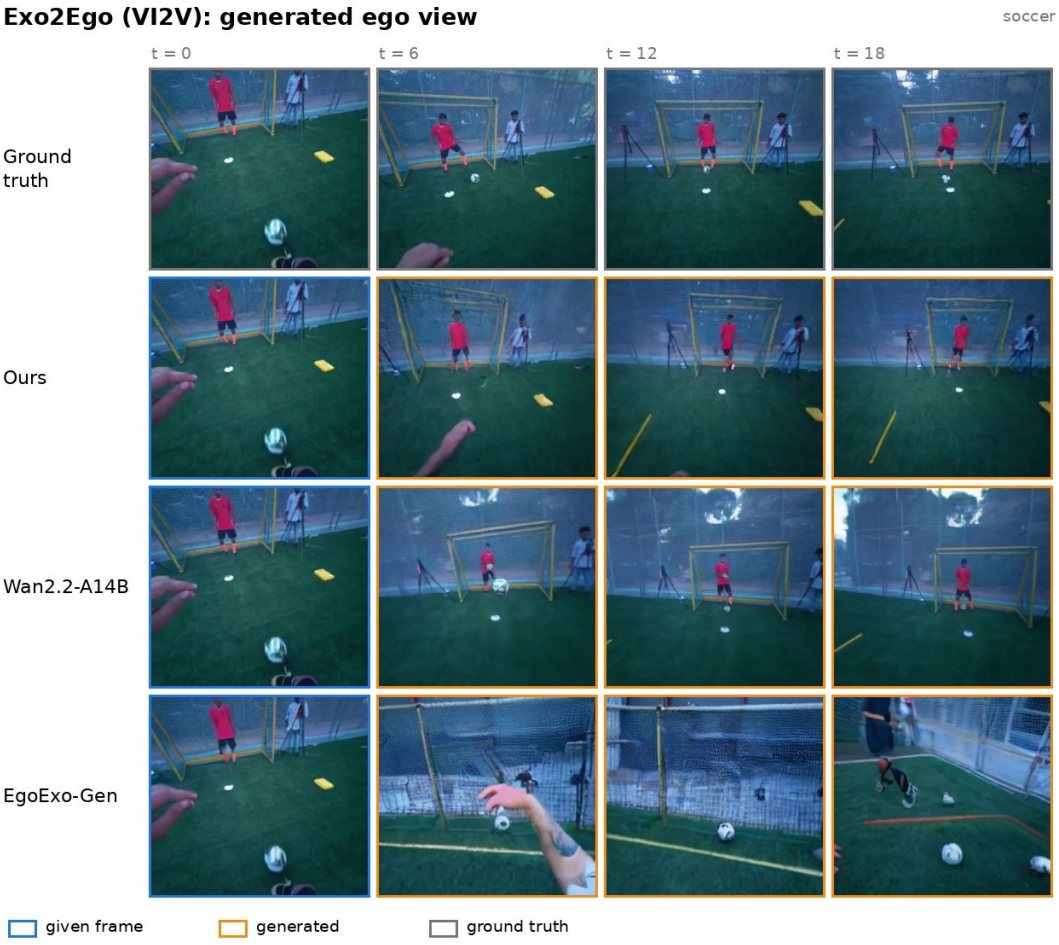


Figure 9: **Exo2Ego with a seed frame, soccer.** Wan2.2-5B is omitted for this group, where its output degenerates to noise. Rows: ground truth, ours, and the baselines of Table 1 on the same inputs; columns: frames $t = 0, 6, 12, 18$ (the first 20 frames) of the generated panel. Blue outlines mark given frames, orange generated frames; the group is drawn from the 20-group user-study subset.

