EGOFORGE: GOAL-DIRECTED EGOCENTRIC WORLD SIMULATOR

Anonymous authors
Paper under double-blind review

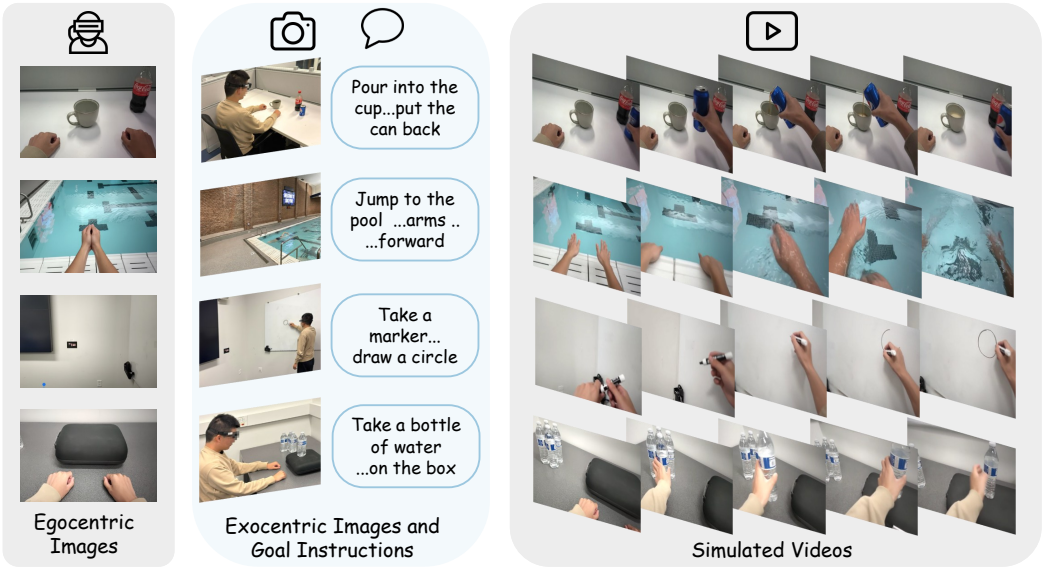


Figure 1: **Real-world smart-glasses deployment of EgoForge.** From a single smart-glasses egocentric frame, a high-level goal instruction, and an auxiliary exocentric image that serves as a scene reference, EgoForge synthesizes egocentric rollouts that follow user intent and preserve scene structure, without relying on dense inference-time conditioning such as camera trajectories, video prefixes, or synchronized multi-view capture.

Qualitative Videos: `https://anonymous.4open.science/r/EgoF-1B40`

ABSTRACT

Generative world models have shown promise for simulating dynamic environments, yet egocentric video remains challenging due to rapid viewpoint changes, frequent hand–object interactions, and goal-directed procedures whose evolution depends on latent human intent. Existing approaches either focus on hand-centric instructional synthesis with limited scene evolution, perform static view translation without modeling action dynamics, or rely on dense supervision, such as camera trajectories, long video prefixes, and synchronized multi-camera capture. In this work, we introduce **EgoForge**, an egocentric goal-directed world simulator that generates coherent, first-person video rollouts from minimal static inputs: a single egocentric image, a high-level instruction, and an optional auxiliary exocentric view. To improve intent alignment and temporal coherence, we introduce GRAFT, a trajectory-level diffusion refinement method that uses positive and negative rollout distributions, derived from goal, temporal, scene-consistency, and perceptual rewards, to steer the diffusion velocity field toward coherent, goal-complete egocentric simulations. Extensive experiments show EgoForge achieves consistent gains in semantic alignment, geometric stability, and motion fidelity over strong baselines, and performs robustly in real-world smart-glasses experiments.

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

Generative world models are redefining how AI systems learn, simulate, and reason about dynamic environments (Matsuo et al., 2022; Hu et al., 2023; Bruce et al., 2024). Recent work has made progress in natural scenes generation, such as autonomous driving (Zhou et al., 2024; Russell et al., 2025), embodied navigation (Bar et al., 2025), and virtual worlds (Ball et al., 2025). However, despite their visual fidelity and predictive power, these models struggle to model human actions in egocentric vision. First-person streams exhibit rapid viewpoint changes, frequent hand–object interactions, and goal-directed behaviors whose evolution depends on latent human intent. Simulating first-person experiences requires understanding affordances and procedural dependencies underlying human actions, which traditional world models overlook.

These limitations have become increasingly important as the demand for immersive and interactive experiences grows across Extended Reality (XR) platforms, including virtual and augmented environments (Tu et al., 2025; Zhang et al., 2025b; Yang et al., 2025). Such applications require human-centric world models capable of generating predictive, controllable, and physically consistent simulations. Egocentric world models could enable next-generation immersive digital experiences. Existing approaches to egocentric generation (Zhang & Shou, 2025; Tu et al., 2025; Zhang et al., 2025b; Agarwal et al., 2025; Zhu et al., 2025) face three fundamental limitations: **(1) Dense supervision requirements.** Existing methods typically require dense motion annotations, calibrated trajectories, video prefixes, or synchronized multi-view recordings, which are costly to obtain at scale and cannot be assumed to be available at inference time in wearable scenarios. **(2) Limited goal-directed control.** Most models are conditioned on short textual prompts or low-level actions (*e.g.*, keyboard or joint controls), producing generic motion patterns and offering limited control over multi-step behaviors. As a result, they struggle to represent semantic intent such as "open the fridge and pour milk." **(3) Weak physical grounding.** Video diffusion models are optimized for visual realism but lack 3D awareness, preventing consistent reasoning about egocentric motion.

To address these challenges, we introduce **EgoForge**, an egocentric world simulator designed for realistic and controllable first-person video generation. Unlike existing approaches that rely on dense motion signals, multi-view capture, or video input, EgoForge generates coherent, goal-directed egocentric rollouts from a single egocentric observation, a high-level instruction, and an optional auxiliary exocentric reference image. Built upon a diffusion-transformer backbone, EgoForge enforces representational alignment (Yu et al., 2025a; Wu et al., 2025) between diffusion latents and 3D geometric features for spatially stable video synthesis. To improve rollout behavior and better align generation with task intent, we further propose GRAFT, a trajectory-level reward-guided refinement stage that uses positive and negative rollout distributions, derived from goal, scene-consistency, temporal-causality, and perceptual rewards, to steer the diffusion velocity field toward coherent, goal-complete egocentric simulations.

To evaluate goal-directed world simulation, we curate **X-Ego**, a benchmark pairing egocentric observations with semantic annotations for grounded video generation. Comprehensive experiments demonstrate that **EgoForge** achieves large gains in semantic alignment while substantially improving video realism and temporal coherence, with 43% lower FVD and 51% lower Flow MSE compared to strong baselines. EgoForge also yields higher structural fidelity (+9.7% SSIM ↑), lower perceptual error (-35% LPIPS ↓), and improved reconstruction quality (+17.8% PSNR ↑).

The contributions of our work are:

- We introduce **EgoForge**, an egocentric world simulator that generates goal-directed first-person video rollouts from minimal inputs, *i.e.*, a single egocentric image, a high-level instruction, and an optional auxiliary exocentric view. To our knowledge, this is the first work to generate goal-directed egocentric rollouts beyond hand-centric instructional motion from minimal static context without pose/trajectory inputs, video prefixes, or synchronized multi-view capture at inference.
- To improve intent alignment and temporal coherence, we propose **GRAFT**, a trajectory-level reward-guided refinement mechanism for video diffusion that fuses goal completion, scene stability, temporal causality, and perceptual fidelity into a unified vector-field update steering sampling toward coherent, goal-consistent rollouts.
- We curate the **X-Ego** benchmark, pairing egocentric observations with detailed event flows, hand–object interactions, object-state changes, and auxiliary visual references, enabling systematic evaluation of goal alignment, temporal coherence, and physical consistency.

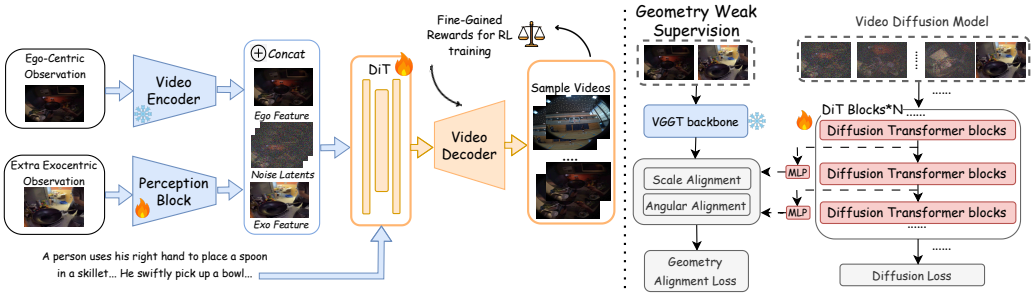


Figure 2: **EgoForge Overview**: Given a single egocentric observation, a high-level instruction expressing user intent, and an auxiliary exo-view reference, EgoForge fuses encoded visual features with noisy video latents at each DiT block to guide generation. Geometry alignment weakly supervises intermediate features using angular and scale consistency to encourage spatially stable rollouts. The resulting rollout videos are further refined via the proposed GRAFT alignment that optimizes goal completion, scene consistency, temporal causality, and perceptual fidelity.

2 Related Work

As summarized in Table 1, prior work relies on continuous exocentric streams for synchronized view generation (Xu et al., 2025; Zhang et al., 2025a; Kang et al., 2025), performs static view translation (Cheng et al., 2024; Park et al., 2025), or generates videos in hand-centric settings with largely static scenes (Li et al., 2024b). Broader egocentric generation typically requires explicit motion signals or synchronized multi-view videos (Zhang et al., 2025a; Kang et al., 2025; Li et al., 2024b; Wang et al., 2024). In contrast, EgoForge predicts controllable, physically consistent egocentric videos from a single first-person image and instruction, without dense motion supervision. See Appendix A for expanded related work.

Table 1: **Representative Related Works.** I/O icons represent Ego-view, Ego Video, Exo Image, Exo Video stream, Text/Action prompt, and Camera parameters. Prior work requires (multi-view) video streams or camera trajectories, while EgoForge generates from minimal static observations. ‡assumes fixed camera poses. ⋆hand motion synthesis.