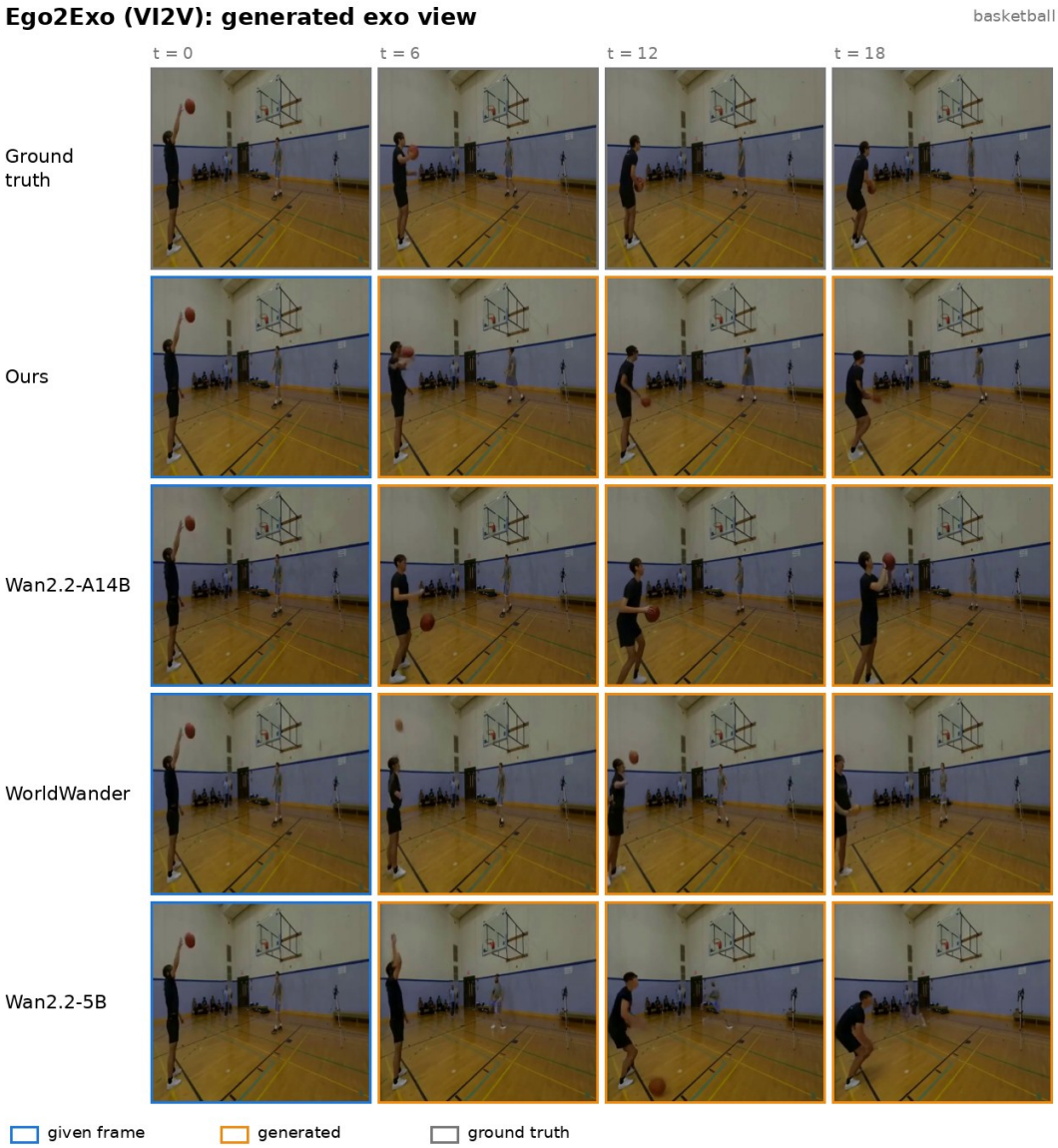


Figure 10: **Ego2Exo, basketball.** The exocentric view is generated from the egocentric video and the exocentric first frame. Rows: ground truth, ours, and the baselines of Table 1 on the same inputs; columns: frames $t = 0, 6, 12, 18$ (the first 20 frames) of the generated panel. Blue outlines mark given frames, orange generated frames; the group is drawn from the 20-group user-study subset.