Model	Inputs	Output
4Diff (Cheng et al., 2024)	+	
EgoWorld (Park et al., 2025)		
Exo2Ego-V (Zhang et al., 2025a)	× 4 sync	
EgoX‡ (Kang et al., 2025)		
Handi⋆ (Li et al., 2024b)	+	
EgoDreamer (Wang et al., 2024)	+ +	
EgoForge (Ours)	+ +	

3 EgoForge Method

EgoForge is an egocentric world simulator designed to generate goal-directed first-person video rollouts that simulate how a scene evolves when a user performs a specified task. Given an initial egocentric frame or short clip $\mathbf{x}_{1:k}$, the objective is to synthesize a plausible future sequence $\mathbf{x}_{k+1:T}$. Formally, EgoForge models

$$p_\theta(\mathbf{x}_{k+1:T} \mid \mathbf{x}_{1:k}) = \prod_{i=k+1}^{T} p_\theta(\mathbf{x}_i \mid \mathbf{x}_{<i}, \mathcal{C}), \tag{1}$$

where $\mathcal{C} = \{\mathbf{x}_{1:k}, y, \mathbf{x}^{exo}\}$ is the conditioning context, which consists of the embeddings derived from $\mathbf{x}_{1:k}$, instruction y, and the exocentric reference $\mathbf{x}^{exo}$. Unlike prior approaches, we do not assume access to camera trajectories, pose signals, or synchronized multi-view streams at inference time. The generated rollout should satisfy three key properties, such as goal alignment (*i.e.*, the sequence reflects the intended instruction), temporal coherence (*i.e.*, motion evolves smoothly across frames), and physical consistency (*i.e.*, scene geometry and interactions remain stable).

3.1 Diffusion-Based Egocentric Generator

Figure 2 illustrates the overall architecture of EgoForge, where video generation is performed in the latent space of a pretrained video autoencoder. Let $\mathbf{z}_0 = \mathrm{Enc}(\mathbf{x}_{k+1:T})$ be the latent trajectory. We

adopt the rectified-flow form of flow matching (Lipman et al., 2022; Liu et al., 2023), consistent with the flow-matching pretraining of the Wan2.2 backbone (Wan et al., 2025). We sample a continuous time $t \sim \mathcal{U}(0,1)$ and a noise tensor $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and form the noisy latent by linear interpolation between data and noise:

$$\mathbf{z}_t = (1-t)\,\mathbf{z}_0 + t\,\epsilon, \qquad \boldsymbol{v} = \frac{\mathrm{d}\mathbf{z}_t}{\mathrm{d}t} = \epsilon - \mathbf{z}_0, \tag{2}$$

where $\boldsymbol{v}$ is the constant velocity of the straight path from $\mathbf{z}_0$ to ϵ, and $t=1$ corresponds to pure noise. A diffusion transformer $\boldsymbol{v}_\theta$ is trained to predict this velocity. Conditioning on $\mathcal{C}$ is implemented through adaptive normalization and cross-attention layers that incorporate the fused embedding. We concatenate the noisy latent with ego and conditioning context features along the channel dimension $\tilde{\mathbf{z}}_t = \mathrm{Concat}(\mathbf{z}_t, \mathbf{f}_{\text{ego}}, \mathbf{f}_{\mathcal{C}})$ and inject the time t through a learned time embedding $\gamma(t)$. The denoising objective is the velocity-matching loss $\mathcal{L}_{\mathrm{D}} = \mathbb{E}_{t,\mathbf{z}_0,\epsilon}\left[\|\boldsymbol{v}_\theta(\tilde{\mathbf{z}}_t, t, \mathcal{C}) - (\epsilon - \mathbf{z}_0)\|_2^2\right]$, and a rollout is generated by integrating the learned ODE $\mathrm{d}\mathbf{z}_t/\mathrm{d}t = \boldsymbol{v}_\theta(\tilde{\mathbf{z}}_t, t, \mathcal{C})$ from $\mathbf{z}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ to $t=0$ and decoding $\mathbf{z}_0$.

Geometry Weak Supervision. To inject 3D reasoning into the diffusion backbone, we align its intermediate representations with geometry features extracted from a pretrained VGGT model (Wang et al., 2025), following REPA (Yu et al., 2025a) and Geometry Forcing (Wu et al., 2025). Let $\mathbf{g}_l \in \mathbb{R}^{N\times Q\times D_g}$ denote the VGGT features at layer l, where L is the number of selected layers, N and Q index temporal and spatial tokens, and D_g denotes the feature dimension. For the diffusion transformer, we extract hidden activations $\mathbf{h}_l \in \mathbb{R}^{N'\times Q'\times D_h}$ from the corresponding layers. Because the two backbones operate at different resolutions, we introduce a learnable projection operator $\Pi_l : \mathbb{R}^{N'\times Q'\times D_h} \to \mathbb{R}^{N\times Q\times D_g}$, implemented via spatiotemporal resampling and channel projection, and obtain projected diffusion features $\mathbf{p}_l = \Pi_l(\mathbf{h}_l)$. To encourage the diffusion features to match the direction of VGGT geometry features, we employ a cosine alignment loss

$$\mathcal{L}^{\text{ang}} = -\frac{1}{LNQ}\sum_{l,n,q}\cos\big(\mathbf{g}_{l,n,q},\,\mathbf{p}_{l,n,q}\big). \tag{3}$$

To constrain the magnitude of geometric features and avoid scale collapse, we introduce a scale alignment loss. We first normalize the projected features as $\tilde{\mathbf{p}}_l = \mathbf{p}_l/(\|\mathbf{p}_l\|_2 + \varepsilon)$ and obtain geometry predictions $\hat{\mathbf{g}}_l = \rho_l(\tilde{\mathbf{p}}_l)$ via a learned linear head ρ_l. The scale alignment loss is then defined as

$$\mathcal{L}^{\text{sca}} = \frac{1}{LNQ}\sum_{l,n,q}\big\|\hat{\mathbf{g}}_{l,n,q} - \mathbf{g}_{l,n,q}\big\|_2^2. \tag{4}$$

The geometry-level coordination objective aggregates both terms across a selected set of layers $\mathcal{L}_{\mathrm{G}} = \zeta_1\mathcal{L}^{\text{ang}} + \zeta_2\mathcal{L}^{\text{sca}}$, where ζ_1 and ζ_2 are coefficients balancing the angular and scale consistency.

3.2 GRAFT: Goal-conditioned Rollout Alignment

Although EgoForge conditions generation on an egocentric observation, an instruction, and an optional exocentric reference image, these signals do not necessarily contribute equally during denoising. One cue may dominate the sampling trajectory, causing the generated rollout to preserve scene appearance while ignoring the goal, follow the instruction while drifting from the environment, or produce locally plausible motion with incorrect temporal order. To mitigate this issue, we introduce GRAFT (Goal-conditioned Rollout Alignment with negative-aware Fine-Tuning), a trajectory-level reward-guided refinement for egocentric video diffusion. GRAFT defines optimality over complete goal-conditioned rollouts rather than individual frames or samples. It scores each rollout with fine-grained rewards for goal completion, scene consistency, temporal causality, and perceptual fidelity. These scores are then converted into per-condition normalized optimality signals that induce positive and negative rollout distributions. The resulting negative-aware update acts directly on the diffusion velocity field. The positive/negative velocity parameterization and the optimality transform of Eq. (5) below follow the forward-process formulation of DiffusionNFT (Zheng et al., 2025); the rollout-level optimality, the multi-reward composition of $\mathcal{R}_{\text{total}}$, and the per-condition reward scale Z_c are specific to GRAFT. Figure 3 illustrates a refinement iteration of GRAFT, where rollout-level rewards guide a negative-aware update of the video generator.

Given the supervised-finetuned policy π^{old}, for each condition $c \in \mathcal{C}$ we draw K candidate rollouts $\mathbf{x}_{1:T}^{(k)} \sim \pi^{\text{old}}(\cdot \mid c)$, $k = 1, \ldots, K$, and assign each rollout a raw reward $R_c^{(k)} = \mathcal{R}_{\text{total}}(\mathbf{x}_{1:T}^{(k)}, c)$.

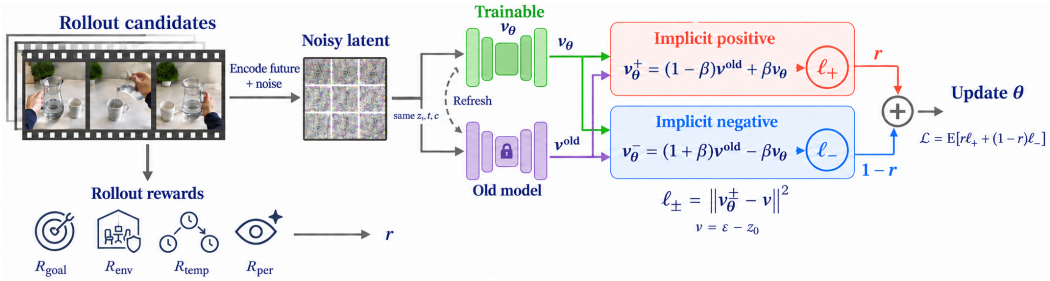


Figure 3: **GRAFT refinement.** Rollout-level rewards yield normalized weights r and $1-r$ for the implicit positive and negative branch losses. Only the trainable model receives gradients, while the old model is refreshed across iterations.

We estimate the per-condition expected reward with the empirical baseline $\mu_c = \frac{1}{K}\sum_{k=1}^{K} R_c^{(k)} \approx \mathbb{E}_{\mathbf{x}\sim\pi^{\text{old}}(\cdot|c)}[\mathcal{R}_{\text{total}}(\mathbf{x}, c)]$, and normalize each centered reward into an optimality probability:

$$\tilde{\mathcal{R}}_{\text{total}}^{(k)} = \frac{1}{2}\left[1 + \text{clip}\left(\frac{\mathcal{R}_{\text{total}}^{(k)}(\mathbf{x}_{1:T}^{(k)}, c) - \mu_c}{Z_c}, -1, 1\right)\right], \tag{5}$$

where $Z_c > 0$ is a per-condition reward scale, the standard deviation of the K rewards of condition c (Appendix E), so that $\tilde{\mathcal{R}}_{\text{total}}^{(k)} \in [0,1]$. For notational convenience, we denote the normalized optimality as $r(\mathbf{x}^{(k)}, c) := \tilde{\mathcal{R}}_{\text{total}}^{(k)}$, so that $r(\mathbf{x}^{(k)}, c) \in [0,1]$. Let $p_{\pi^{\text{old}}}(o = 1 \mid c) := \mathbb{E}_{\mathbf{x}\sim\pi^{\text{old}}(\cdot|c)}[r(\mathbf{x}, c)]$ denote the expected per-condition optimality mass, which normalizes the reweighted positive and negative posteriors:

$$\begin{aligned}
\pi^+(\mathbf{x} \mid c) &= \frac{r(\mathbf{x}, c)}{p_{\pi^{\text{old}}}(o = 1 \mid c) + \varepsilon}\, \pi^{\text{old}}(\mathbf{x} \mid c), \\
\pi^-(\mathbf{x} \mid c) &= \frac{1 - r(\mathbf{x}, c)}{1 - p_{\pi^{\text{old}}}(o = 1 \mid c) + \varepsilon}\, \pi^{\text{old}}(\mathbf{x} \mid c),
\end{aligned} \tag{6}$$

where $\varepsilon > 0$ is a small constant to avoid division by zero. By construction, these posteriors satisfy $\pi^+ \succ \pi^{\text{old}} \succ \pi^-$ in expected reward. We define reinforcement guidance as a vector field acting on the intermediate forward states $\mathbf{z}_t$, steering the model toward π^+ while repelling it from π^-.