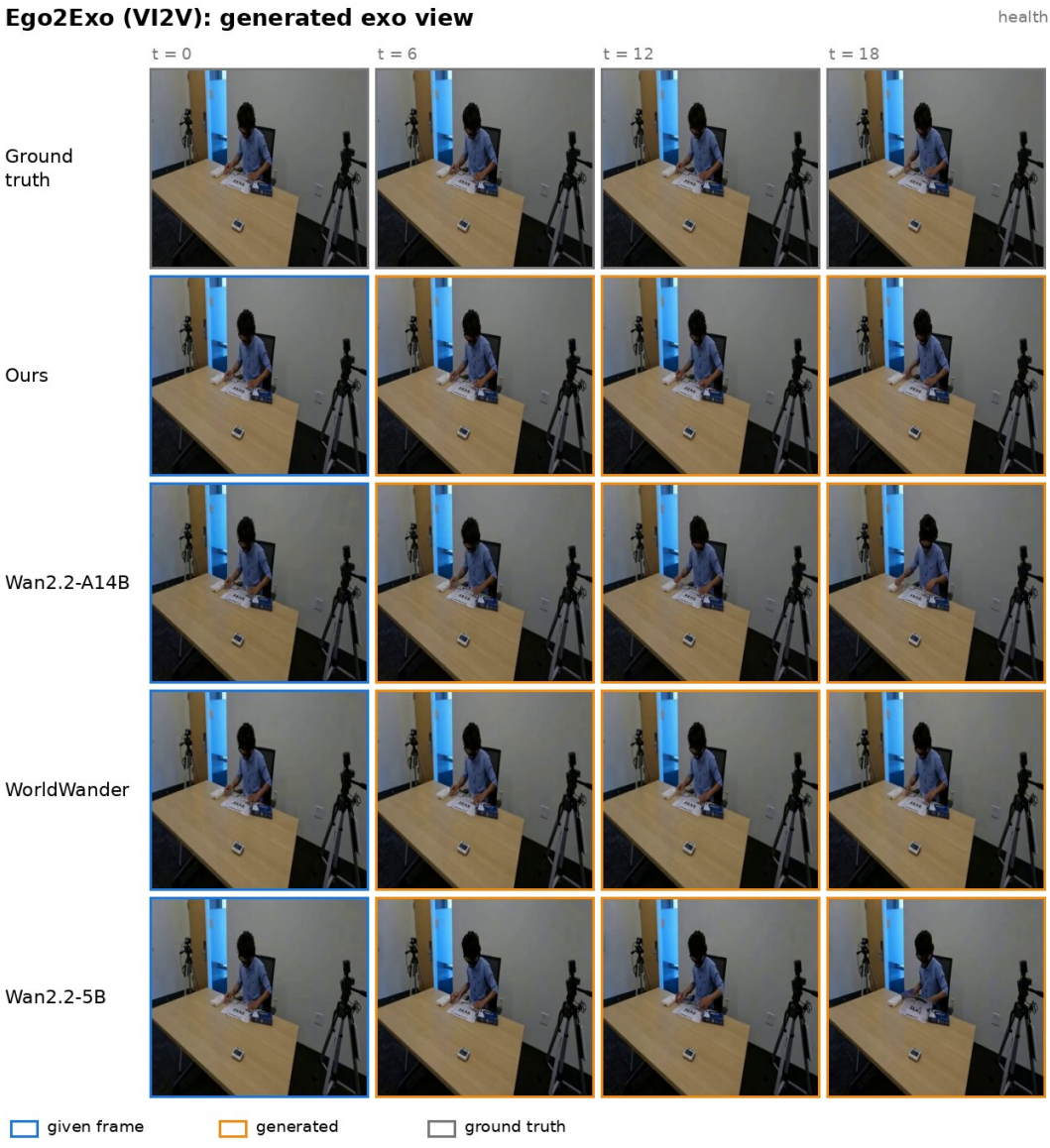


Figure 11: **Ego2Exo, health care.** Rows: ground truth, ours, and the baselines of Table 1 on the same inputs; columns: frames $t = 0, 6, 12, 18$ (the first 20 frames) of the generated panel. Blue outlines mark given frames, orange generated frames; the group is drawn from the 20-group user-study subset.