Let $\boldsymbol{v}^+$, $\boldsymbol{v}^-$, and $\boldsymbol{v}^{\text{old}}$ denote the velocity fields of the respective policies, and let $\alpha(\mathbf{z}_t, c) = \mathbb{E}[r(\mathbf{x}, c) \mid \mathbf{z}_t, c]$ denote the conditional optimality at the intermediate state $\mathbf{z}_t$. The improvement direction is $\Delta(\mathbf{z}_t, c, t) = \big[1 - \alpha(\mathbf{z}_t, c)\big]\big(\boldsymbol{v}^{\text{old}} - \boldsymbol{v}^-\big) = \alpha(\mathbf{z}_t, c)\big(\boldsymbol{v}^+ - \boldsymbol{v}^{\text{old}}\big)$, yielding the guided target field $\boldsymbol{v}^*(\mathbf{z}_t, c, t) = \boldsymbol{v}^{\text{old}}(\mathbf{z}_t, c, t) + \lambda\,\Delta(\mathbf{z}_t, c, t)$ with guidance strength $\lambda > 0$. Rather than learning Δ as a separate model, we optimize the policy directly via a negative-aware flow-matching loss

$$\mathcal{L}(\theta) = \mathop{\mathbb{E}}_{c,\, \mathbf{x}\sim\pi^{\text{old}}(\cdot|c),\, t,\, \epsilon}\Big[r(\mathbf{x}, c)\, \big\|\boldsymbol{v}_\theta^+ - \boldsymbol{v}\big\|_2^2 + \big(1 - r(\mathbf{x}, c)\big)\, \big\|\boldsymbol{v}_\theta^- - \boldsymbol{v}\big\|_2^2\Big], \tag{7}$$

where $\boldsymbol{v} = \epsilon - \mathbf{z}_0$ is the forward-process velocity target of Eq. (2) evaluated on the generated frames of the collected rollout, $\mathbf{z}_0 = \text{Enc}(\mathbf{x}_{k+1:T})$, with the observed frames entering only through $\mathcal{C}$, and $\boldsymbol{v}_\theta^+ = (1-\beta)\boldsymbol{v}^{\text{old}} + \beta\boldsymbol{v}_\theta$, $\boldsymbol{v}_\theta^- = (1+\beta)\boldsymbol{v}^{\text{old}} - \beta\boldsymbol{v}_\theta$ are implicit positive and negative policies expressed through the single network $\boldsymbol{v}_\theta$ around the frozen $\boldsymbol{v}^{\text{old}}$. Only clean rollouts and their rewards enter the loss, so no likelihood of the sampling chain is required. For a rollout with optimality r, the minimizer of Eq. (7) satisfies

$$\boldsymbol{v}_{\theta^*} = \boldsymbol{v}^{\text{old}} + \frac{2\,r(\mathbf{x}, c) - 1}{\beta}\big(\boldsymbol{v} - \boldsymbol{v}^{\text{old}}\big), \tag{8}$$

so rollouts with $r > \frac{1}{2}$ pull the velocity field toward their own trajectories while rollouts with $r < \frac{1}{2}$ push it away; in expectation over π^{old} the minimizer of Eq. (7) is the guided target field $\boldsymbol{v}^*$ with $\lambda = 2/\beta$, so β sets the guidance strength. Appendix B provide more information about GRAFT.

Rewards. Given a set of rollout candidates $\mathcal{X} = \{\mathbf{x}_{1:T}^{(k)}\}_{k=1}^{K}$, we evaluate each generated video trajectory $\mathbf{x}_{1:T}^{(k)}$ of length T with a set of scalar rewards that measure goal completion, environment preservation, temporal causality, and perceptual fidelity:

- **Goal Completion** ($\mathcal{R}_{\text{goal}}$) evaluates task completion, *i.e.*, whether the trajectory successfully achieves the task outcome, measured by the similarity of the final state to the target reference.

Table 2: **Quantitative comparisons on the X-Ego benchmark.** EgoForge outperforms all baselines across semantic, perceptual, and temporal metrics.

Model	DINO-Score↑	CLIP-Score↑	SSIM ↑	LPIPS ↓	FVD ↓	flow MSE ↓	PSNR ↑
EgoDreamer (Wang et al., 2024)	42.35	25.40	0.58	0.35	580.45	8.15	15.20
Handi (Li et al., 2024b)	31.12	18.25	0.42	0.52	912.30	14.50	12.85
Cosmos (NVIDIA et al., 2025)	49.42	29.77	0.70	0.26	448.12	6.40	18.73
HunyuanVideo (Kong et al., 2024)	53.54	29.43	0.71	0.26	384.31	6.10	18.88
WAN2.2 (Wan et al., 2025)	53.99	35.69	0.72	0.23	322.17	5.78	20.44
EgoForge (Ours)	**61.25**	**39.30**	**0.79**	**0.15**	**182.25**	**2.83**	**24.08**

Table 3: **Quantitative comparisons on X-Ego between EgoForge and other finetuned baseline variants.** +EV: with exo-view img, +TT: text-only domain adaptation; +CI: our conditioning inputs (exo-view and goal instructions) with our structured injection with Geometry Weak Supervision.

Model	DINO-Score↑	CLIP-Score↑	SSIM ↑	LPIPS ↓	FVD ↓	flow MSE ↓	PSNR ↑
Cosmos+EV	48.60	29.60	0.67	0.28	485.75	6.82	18.30
Cosmos+TT	50.80	30.40	0.71	0.25	433.90	6.31	18.88
HunyuanVideo+EV	52.80	29.20	0.70	0.27	405.87	6.30	18.61
HunyuanVideo+TT	54.10	29.86	0.72	0.24	365.80	5.95	19.10
WAN2.2+EV	52.91	35.11	0.71	0.27	352.41	6.25	20.05
WAN2.2+TT	54.80	36.20	0.73	0.25	310.57	5.60	20.64
WAN2.2+CI	58.92	38.05	0.76	0.18	218.72	3.92	22.87
EgoForge (Ours)	**61.25**	**39.30**	**0.79**	**0.15**	**182.25**	**2.83**	**24.08**

- **Scene Consistency** ($\mathcal{R}_{\text{env}}$) measures consistency with the initial scene, penalizing drift, misplaced objects, or transitions into unrelated environments.
- **Temporal Causality** ($\mathcal{R}_{\text{temp}}$) assesses whether the motion evolves in a physically plausible, coherent, and causal manner without temporal artifacts.
- **Perceptual Fidelity** ($\mathcal{R}_{\text{per}}$) captures visual clarity and absence of distortions or artifacts.

Rewards are combined into a single total reward for each rollout, $\mathcal{R}_{\text{total}}(\mathbf{x}_{1:T}^{(k)}, c)$. We use strong vision–language models as non-parametric evaluators to score the generated videos. Detailed reward definitions and complete evaluation prompts are provided in Appendices C and N, respectively.

4 Experiments

X-Ego Benchmark. We introduce X-Ego, the first large-scale benchmark designed to evaluate the ability of egocentric world models to synthesize complex scenes involving multiple, fine-grained conditioning signals. X-Ego is curated from the Nymeria (Ma et al., 2024) and Ego-Exo4D (Grauman et al., 2024) datasets. To better understand human–environment interaction, we further enrich this benchmark with dense annotations of hand–object dynamics, object state changes, and step-level action semantics. The corpus contains 15,000 training samples and a separate, non-overlapping set of 100 held-out test clips covering all interaction categories in our taxonomy. Details about the annotation and dataset statistics are provided in Appendix D; caption and action-segmentation prompts appear in Appendix N.

Implementation Details and Metrics We provide details in Appendix E. We measure low-level visual fidelity using PSNR (Huynh-Thu & Ghanbari, 2008) and SSIM (Wang et al., 2004). To assess perceptual alignment, we employ LPIPS (Zhang et al., 2018), DINO-Score (Zhang et al., 2022) and CLIP-Score (Radford et al., 2021), which captures higher-level semantic correspondence. We further measure distributional realism using FVD (Rakhimov et al., 2020). More details are in Appendix F.

4.1 Experimental Results

Quantitative comparisons. We compare EgoForge with state-of-the-art general-purpose video models, including Cosmos (NVIDIA et al., 2025), HunyuanVideo (Kong et al., 2024), and WAN2.2 (Wan et al., 2025), as well as models designed for egocentric generation such as EgoDreamer (Wang et al., 2024) and Handi (Li et al., 2024b). These models represent the strongest available general-purpose video generators and the most relevant publicly available egocentric baselines. To ensure a fair evaluation, we fine-tune video models on our X-Ego dataset to bridge the domain gap.

Table 4: **Ablation on EgoForge modules.** We evaluate the impact of diffusion finetuning (FT), geometry weak supervision (GWS), and the proposed GRAFT refinement. Each component consistently improves performance across semantic, perceptual, and temporal metrics, with the full EgoForge model achieving the best overall results.

FT	GWS	GRAFT	DINO-Score↑	CLIP-Score↑	SSIM↑	LPIPS↓	FVD↓	flow MSE↓	PSNR↑
✓	✗	✗	56.81	37.10	0.74	0.21	260.89	4.82	21.92
✓	✓	✗	58.92	38.05	0.76	0.18	218.72	3.92	22.87
✓	✓	✓	**61.25**	**39.30**	**0.79**	**0.15**	**182.25**	**2.83**	**24.08**

As shown in Table 2, **EgoForge** consistently outperforms all baselines across all metrics. Compared with the strongest baseline, it improves semantic alignment by +13.4% DINO-Score and +10.1% CLIP-Score, while increasing structural fidelity (+9.7% SSIM) and reconstruction quality (+17.8% PSNR). At the same time, perceptual error is substantially reduced (35% lower LPIPS). Most notably, EgoForge achieves large gains in temporal modeling, reducing FVD by 43% and flow MSE by 51%, indicating significantly more coherent motion and stable scene dynamics. These improvements demonstrate that EgoForge more effectively captures egocentric motion patterns than baselines.

To ensure a fair and rigorous comparison, we further enhance the baseline models using three progressive strategies under matched training settings. Native exo-image conditioning (+EV) supplies the auxiliary exocentric image through each backbone's existing visual-conditioning interface, then fine-tunes the model on X-Ego with its standard denoising objective. Goal-text domain fine-tuning (+TT) fine-tunes the backbone on the same X-Ego videos through its native text-conditioning pathway, using the high-level goal instruction and no exocentric image. Structured conditioning injection with geometry weak supervision (+CI) replaces both pathways with our conditioning mechanism and geometry supervision. As shown in Table 3, both baseline strategies help, and +EV performs worse than +TT on all three backbones, which shows that supplying the exocentric image through a native interface is not by itself sufficient. EgoForge achieves the best results across all seven metrics, reaching an FVD of 182.25 and a flow MSE of 2.83.

One row in Table 3 requires clarification. Wan2.2+CI is exactly the Stage-1 model of EgoForge before reward-guided refinement, so we report it as our own ablation rather than as an independent baseline. We therefore separate two quantities throughout the paper. The full-system gain compares EgoForge with the off-the-shelf Wan2.2 backbone and reaches $+13.4\%$ DINO-Score and 43.4% lower FVD. The method-only gain compares the GRAFT stage with our equally conditioned Stage-1 model and reaches $+3.95\%$ DINO-Score and 16.7% lower FVD. Appendix I gives the per-metric gain breakdown (Table 9) and independent matched-conditioning baselines on Cosmos and HunyuanVideo (Table 17), where GRAFT lowers FVD by about 16% on every backbone.

Human evaluation. We conduct a human study to assess quality and goal alignment; the protocol and results, including the matched Stage-1 comparison are reported in Appendix G, Tables 6 and 7.

4.2 Ablations

Effectiveness of GRAFT. Table 4 isolates reward-guided refinement by comparing FT+GWS with the full FT+GWS+GRAFT model, keeping conditioning and geometry supervision fixed. Adding GRAFT increases DINO-Score from 58.92 to 61.25 and CLIP-Score from 38.05 to 39.30, while reducing FVD from 218.72 to 182.25 and flow MSE from 3.92 to 2.83. These gains support the effectiveness of trajectory-level refinement for semantic alignment and temporal coherence. A naive per-frame adaptation of an image-level reward objective is worse than no refinement at all (FVD 231.60 vs. 218.72, Appendix H.2, Table 10).

Effect of Geometric Alignment Loss. We compare FT with FT+GWS in Table 4, with GRAFT disabled in both variants. Geometry alignment increases DINO-Score by 2.11 (56.81 to 58.92) and CLIP-Score by 0.95 (37.10 to 38.05), while reducing LPIPS from 0.21 to 0.18 and flow MSE from 4.82 to 3.92. Appendix H.1, Table 8, shows complementary gains from geometry supervision and GRAFT, while extra denoising fine-tuning remains below the full model. Appendix H.3, Table 11, shows that geometric alignment outperforms semantic and random feature alignment.

Conditioning inputs. Table 5 studies whether EgoForge uses each conditioning signal: the egocentric observation, the goal instruction, and the auxiliary exocentric image. Each missing modality is

Table 5: **Ablation of conditioning inputs.** Missing modalities are replaced with learned null embeddings. For counterfactual conditions, *random Exo* uses an exocentric view sampled from another test example, and *wrong Text* uses an action-incompatible instruction. Metrics are computed against the original held-out video and instruction.

Conditioning	Inputs			Generation Quality						
	Ego	Text	Exo	DINO ↑	CLIP ↑	SSIM ↑	LPIPS ↓	FVD ↓	Flow MSE ↓	PSNR ↑
Single-modality conditioning										
Text only	✗	✓	✗	49.84	36.75	0.62	0.33	398.52	7.01	17.18
Ego only	✓	✗	✗	54.26	28.41	0.70	0.25	310.44	5.73	20.14
Exo only	✗	✗	✓	50.91	27.86	0.65	0.30	365.73	6.38	18.56
Two-modality conditioning										
Ego + Text	✓	✓	✗	58.73	38.72	0.77	0.17	204.62	3.19	23.30
Ego + Exo	✓	✗	✓	56.88	30.95	0.76	0.18	231.48	3.95	22.56
Text + Exo	✗	✓	✓	55.64	37.91	0.69	0.25	282.36	4.91	19.69
Counterfactual conditioning										
Ego + Text + random Exo	✓	✓	rand.	57.20	37.89	0.74	0.21	237.44	4.13	21.85
Ego + wrong Text + Exo	✓	wrong	✓	55.31	31.68	0.75	0.20	254.18	4.55	21.84
Ego + Text + Exo	✓	✓	✓	**61.25**	**39.30**	**0.79**	**0.15**	**182.25**	**2.83**	**24.08**

replaced with a learned null embedding. Ego-only generation preserves more low-level scene structure but lacks goal alignment. Text-only generation achieves better text alignment but suffers from poor visual and temporal fidelity because it lacks the first-person initialization. Ego+Text already produces plausible rollouts, showing that the exocentric view is optional at inference time. Adding the correct exocentric reference further improves all metrics, reducing FVD from 204.62 to 182.25 and flow MSE from 3.19 to 2.83. We also evaluate two counterfactual settings. Replacing the paired exocentric image with a random exo-view hurts all metrics, indicating that EgoForge does not ignore the exocentric input. Replacing the instruction with an incorrect goal causes a large drop in CLIP-Score and temporal fidelity, showing that the rollout is sensitive to high-level intent rather than being determined only by visual context. The exocentric image used in this table is taken from the end of the clip. Appendix H.6, Table 14, shows that a clip-start reference provides most of the benefit; Appendix H.7, Table 15, repeats the component ablation under that setting. Appendix H.4, Table 12, shows that refinement increases sensitivity to the goal instruction. Appendix H.5, Table 13, shows that refinement gains persist without exocentric input.

Reward Ablation. Ablating $\mathcal{R}_{\text{goal}}$, $\mathcal{R}_{\text{env}}$, $\mathcal{R}_{\text{temp}}$, or $\mathcal{R}_{\text{per}}$ degrades performance, with the full reward composition performing best across all reported metrics (Appendix H.8, Table 16). Appendix J evaluates cross-judge agreement and reliability (Tables 18 and 19), robustness to judge replacement, agreement with human ratings (Table 20), and reward-hacking probes.

4.3 Qualitative Analysis

Figure 4 provides qualitative comparisons of our model, EgoForge, against Cosmos (Agarwal et al., 2025), Hunyuan (Kong et al., 2024), and Wan2.2 (Wan et al., 2025) for two complex, long-horizon egocentric tasks. Results demonstrate that EgoForge generates more plausible and coherent video sequences than the baseline models. Additional examples are provided in Appendix K.

Figure 5 provides a qualitative comparison of EgoForge's generation with and without an auxiliary exocentric image. Results show that EgoForge generates plausible egocentric rollouts using instruction guidance alone (Rows 1 & 3), while incorporating an auxiliary exocentric image further improves spatial grounding and scene consistency by anchoring the simulation to the reference environment (Rows 2 & 4), such as the *potted plants on the windowsill* and the distinctive *red and green rubberized surface* of the court. This confirms that an exo-view frame can serve as effective guidance for EgoForge. Dense rollout visualizations across diverse activities are provided in Appendix K, Figure 6, alongside additional baseline comparisons and long-duration sequences (Figures 7 and 8).

Real-world Smart-Glasses Experiments Previous egocentric world models have been tested only with in-domain data, ignoring the complexity of real-world out-of-domain (OOD) data. To fill this gap, we deploy EgoForge in a real-world environment using ARGO smartglasses. The selected tasks are shown in Figure 1. For each task, a single egocentric frame is captured directly from

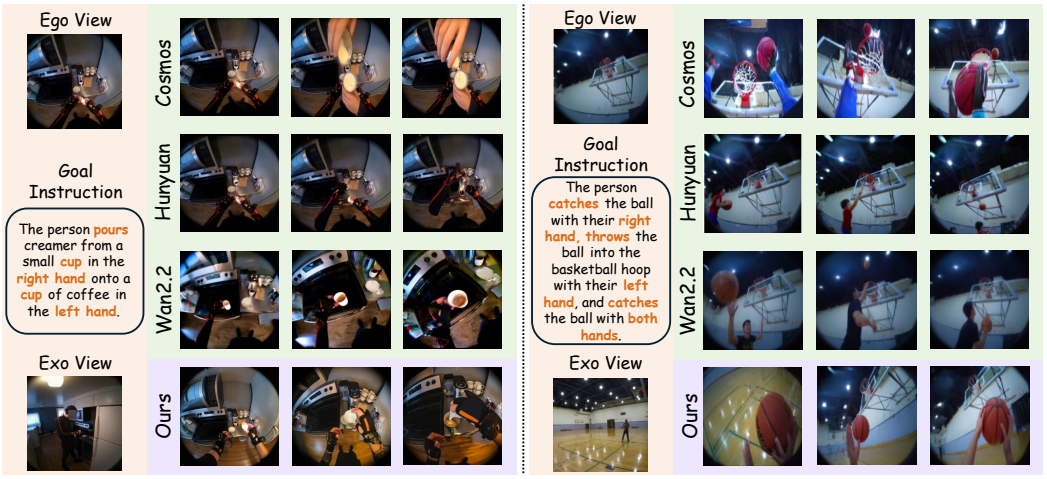


Figure 4: **Qualitative Comparison between EgoForge and baselines**. EgoForge accurately reconstructs multi-step, causally ordered actions, preserving hand–object geometry, temporal consistency, and goal alignment. For instance, in the first example, Cosmos erroneously generates a third hand, Hunyuan depicts a disconnected arm, and Wan2.2 fails to complete the coffee-pouring task. In the second example, Cosmos generates multiple balls, while both Hunyuan and Wan2.2 generate an incorrect person to perform the action. In contrast, EgoForge accurately completes both tasks.



Figure 5: **Qualitative comparison of EgoForge with and without exocentric input.** EgoForge can be steered using an auxiliary exo-view to improve spatial grounding by anchoring the simulation to the reference environment.

the smartglasses, while an auxiliary exo-view reference image from another participant is provided alongside the text-action guidance. Details are provided in Appendix L, Figure 9.

5 Conclusion

We introduce EgoForge, a goal-directed egocentric world simulator that generates coherent first-person video rollouts from minimal context: an egocentric observation, a high-level instruction, and an optional exocentric reference. EgoForge combines geometry-aware grounding for spatial consistency with GRAFT, a trajectory-level reward-guided refinement method that balances goal completion, scene stability, temporal causality, and perceptual fidelity through a negative-aware update of the diffusion velocity field. Experiments, including smart-glasses scenarios, show that EgoForge outperforms strong baselines, producing rollouts with stronger instruction alignment, more realistic egocentric motion, and more stable scene evolution. We hope this direction will facilitate future research on immersive world models, interactive simulation, and human-centered XR systems.

AI USE STATEMENT

We used generative AI tools to help with writing. Their use was limited to language polishing, such as correcting grammar, improving readability, and shortening sentences to fit the page limit. We did not use generative AI tools for research ideation, method design, experiment design, or the interpretation of experimental results. The model design, the GRAFT formulation, the benchmark construction, and all experimental settings were developed by the authors. All reported numbers come from experiments that the authors implemented and ran.