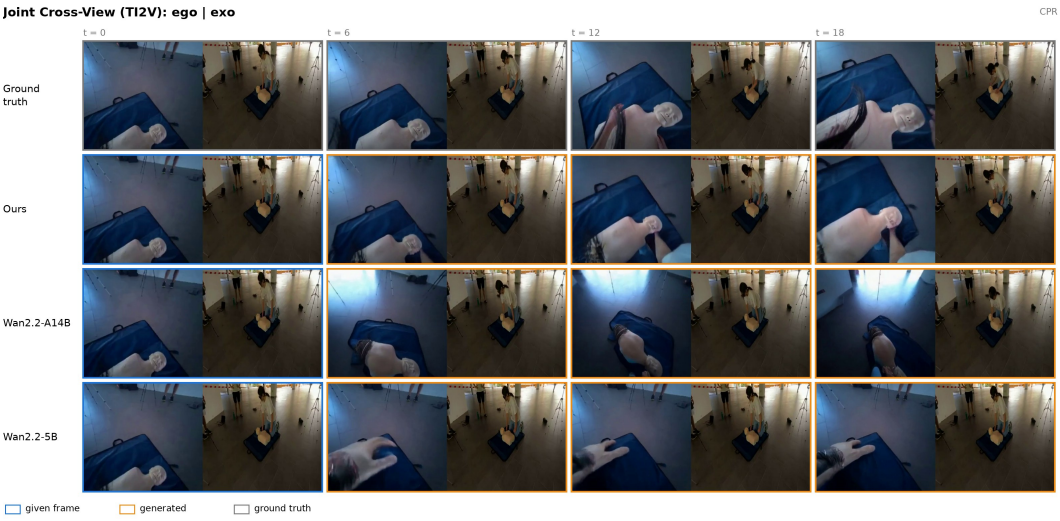


Figure 12: **Joint generation, CPR.** Both views are generated from their first frames and the caption; each tile shows the egocentric (left) and exocentric (right) panel. Rows: ground truth, ours, and the baselines of Table 1 on the same inputs; columns: frames $t = 0, 6, 12, 18$ (the first 20 frames) of the generated panel. Blue outlines mark given frames, orange generated frames; the group is drawn from the 20-group user-study subset.