REPRODUCIBILITY STATEMENT

Section 3 describes the model architecture, geometry weak supervision, and GRAFT refinement objective. Appendix D details the construction and annotation of X-Ego, including the separation of training and test data. Appendix E reports the backbone, optimization settings, LoRA configuration, training hardware, and two-stage training procedure, including rollout collection and reward normalization. Reward definitions and evaluation and annotation prompts are provided in Appendices C and O, respectively. Appendix F describes the automatic evaluation metrics, and Appendix G provides the human-evaluation protocol and participant instructions. We separately evaluate generation without an exocentric image and with clip-start or clip-end references in Appendices H.5 and H.6, and report the component ablation with clip-start references in Appendix H.7. We intent to release all code, model checkpoints, and the X-Ego benchmark upon publication.

ETHICS STATEMENT

EgoForge is developed for research on egocentric video generation and goal-directed simulation. X-Ego is constructed from Nymeria and Ego-Exo4D. Further use of these data should respect the source datasets' access conditions and participants' privacy. Our human evaluation involved voluntary, unpaid participation. Raters were informed of the evaluation purpose and rating criteria, and their anonymous responses were analyzed only in aggregate. No sensitive personal information was collected from raters. Realistic first-person video generation presents potential risks, including fabricated recordings, depictions of private spaces, and misuse in surveillance. Generated rollouts should be clearly identified as synthetic rather than presented as recordings of actual events. Wearable inputs may also contain bystanders or sensitive information, requiring appropriate consent and privacy safeguards. We intend EgoForge for simulation and evaluation, not for surveillance or decisions about individuals. Deployment should include disclosure of synthetic content and human oversight, particularly in safety-critical settings.

REFERENCES

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos J Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in atari. *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 58757–58791, 2024.

Kumar Ashutosh, Rohit Girdhar, Lorenzo Torresani, and Kristen Grauman. Hiervl: Learning hierarchical video-language embeddings. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23066–23078, 2023.

Philip J. Ball, Jakob Bauer, Frank Belletti, Bethanie Brownfield, Ariel Ephrat, Shlomi Fruchter, Agrim Gupta, et al. Genie 3: A new frontier for world models. 2025.

Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15791–15801, 2025.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Maria Elisabeth Bechtle, Feryal Behbahani, Stephanie C.Y. Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments. In *International Conference on Machine Learning (ICML)*, 2024.

Xu Cao, Yifan Shen, Bolin Lai, Wenqian Ye, Yunsheng Ma, Joerg Heintz, Jintai Chen, Meihuan Huang, Jianguo Cao, Aidong Zhang, et al. What is the visual cognition gap between humans and multimodal llms? *arXiv preprint arXiv:2406.10424*, 2024.

Haoxuan Che, Xuanhua He, Quande Liu, Cheng Jin, and Hao Chen. Gamegen-x: Interactive open-world game video generation. In *International Conference on Learning Representations (ICLR)*, 2025.

Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7310–7320, 2024a.

Junsong Chen, Jincheng YU, Chongjian GE, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *International Conference on Learning Representations (ICLR)*, 2024b.

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 (ECCV)*, 2024.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, pp. 720–736, 2018.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):4125–4141, 2020.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal on Computer Vision (IJCV)*, 130(1):33–55, 2022.

Andrey Gorodetskiy, Konstantin Mironov, and Aleksandr Panov. Model-based policy optimization using symbolic world model. In *International Conference on Intelligent Robots and Systems (IROS)*, pp. 664–669, 2024.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18995–19012, 2022.

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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19383–19400, 2024.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International Conference on Machine Learning (ICML)*, pp. 2555–2565, 2019.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:6840–6851, 2020.

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

Yifei Huang, Minjie Cai, Zhenqiang Li, and Yoichi Sato. Predicting gaze in egocentric video by learning task-dependent attention transition. In *European Conference on Computer Vision (ECCV)*, pp. 754–769, 2018.

Quan Huynh-Thu and Mohammed Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics letters*, 44(13):800–801, 2008.

Wenqi Jia, Miao Liu, and James M Rehg. Generative adversarial network for future hand segmentation from egocentric video. In *European Conference on Computer Vision (ECCV)*, pp. 639–656, 2022.

Taewoong Kang, Kinam Kim, Dohyeon Kim, Minho Park, Junha Hyung, and Jaegul Choo. Egox: Egocentric video generation from a single exocentric video. *arXiv preprint arXiv:2512.08269*, 2025.

Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *International Conference on Computer Vision (ICCV)*, pp. 5492–5501, 2019.

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

Jiaman Li, Karen Liu, and Jiajun Wu. Ego-body pose estimation via ego-head pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17142–17151, 2023.

Qifeng Li, Xiaosong Jia, Shaobo Wang, and Junchi Yan. Think2drive: Efficient reinforcement learning by thinking with latent world model for autonomous driving (in carla-v2). In *European Conference on Computer Vision (ECCV)*, pp. 142–158, 2024a.

Yayuan Li, Zhi Cao, and Jason J Corso. Handi: Hand-centric text-and-image conditioned video generation. *arXiv preprint arXiv:2412.04189*, 2024b.

Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, et al. Open-sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*, 2024.

Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z Xu, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:7575–7586, 2022.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

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

Shaowei Liu, Subarna Tripathi, Somdeb Majumdar, and Xiaolong Wang. Joint hand motion and interaction hotspots prediction from egocentric videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3282–3292, 2022.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

Yuanzhe Liu, Jingyuan Zhu, Yuchen Mo, Gen Li, Xu Cao, Jin Jin, Yifan Shen, Zhengyuan Li, Tianjiao Yu, Wenzhen Yuan, et al. Palm: Progress-aware policy learning via affordance reasoning for long-horizon robotic manipulation. *arXiv preprint arXiv:2601.07060*, 2026.

Zhengyi Luo, Ryo Hachiuma, Ye Yuan, and Kris Kitani. Dynamics-regulated kinematic policy for egocentric pose estimation. *Advances in Neural Information Processing Systems (NeurIPS)*, 34: 25019–25032, 2021.

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 (ECCV)*, pp. 445–465, 2024.

Yutaka Matsuo, Yann LeCun, Maneesh Sahani, Doina Precup, David Silver, Masashi Sugiyama, Eiji Uchibe, and Jun Morimoto. Deep learning, reinforcement learning, and world models. *Neural Networks*, 152:267–275, 2022.

NVIDIA, Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, et al. World simulation with video foundation models for physical ai, 2025.

OpenAI. Creating video from text. In *https://openai.com/index/sora/*, 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.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *International Conference on Computer Vision (ICCV)*, pp. 4195–4205, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations (ICLR)*, 2024.

Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang. Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. In *International Conference on Computer Vision (ICCV)*, pp. 5285–5297, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Francesco Ragusa, Giovanni Maria Farinella, and Antonino Furnari. Stillfast: An end-to-end approach for short-term object interaction anticipation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3636–3645, 2023.

Ruslan Rakhimov, Denis Volkhonskiy, Alexey Artemov, Denis Zorin, and Evgeny Burnaev. Latent video transformer. *arXiv preprint arXiv:2006.10704*, 2020.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.

Lloyd Russell, Anthony Hu, Lorenzo Bertoni, George Fedoseev, Jamie Shotton, Elahe Arani, and Gianluca Corrado. Gaia-2: A controllable multi-view generative world model for autonomous driving. *arXiv preprint arXiv:2503.20523*, 2025.

Yifan Shen, Yuanzhe Liu, Jingyuan Zhu, Xu Cao, Xiaofeng Zhang, Yixiao He, Wenming Ye, James Matthew Rehg, and Ismini Lourentzou. Fine-grained preference optimization improves spatial reasoning in vlms. *arXiv preprint arXiv:2506.21656*, 2025.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021.

Yuanpeng Tu, Hao Luo, Xi Chen, Xiang Bai, Fan Wang, and Hengshuang Zhao. Playerone: Egocentric world simulator. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. In *International Conference on Learning Representations (ICLR)*, 2025.

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

Jian Wang, Diogo Luvizon, Weipeng Xu, Lingjie Liu, Kripasindhu Sarkar, and Christian Theobalt. Scene-aware egocentric 3d human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13031–13040, 2023.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. *arXiv preprint arXiv:2503.11651*, 2025.

Xiaofeng Wang, Kang Zhao, Feng Liu, Jiayu Wang, Guosheng Zhao, Xiaoyi Bao, Zheng Zhu, Yingya Zhang, and Xingang Wang. Egovid-5m: A large-scale video-action dataset for egocentric video generation. *arXiv preprint arXiv:2411.08380*, 2024.

Xiaohan Wang, Linchao Zhu, Heng Wang, and Yi Yang. Interactive prototype learning for egocentric action recognition. In *International Conference on Computer Vision (ICCV)*, pp. 8168–8177, 2021.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

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. *arXiv preprint arXiv:2507.07982*, 2025.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, et al. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. *arXiv preprint arXiv:2410.10629*, 2024.

Jingqiao Xiu, Fangzhou Hong, Yicong Li, Mengze Li, Wentao Wang, Sirui Han, Liang Pan, and Ziwei Liu. Egotwin: Dreaming body and view in first person. *arXiv preprint arXiv:2508.13013*, 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.

Bufang Yang, Lilin Xu, Liekang Zeng, Kaiwei Liu, Siyang Jiang, Wenrui Lu, Hongkai Chen, Xiaofan Jiang, Guoliang Xing, and Zhenyu Yan. Contextagent: Context-aware proactive llm agents with open-world sensory perceptions. *arXiv preprint arXiv:2505.14668*, 2025.

Yufei Ye, Xueting Li, Abhinav Gupta, Shalini De Mello, Stan Birchfield, Jiaming Song, Shubham Tulsiani, and Sifei Liu. Affordance diffusion: Synthesizing hand-object interactions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22479–22489, 2023.

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*, 2025a.

Tianjiao Yu, Xinzhuo Li, Yifan Shen, Yuanzhe Liu, and Ismini Lourentzou. Core3d: Collaborative reasoning as a foundation for 3d intelligence. *arXiv preprint arXiv:2512.12768*, 2025b.

Binjie Zhang and Mike Zheng Shou. Ego-centric predictive model conditioned on hand trajectories. *arXiv preprint arXiv:2508.19852*, 2025.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.

Haoyu Zhang, Qiaohui Chu, Meng Liu, Haoxiang Shi, Yaowei Wang, and Liqiang Nie. Exo2ego: Exocentric knowledge guided mllm for egocentric video understanding. *arXiv preprint arXiv:2503.09143*, 2025a.

Mengmi Zhang, Keng Teck Ma, Joo Hwee Lim, Qi Zhao, and Jiashi Feng. Deep future gaze: Gaze anticipation on egocentric videos using adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4372–4381, 2017.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Yifan Zhang, Chunli Peng, Boyang Wang, Puyi Wang, Qingcheng Zhu, Fei Kang, Biao Jiang, Zedong Gao, Eric Li, Yang Liu, et al. Matrix-game: Interactive world foundation model. *arXiv preprint arXiv:2506.18701*, 2025b.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. Diffusionnft: Online diffusion reinforcement with forward process. *arXiv preprint arXiv:2509.16117*, 2025.

Yunsong Zhou, Michael Simon, Zhenghao Peng, Sicheng Mo, Hongzi Zhu, Minyi Guo, and Bolei Zhou. Simgen: Simulator-conditioned driving scene generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 48838–48874, 2024.

Haoyi Zhu, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Chunhua Shen, Jiangmiao Pang, and Tong He. Aether: Geometric-aware unified world modeling. In *International Conference on Computer Vision (ICCV)*, 2025.

APPENDIX

A Related Work 17

B Why Trajectory-Level Alignment 17

C GRAFT Reward Details 18

D X-Ego Details 19

E Implementation Details 19

F Video Generation Evaluation 20

G Details of User Study 20

H Additional Ablation Studies 21

H.1 Component Factorization . . . 21
H.2 GRAFT Design . . . 22
H.3 Geometry Weak Supervision . . . 23
H.4 Instruction Sensitivity . . . 23
H.5 Evaluation without Exocentric Input . . . 23
H.6 Effect of Exocentric Reference Timestamp . . . 24
H.7 Component Ablation with a Clip-Start Reference . . . 24
H.8 Reward Component Ablation . . . 25

I Refinement across Video Backbones 25

J Reliability of the VLM Reward Judge 25

J.1 Agreement across judges . . . 26
J.2 Robustness to the choice of judge . . . 26
J.3 Agreement with human raters . . . 26
J.4 Reward hacking . . . 27

K Additional Qualitative Examples 27

L Device for Real World Experiments 27

M Broader Impacts 29

N Evaluation and Annotation Prompts 30

A Related Work

Egocentric Vision. Egocentric vision research has rapidly advanced with improvements in both dataset quality and model scaling. Foundational benchmarks such as EPIC-KITCHENS (Damen et al., 2022; 2020; 2018), Ego4D (Grauman et al., 2022), and EgoExo4D (Grauman et al., 2024) have enabled large-scale analysis of daily tasks and multi-view human activity. On the modeling side, prior work focuses on recognizing actions, gaze, attention, and human-object interactions from egocentric views (Huang et al., 2018; Kazakos et al., 2019; Wang et al., 2021; Liu et al., 2022; Ragusa et al., 2023), or on estimating human body pose (Li et al., 2023; Luo et al., 2021; Wang et al., 2023). There is also increasing attention to vision–language models for learning multimodal representations from egocentric data (Lin et al., 2022; Pramanick et al., 2023; Ashutosh et al., 2023; Cao et al., 2024; Shen et al., 2025) and forecasting plausible hand motions (Jia et al., 2022), hand-object interactions (Ye et al., 2023), and gaze (Zhang et al., 2017). Methods for cross-view exocentric-to-egocentric translation and joint egocentric video-motion synthesis depend on explicit motion supervision, such as camera trajectories or synchronized multi-camera exocentric recordings (Cheng et al., 2024; Liu et al., 2024; Xiu et al., 2025; Park et al., 2025). However, such data are difficult to acquire and still fail to capture high-level, goal-driven human intent, limiting controllability and semantic diversity in generated egocentric videos. In contrast, EgoForge generates egocentric videos without requiring motion or view supervision.

Video Generation. Advances in latent diffusion (Rombach et al., 2022; Podell et al., 2024; Chen et al., 2024b) and score-based generative modeling (Song & Ermon, 2019; Song et al., 2021; Ho et al., 2020) established the principles of iterative denoising and latent-space synthesis that now underpin most video generators. Building on these foundations, transformer-based video diffusion architectures enable greater temporal coherence, longer video duration, and improved realism (OpenAI, 2024; Peebles & Xie, 2023; Kong et al., 2024; Xie et al., 2024). Large-scale frameworks such as Stable Video Diffusion (Blattmann et al., 2023), VideoCrafter (Chen et al., 2023; 2024a), and Open-Sora (Lin et al., 2024) demonstrate strong cross-domain synthesis through efficient latent modeling and scaling. While these models excel at general-purpose video generation, they lack representations of agent intent, viewpoint dynamics, and causal continuity required for egocentric simulation.

World Models. World models are predictive models trained to simulate the environment dynamics, serving either as tools for agent policy optimization (Hafner et al., 2019; Gorodetskiy et al., 2024; Li et al., 2024a) or as high-fidelity world simulators and game engines (Alonso et al., 2024; Valevski et al., 2025; Bruce et al., 2024; Che et al., 2025; Liu et al., 2026; Yu et al., 2025b). Recent efforts such as Matrix-Game (Zhang et al., 2025b), Cosmos (Agarwal et al., 2025), and Aether (Zhu et al., 2025) generate unlimited and interactive video streams of 3D worlds, but lack explicit understanding of user intent and task context. Reward-based post-training of diffusion models has been explored for images through likelihood-free or forward-process objectives, as in DiffusionNFT (Zheng et al., 2025), but these methods score individual samples and do not address goal completion over temporally extended rollouts. In contrast, EgoForge generates an egocentric video trajectory that fulfills a specified goal through our proposed GRAFT trajectory-level reward-guided refinement that optimizes goal completion and temporal coherence.

B Why Trajectory-Level Alignment

In the main paper, we introduce GRAFT as a trajectory-level reinforcement stage to guide generation using multiple, possibly conflicting conditioning signals. This appendix elaborates on why such an alignment is a non-trivial challenge, moving beyond standard per-frame conditioning. The difficulty arises from coordinating heterogeneous semantic and visual goals over entire, temporally extended video rollouts. We highlight three specific complexities that necessitate a trajectory-level optimization approach.

Reward Sparsity and Temporal Credit Assignment. A primary challenge in egocentric simulation is that the criteria for success are often holistic and only measurable at an action sequence's conclusion. For instance, our goal completion reward ($\mathcal{R}_{\text{goal}}$) evaluates whether a multi-step task (*e.g.*, "put the can back") was *ultimately* successful. This introduces a temporal credit assignment problem. If a generated trajectory of T frames fails, it is non-trivial to pinpoint which frames or input-fusion errors contributed to the failure, and when those errors occurred during the rollout, contributed to the failure. Simple per-frame rewards (*e.g.*, pixel-wise losses) cannot capture this procedural, long-horizon

objective. Therefore, a reinforcement learning mechanism is required to optimize for this sparse, trajectory-wide reward, propagating the global success signal back to the entire generative process.

Preventing Goals Drift. In long-duration video synthesis, generative models are prone to "goals drift," where the generated sequence gradually deviates from the initial conditioning signals. For example, while DiT blocks are conditioned on inputs at each step, there is no explicit mechanism in the standard diffusion objective to guarantee that a scene's geometric structure or background elements (part of $\mathcal{R}_{\text{env}}$) remain consistent from frame 1 to frame T. Without a trajectory-level regularizer, the model might "forget" or "mutate" these goals, leading to artifacts such as a drifting environment or object flickering. The GRAFT stage acts as this global supervisor, evaluating the entire rollout and penalizing sequences that exhibit such drift, thereby enforcing long-term temporal coherence.

The "Shortcut" Problem in Multimodal Fusion. As we note in the main paper, heterogeneous conditioning signals may dominate, conflict, or cancel out. A significant challenge is that the model can learn "shortcuts" to minimize the training loss without effectively combining information from the conditioning inputs. For example, the model might find it easier to overfit to the auxiliary visual inputs, producing a stylistically similar video while entirely ignoring the complex procedural instructions in the text input. This results in a visually plausible rollout that fails the semantic task. Our multi-dimensional reward function (assessing goal, environment, causality, and fidelity) and the negative-aware finetuning structure address this problem. By evaluating rollouts across multiple axes and explicitly repelling the model from "suboptimal" samples (*e.g.*, low $\mathcal{R}_{\text{goal}}$ but high $\mathcal{R}_{\text{per}}$), GRAFT compels the model to find a policy that balances all inputs, discouraging shortcut solutions and enforcing the joint use of conditioning inputs.

C GRAFT Reward Details

This appendix details reward design choices for our proposed GRAFT.

Goal Completion Reward ($\mathcal{R}_{\text{goal}}$). The goal completion reward ($\mathcal{R}_{\text{goal}}$) evaluates the success of the manipulation task based on both the process and the final outcome. This 2.0-point metric is a composite function, $\mathcal{R}_{\text{goal}} = \mathcal{R}_{\text{task}} + \mathcal{R}_{\text{align}}$, where each component has a maximum value of 1.0. The Task Completion score ($\mathcal{R}_{\text{task}}$) assesses the dynamic fidelity of the video, rewarding the agentic world model for performing the state changes and manipulations required to transition from the initial state to the final one. Concurrently, the Visual Alignment score ($\mathcal{R}_{\text{align}}$) evaluates the target outcome of the action by measuring the semantic correspondence and spatial relationship of key elements between the video's final frame and the target state image, ensuring the end state is verifiably correct.

Scene Consistency Reward ($\mathcal{R}_{\text{env}}$). To ensure the model focuses specifically on the manipulation task without altering the surrounding scene, we introduce the scene consistency reward ($\mathcal{R}_{\text{env}}$). This 2.0-point metric is critical for disentangling the required action from the background context, particularly when the target image depicts in a different environment. It is defined as $\mathcal{R}_{\text{env}} = \mathcal{R}_{\text{consist}} + \mathcal{R}_{\text{contam}}$. The Consistency Score ($\mathcal{R}_{\text{consist}}$) measures the temporal stability of static environmental features between the initial state image and the entire video sequence, penalizing any unmotivated environmental drift. The Contamination Score ($\mathcal{R}_{\text{contam}}$) specifically penalizes the model for "hallucinating" or leaking environmental features (*e.g.*, lighting, background objects, textures) that are present only in the target image but not in the initial state.

Temporal Causality Reward ($\mathcal{R}_{\text{temp}}$). The temporal causality reward ($\mathcal{R}_{\text{temp}}$) assesses the physical and logical coherence of the generated video's dynamics. A video that successfully achieves the goal but does so through impossible or nonsensical motion is penalized. This 2.0-point metric, $\mathcal{R}_{\text{temp}} = \mathcal{R}_{\text{phys}} + \mathcal{R}_{\text{logic}}$, evaluates how plausibly the task is performed. The Physics Plausibility score ($\mathcal{R}_{\text{phys}}$) scrutinizes the sequence for adherence to real-world physical principles, such as momentum, gravity, and object permanence, penalizing unrealistic motion or object interactions. To complement this, the Causal Logic score ($\mathcal{R}_{\text{logic}}$) ensures that all state changes (effects) are preceded by logical and visible interactions (causes), such as an agent's hand grasping an object before it moves, thereby preventing spurious correlations or "action-at-a-distance" phenomena.

Perceptual Fidelity Reward ($\mathcal{R}_{\text{per}}$). This score is computed as the sum of three clipped and normalized components, $\mathcal{R}_{\text{PSNR}}$, $\mathcal{R}_{\text{FVD}}$, and $\mathcal{R}_{\text{LPIPS}}$, each contributing a maximum of $\frac{2}{3}$ points. PSNR measures low-level, pixel-based reconstruction fidelity, LPIPS assesses perceptual similarity by comparing

deep features in a manner aligned with human judgment, and FVD is a distributional metric that evaluates the overall realism of both visual appearance and temporal motion against a set of real videos.