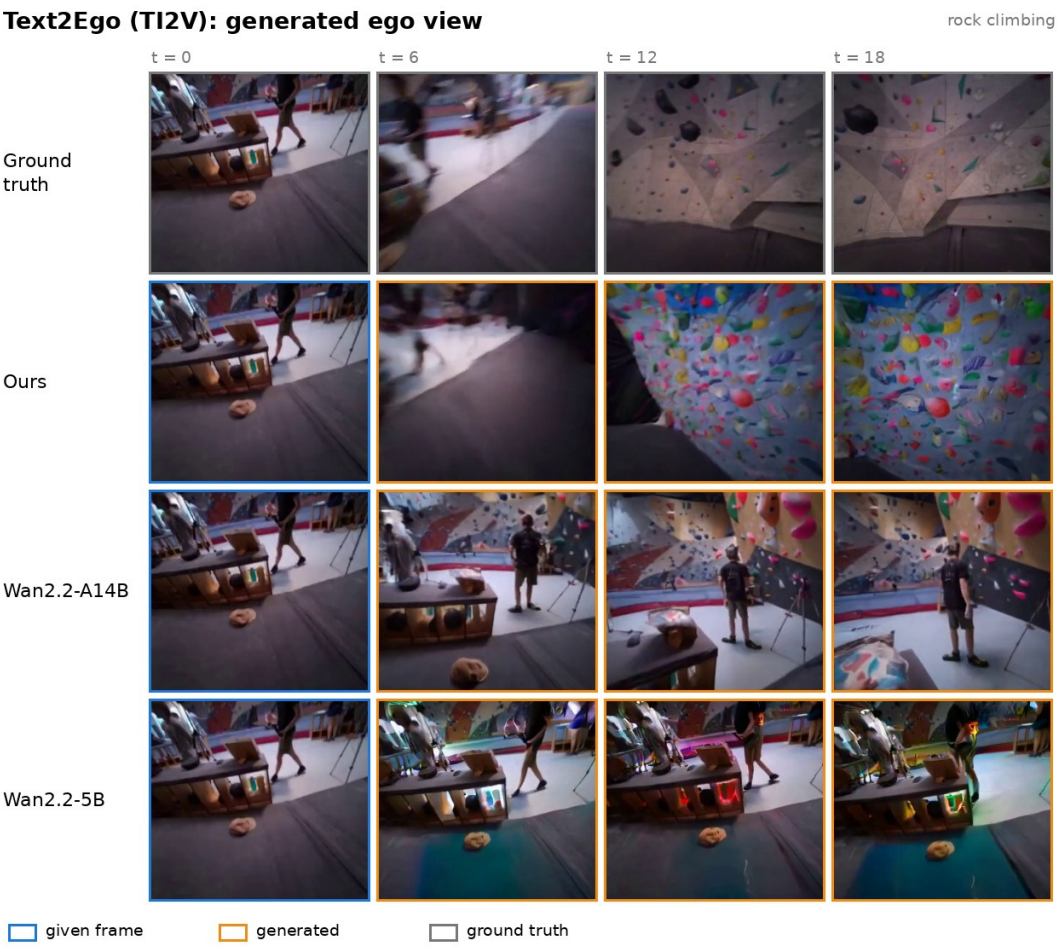


Figure 13: **Text2Ego, rock climbing.** Only the egocentric seed frame and the caption are given. Rows: ground truth, ours, and the baselines of Table 1 on the same inputs; columns: frames $t = 0, 6, 12, 18$ (the first 20 frames) of the generated panel. Blue outlines mark given frames, orange generated frames; the group is drawn from the 20-group user-study subset.

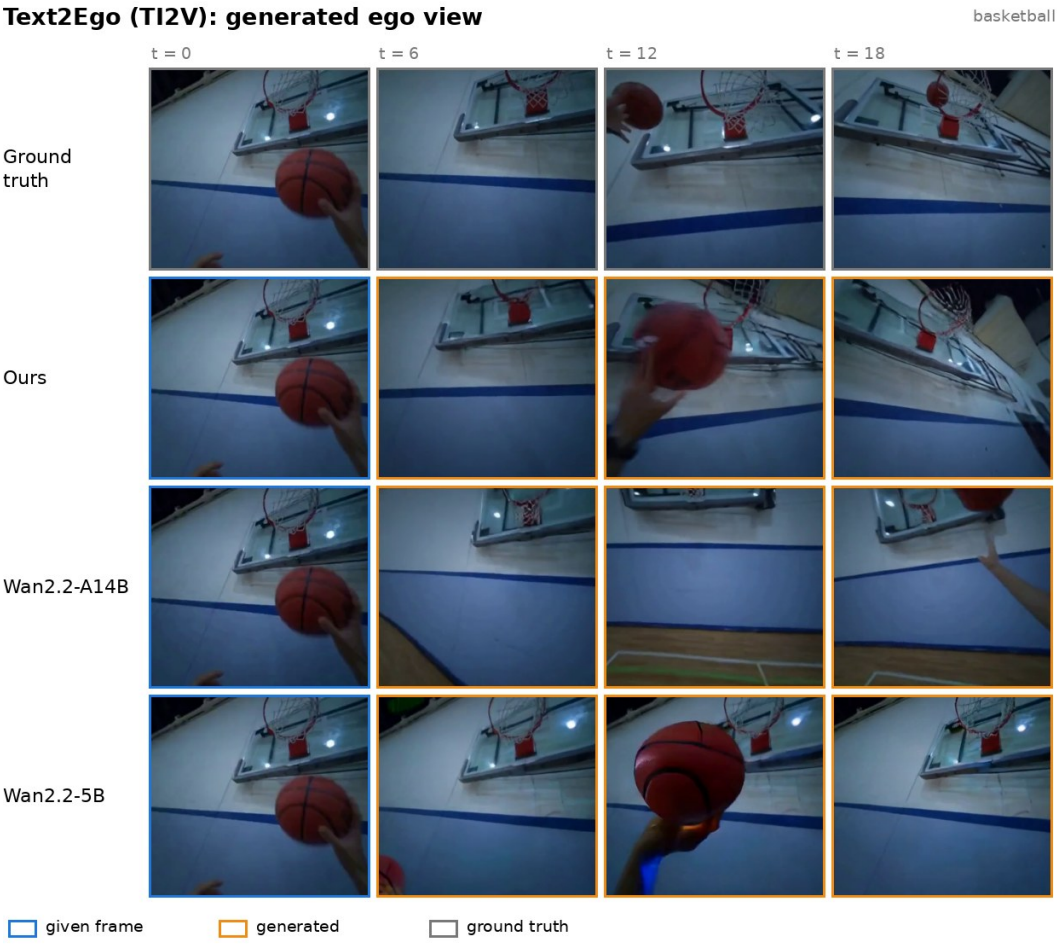


Figure 14: **Text2Ego, basketball.** Rows: ground truth, ours, and the baselines of Table 1 on the same inputs; columns: frames $t = 0, 6, 12, 18$ (the first 20 frames) of the generated panel. Blue outlines mark given frames, orange generated frames; the group is drawn from the 20-group user-study subset.