D X-EGO DETAILS

We introduce X-Ego, the first large-scale dataset and benchmark designed to evaluate the capability of egocentric world models to synthesize complex, goal-directed scenes from sparse static context. X-Ego is constructed from the Nymeria (Ma et al., 2024) and Ego-Exo4D (Grauman et al., 2024) datasets. We segment the videos based on the action annotations provided with these two datasets, resulting in clips that are each 10 seconds long. We then instruct an expert temporal-action summarizer to generate concise descriptions of stationary atomic actions lasting 10 seconds. The summarizer is specifically constrained to select hand-on-object manipulations while strictly avoiding non-stationary actions, locomotion, speech, and idle states, relying primarily on the provided filtered annotations to ensure the selected action dominates the segment. This setup guarantees that the synthesized descriptions focus exclusively on fine-grained in-place manipulation tasks.

To better understand human–environment interaction, we enrich this benchmark with dense annotations that detail fine-grained hand-object dynamics, object state changes, and step-level semantics. Instead of simple textual expansion, we instruct a multimodal language model to refine the concise atomic action description by grounding it in the actual video content, ensuring that the generated caption strictly adheres to the visual evidence. The output follows a structured four-sentence format, sequentially detailing the initial visual setup, the micro-dynamics of the motion, the physical reaction of the object, and the outcome. The dataset contains 15,000 training samples plus 100 separately held-out test clips, for a total of 15,100 samples. The test set is disjoint from the training set: none of its 100 clips is included in either the 13,000-sample Stage-1 training set or the 2,000-sample Stage-2 training set. The test set contains samples that span the benchmark's interaction taxonomy, balancing broad coverage with the cost of evaluating long-horizon, high-fidelity video generation under multiple complementary metrics.

E IMPLEMENTATION DETAILS

Backbone and data. EgoForge uses the Wan2.2-5B (Wan et al., 2025) model as the base generator. The 15,000 training samples are allocated to two stages; the separate 100-clip test set is used only for evaluation. We train with a batch size of 1 at a resolution of 720p, processing input videos at 24 fps and using 241 frames per sequence. Both stages fine-tune the diffusion transformer with Low-Rank Adaptation (LoRA) of rank 32, optimized with Adam at a learning rate of $1e^{-4}$ under mixed-precision (bf16) training on 8 H100 GPUs.

Stage 1: denoising fine-tuning. The first stage fine-tunes the model on 13,000 samples with the velocity-matching loss $\mathcal{L}_{\mathrm{D}}$ and the geometry weak supervision $\mathcal{L}_{\mathrm{G}}$ of Section 3.1, sampling $t \sim \mathcal{U}(0,1)$ as in Eq. (2). The DINOv3 and VGGT backbones are frozen. This stage runs for 10 epochs and requires approximately 108 hours.

Stage 2: GRAFT. The second stage uses the remaining 2,000 conditions. Only the diffusion transformer is trained, while all other components are frozen. The sampling policy π^{old} is the frozen Stage-1 model. For each condition we draw $K=6$ rollouts from π^{old} with the same sampler used at evaluation, giving 12,000 scored rollouts, and score each rollout with $\mathcal{R}_{\mathrm{total}}$ (Appendix C). Only the clean rollouts and their rewards are stored; the sampling trajectories are not needed. Rewards are normalized within each group of K rollouts: μ_c in Eq. (5) is the group mean and Z_c is the standard deviation of the K raw rewards, floored at a small constant to avoid division by zero. A condition whose rollouts all receive the same reward therefore maps every rollout to $r=\frac{1}{2}$ and, by Eq. (8), carries no reward signal: such groups only anchor $\boldsymbol{v}_\theta$ to the frozen $\boldsymbol{v}^{\mathrm{old}}$. We set the guidance parameter to $\beta=0.1$ in Eq. (7), sample $t \sim \mathcal{U}(0,1)$ and $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ exactly as in Stage 1, and evaluate $\boldsymbol{v}^{\mathrm{old}}$ and $\boldsymbol{v}_\theta$ on the same noisy latent $\mathbf{z}_t$ so that the implicit positive and negative policies share one forward pass of the frozen model.

Table 6: **User study.** Each rating dimension targets one property of the rollout. Q. is Quality (overall synthesis quality), F. is Fidelity (object consistency), M. is Smooth Motion (physical plausibility), E. is Smooth Environment (scene consistency), and A. is Alignment (goal completion). Each dimension is rated from 1 (worst) to 5 (best). *Best model variants from Table 3.

Model	Q.↑	F.↑	M.↑	E.↑	A.↑
Cosmos* (NVIDIA et al., 2025)	3.29	2.54	3.07	2.47	2.19
Hunyuan* (Kong et al., 2024)	3.46	2.86	3.72	3.16	3.08
WAN2.2* (Wan et al., 2025)	3.22	3.48	3.82	4.07	3.15
EgoForge (Ours)	**4.58**	**4.71**	**4.25**	**4.48**	**4.75**

F Video Generation Evaluation

As summarized in the main paper, we employ a comprehensive suite of metrics to evaluate the performance of EgoForge on the X-Ego benchmark. We describe how each metric is computed below.. For semantic and perceptual alignment, we compute the **DINO-Score** to evaluate frame-level semantic similarity, which is calculated as the average cosine similarity between the DINOv2 ViT-g embeddings of the predicted video frames and the ground-truth video frames. We also compute the **CLIP-Score** to measure text-video alignment. This is calculated as the cosine similarity between the CLIP embedding of the generated video frames and the embedding of the corresponding textual guidance. Additionally, **LPIPS** (Learned Perceptual Image Patch Similarity) assesses perceptual distance by computing the L_2 distance between deep features extracted from a pretrained network (*e.g.*, VGG), aligning more closely with human judgments of visual similarity than pixel-based metrics. For low-level visual fidelity, **PSNR (Peak Signal-to-Noise Ratio)** measures pixel-level reconstruction quality by comparing the maximum possible pixel value to the Mean Squared Error (MSE) between the generated and ground-truth frames, where a higher PSNR indicates lower pixel-wise error. **SSIM (Structural Similarity Index)** assesses visual fidelity by comparing three components: luminance, contrast, and structural information between the generated and ground-truth frames. Finally, for distributional and temporal fidelity, **FVD (Fréchet Video Distance)** assesses distributional realism by measuring the Fréchet Inception Distance between distributions of real and generated videos. Features are extracted from a pretrained video classification model, and the distance between the two multivariate Gaussian distributions (one for real videos, one for generated) is computed. **Flow MSE** measures temporal and motion fidelity. We first compute the optical flow fields for both the generated and ground-truth video sequences, and then calculate the mean squared error (MSE) between the two sets of flow fields, penalizing discrepancies in predicted motion.

G Details of User Study

We recruit 20 annotators to evaluate 25 groups of videos, where each group contains outputs from all competing methods paired with the corresponding text prompt. Participation was voluntary, and no monetary or non-monetary compensation was provided. Participants were provided with a detailed annotation protocol to rate each video on a 1-to-5 scale across five dimensions: Quality (overall visual coherence), Fidelity (identity preservation and absence of artifacts or distortions), Smooth Motion (temporal consistency and fluidity of subject movement), Smooth Environment (background stability and smoothness), and Alignment (semantic correspondence between video and input). As shown in Table 6, our method achieves substantial gains in Alignment (4.75) and Fidelity (4.71), surpassing the strongest competitors (*e.g.*, +1.60 in Alignment over WAN2.2). These results, combined with superior scores in temporal and environmental smoothness, validate the effectiveness of our approach in generating high-quality, prompt-aligned videos.

Table 7 compares EgoForge before and after GRAFT refinement. The Stage-1 model already includes our conditioning mechanism and geometry weak supervision, so this comparison isolates the contribution of GRAFT. Human ratings improve across all four reported dimensions. The largest gain is in alignment with the goal instruction, which increases from 3.94 to 4.75, followed by fidelity (4.16 to 4.71), environment stability (4.21 to 4.48), and motion smoothness (4.02 to 4.25). These results show that GRAFT improves instruction following while also producing more consistent objects, more stable backgrounds, and smoother motion.

Table 7: **Effect of GRAFT on human ratings.** Stage-1 (+CI) uses the same conditioning mechanism and geometry supervision as EgoForge but has not undergone GRAFT refinement. Scores range from 1 (worst) to 5 (best), and Δ denotes the improvement over Stage-1. GRAFT improves all four reported dimensions, with the largest gain in alignment with the goal instruction.

Rating dimension	Property measured	Wan2.2*	Stage-1 (+CI)	EgoForge	Δ over Stage-1
Alignment	Goal completion	3.15	3.94	**4.75**	+0.81
Fidelity	Object consistency	3.48	4.16	**4.71**	+0.55
Smooth Environment	Scene consistency	4.07	4.21	**4.48**	+0.27
Smooth Motion	Physical plausibility	3.82	4.02	**4.25**	+0.23

Under identical conditioning and geometry supervision, GRAFT gives the largest gain in goal completion, and also improves object consistency, scene consistency, and physical plausibility. This ordering matches the design of the reward, where the goal term carries the largest weight in distinguishing successful from failed rollouts.

User Study: Participant Instructions

The following instructions were provided to participants:
"You will be shown a text prompt and several generated egocentric videos corresponding to the same prompt. Please evaluate each video independently. You may replay the videos as needed. For each video, assign a score from 1 (worst) to 5 (best) for each of the following criteria:
1. Quality: overall visual quality and coherence of the generated video. A score of 1 indicates severe visual artifacts or an unrealistic video; a score of 5 indicates a visually coherent and realistic video.
2. Fidelity: preservation of object identity and absence of artifacts or distortions. A score of 1 indicates severe object deformation, identity inconsistency, or hallucinated artifacts; a score of 5 indicates stable object appearance and few or no artifacts.
3. Smooth Motion: temporal consistency and smoothness of the subject or hand/object motion. A score of 1 indicates jittery, discontinuous, or physically implausible motion; a score of 5 indicates smooth and temporally coherent motion.
4. Smooth Environment: stability and smoothness of the background environment. A score of 1 indicates severe background flickering, drifting, or scene changes; a score of 5 indicates a stable and consistent environment.
5. Alignment: semantic correspondence between the generated video and the input text prompt. A score of 1 indicates that the video is unrelated to or fails to follow the prompt; a score of 5 indicates that the video clearly follows and completes the described action.
Please use the full 1-to-5 scale when appropriate and base your judgment only on the displayed prompt and video content."