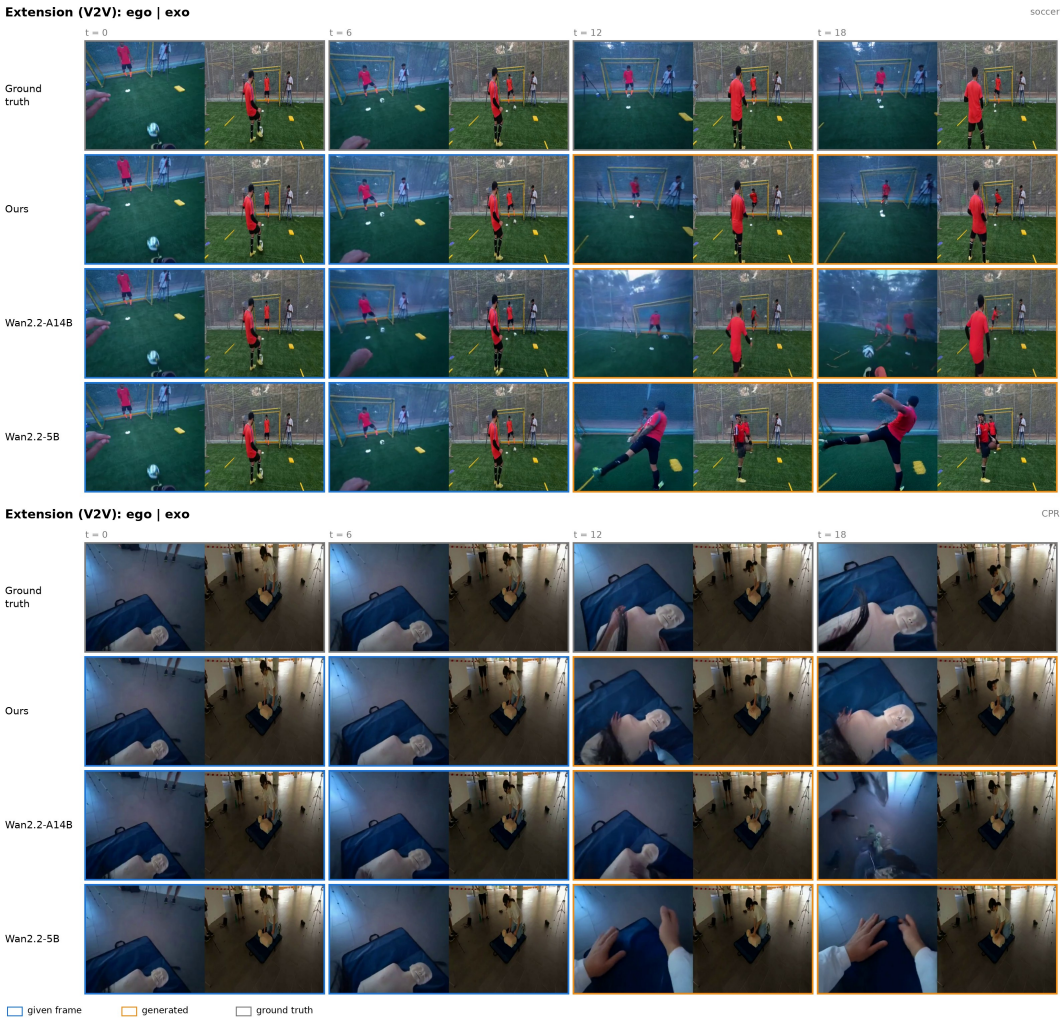


Figure 15: **Extension, soccer and CPR.** Both views are continued from their first nine frames; each tile shows the egocentric (left) and exocentric (right) panel. Rows: ground truth, ours, and the baselines of Table 1 on the same inputs; columns: frames $t = 0, 6, 12, 18$ (the first 20 frames) of the generated panel. Blue outlines mark given frames, orange generated frames; the group is drawn from the 20-group user-study subset.

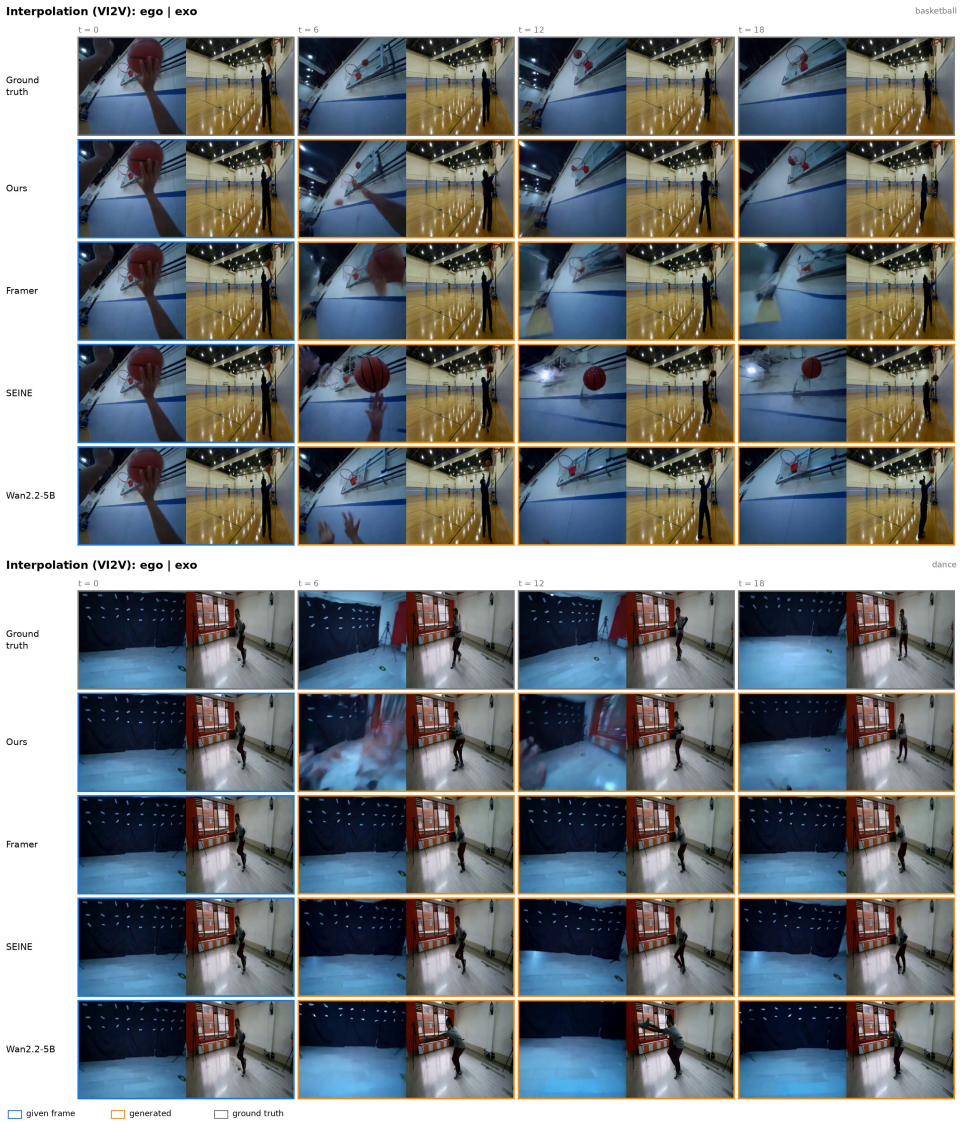


Figure 16: **Interpolation, basketball and dance.** Both views are completed between the first frame and the last four; each tile shows the egocentric (left) and exocentric (right) panel. Rows: ground truth, ours, and the baselines of Table 1 on the same inputs; columns: frames $t = 0, 6, 12, 18$ (the first 20 frames) of the generated panel. Blue outlines mark given frames, orange generated frames; the group is drawn from the 20-group user-study subset.