The study used a simple rating form displaying the text prompt, generated videos, and the five rating questions above. No additional custom interface was used. The user study involved only anonymous subjective ratings of generated videos and text prompts. Participants were informed of the evaluation purpose, the rating criteria, and that their responses would be analyzed only in aggregate. No sensitive personal information was collected. The potential risks were minimal and limited to ordinary visual fatigue from viewing short videos.

H Additional Ablation Studies

H.1 Component Factorization

We first factorize the effects of denoising fine-tuning, geometry weak supervision (GWS), and GRAFT. The structured ego/exo/text conditioning pathway is kept fixed for all fine-tuned variants. We also include an extra-SFT control that uses the same additional optimization budget as GRAFT but continues training only with the standard denoising objective.

Table 8: **Component ablation.** GWS and GRAFT provide complementary gains. Extra SFT uses the same additional training budget as GRAFT but does not use reward-guided refinement.

Variant	DINO↑	CLIP↑	SSIM↑	LPIPS↓	FVD↓	Flow MSE↓	PSNR↑
Denoising FT	56.81	37.10	0.74	0.21	260.89	4.82	21.92
Denoising FT + GWS	58.92	38.05	0.76	0.18	218.72	3.92	22.87
Denoising FT + GRAFT	59.46	38.61	0.77	0.17	203.84	3.41	23.35
Denoising FT + GWS + extra SFT	59.18	38.22	0.77	0.17	211.36	3.74	23.14
EgoForge	**61.25**	**39.30**	**0.79**	**0.15**	**182.25**	**2.83**	**24.08**

Table 9: **Gain attribution.** The left column compares EgoForge with the off-the-shelf Wan2.2 backbone. The right column compares the GRAFT stage with our equally conditioned Stage-1 model.

Metric	**Full-system gain**	**Method-only gain**
DINO↑	$+13.4\%$ $(53.99 \rightarrow 61.25)$	$+3.95\%$ $(58.92 \rightarrow 61.25)$
CLIP↑	$+10.1\%$ $(35.69 \rightarrow 39.30)$	$+3.29\%$ $(38.05 \rightarrow 39.30)$
SSIM↑	$+9.7\%$ $(0.72 \rightarrow 0.79)$	$+3.95\%$ $(0.76 \rightarrow 0.79)$
LPIPS↓	-34.8% $(0.23 \rightarrow 0.15)$	-16.7% $(0.18 \rightarrow 0.15)$
FVD↓	-43.4% $(322.17 \rightarrow 182.25)$	-16.7% $(218.72 \rightarrow 182.25)$
Flow MSE↓	-51.0% $(5.78 \rightarrow 2.83)$	-27.8% $(3.92 \rightarrow 2.83)$
PSNR↑	$+17.8\%$ $(20.44 \rightarrow 24.08)$	$+5.3\%$ $(22.87 \rightarrow 24.08)$

As shown in Table 8, denoising fine-tuning alone improves the base generator, but still leaves a large gap in temporal coherence and structural fidelity. Adding GWS reduces FVD from 260.89 to 218.72 and Flow MSE from 4.82 to 3.92, indicating better geometric and temporal stability. GRAFT also gives a strong improvement without GWS, reducing FVD to 203.84 and Flow MSE to 3.41. Combining both achieves the best results across all metrics. The extra-SFT control remains clearly below the full model, suggesting that the improvement is not simply due to additional training steps.

H.2 GRAFT Design

Table 9 distinguishes the overall improvement of EgoForge from the additional contribution of GRAFT. The full-system gain compares EgoForge with off-the-shelf Wan2.2, capturing the combined effect of our conditioning mechanism, denoising fine-tuning, geometry weak supervision, and refinement. The method-only gain compares EgoForge with Stage-1 (Wan2.2+CI), which already includes all components except GRAFT. Relative to off-the-shelf Wan2.2, the full system reduces FVD by 43.4% and Flow MSE by 51.0%. Relative to Stage-1, GRAFT provides further reductions of 16.7% and 27.8%, respectively, while also improving all five remaining metrics.

We next examine each design choice in GRAFT. All variants in Table 10 start from the same Denoising FT + GWS checkpoint. We compare against reward re-ranking, positive-only refinement, removing per-condition reward normalization, and a shuffled-reward control that breaks the correspondence between rollouts and their rewards. Table 10 shows that full GRAFT outperforms both per-frame refinement and inference-time reward re-ranking. Applying the image-level Diffusion-NFT objective independently to video frames increases FVD from 218.72 to 231.60, performing worse than the unrefined model. This result highlights the limitations of frame-level evaluation: realistic individual frames do not ensure that a video performs the required actions in the correct order, maintains hand–object contact, or reaches the goal through physically plausible transitions. Best-of-6 reward re-ranking improves FVD to 196.44 but remains worse than full GRAFT at 182.25, despite requiring six generated candidates per condition at inference. Positive-only refinement also improves over the unrefined model but falls short of the full negative-aware objective, supporting the benefit of explicitly discouraging low-reward trajectories rather than only reinforcing high-reward ones. Performance also declines without per-condition reward normalization, which accounts for differences in reward scale across conditions. Shuffling the rewards across rollouts raises FVD to 229.55, worse than no refinement, showing that the gains depend on meaningful correspondence between rollouts and their rewards. Finally, removing future-reference reward terms retains most of the improvement, achieving an FVD of 188.94. These terms therefore provide an additional benefit, but GRAFT still substantially improves generation without them.

Table 10: **Ablation of GRAFT design choices.** All variants start from the same Denoising FT + GWS checkpoint. Best-of-6 re-ranking selects the highest-reward rollout from six candidates at inference, without further training. Per-frame DiffusionNFT scores and refines individual frames rather than complete rollouts. Full GRAFT achieves the best performance across all reported metrics, outperforming both alternatives and its ablated variants.

Refinement method	DINO↑	CLIP↑	SSIM↑	LPIPS↓	FVD↓	Flow MSE↓	PSNR↑
No GRAFT	58.92	38.05	0.76	0.18	218.72	3.92	22.87
Per-frame DiffusionNFT (Zheng et al., 2025) (naive port)	58.13	–	–	–	231.60	4.05	–
Shuffled-reward GRAFT	58.47	37.89	0.75	0.20	229.55	4.21	22.41
Best-of-6 reward re-ranking	60.04	38.73	0.78	0.16	196.44	3.26	23.64
Positive-only GRAFT	60.29	38.84	0.78	0.16	190.83	3.08	23.70
GRAFT w/o per-condition normalization	60.12	38.70	0.77	0.17	194.38	3.18	23.56
GRAFT w/o future-reference reward terms	60.72	39.04	0.78	0.16	188.94	3.02	23.77
Full GRAFT	**61.25**	**39.30**	**0.79**	**0.15**	**182.25**	**2.83**	**24.08**

H.3 Geometry Weak Supervision

We further study the geometry alignment objective. All variants in Table 11 use the same denoising fine-tuning and GRAFT stages, differing only in the auxiliary alignment target or loss form. Random feature alignment uses a frozen randomly initialized network with matched feature dimensionality. DINO feature alignment replaces VGGT features with semantic image features.

Table 11: **Ablation of geometry weak supervision.** VGGT-based geometry alignment is more effective than random or semantic feature alignment. Angular and scale consistency are complementary.

Geometry variant	DINO↑	CLIP↑	SSIM↑	LPIPS↓	FVD↓	Flow MSE↓	PSNR↑
No GWS	59.46	38.61	0.77	0.17	203.84	3.41	23.35
Random frozen feature alignment	58.74	38.02	0.75	0.20	207.92	3.58	22.60
DINO feature alignment	59.88	38.76	0.77	0.17	198.70	3.30	23.43
VGGT angular only	60.36	39.02	0.78	0.16	191.85	3.14	23.59
VGGT scale only	60.08	38.86	0.77	0.16	196.44	3.27	23.47
VGGT angular + scale	**61.25**	**39.30**	**0.79**	**0.15**	**182.25**	**2.83**	**24.08**

The random-feature control does not improve performance, showing that the gains are not caused by simply adding an auxiliary loss. DINO feature alignment brings moderate improvements, but remains weaker than VGGT-based alignment, especially on FVD and Flow MSE. This suggests that semantic feature alignment alone is insufficient for stable egocentric rollout generation. Between the two geometry losses, angular alignment provides stronger temporal gains, while scale alignment contributes to perceptual and structural stability. Their combination gives the best result across all metrics, supporting the role of geometry-aware representation alignment in maintaining coherent first-person scene evolution.

H.4 Instruction Sensitivity

We evaluate how strongly the goal instruction influences the generated rollout by replacing the correct instruction with one specifying an incompatible action, while keeping the initial egocentric frame fixed. In both settings, we compute CLIP-Score against the original instruction. The resulting score decrease measures how much changing the instruction reduces alignment with the original goal, with a larger decrease indicating greater instruction sensitivity. As shown in Table 12, replacing the instruction decreases CLIP-Score by only 1.09 points for off-the-shelf Wan2.2 and 1.29 points for Wan2.2+TT, indicating limited sensitivity to the instruction change. In contrast, the decrease is 4.21 points for Stage-1 and 7.62 points for EgoForge. The larger gap after GRAFT refinement indicates that the generated rollouts are more responsive to changes in the goal instruction.

H.5 Evaluation without Exocentric Input

We evaluate EgoForge without the auxiliary exocentric image at inference, using only the initial egocentric frame and goal instruction. Table 13 compares this setting with the Stage-1 model, which retains all three conditioning inputs. Despite receiving less visual context, EgoForge outperforms Stage-1 on six of the seven metrics. FVD decreases from 218.72 to 204.62 and Flow MSE from 3.92 to 3.19, while CLIP-Score increases from 38.05 to 38.72. DINO-Score is slightly lower (58.73 vs.

Table 12: **Counterfactual instruction test.** CLIP-Score under the correct instruction and under an incompatible one, with the initial egocentric frame held fixed. A larger Δ means the model responds more strongly to the goal.

Model	Correct instruction	Counterfactual instruction	$\Delta\uparrow$
Wan2.2 off-the-shelf	35.69	34.60	1.09
Wan2.2+TT	36.20	34.91	1.29
Wan2.2+CI (Stage-1)	38.05	33.84	4.21
EgoForge	**39.30**	31.68	**7.62**

Table 13: **Generation without an exocentric reference.** Stage-1 (Wan2.2+CI) receives the initial egocentric frame, goal instruction, and exocentric image, whereas EgoForge receives only the egocentric frame and instruction at inference. Despite the reduced visual input, EgoForge outperforms Stage-1 on six of seven metrics. Δ denotes the score for EgoForge minus the corresponding Stage-1 score.

Config	DINO↑	CLIP↑	SSIM↑	LPIPS↓	FVD↓	Flow MSE↓	PSNR↑
Wan2.2+CI, with exo	58.92	38.05	0.76	0.18	218.72	3.92	22.87
EgoForge, exo deleted	58.73	**38.72**	**0.77**	**0.17**	**204.62**	**3.19**	**23.30**
Δ	-0.19	$+0.67$	$+0.01$	-0.01	-14.10	-0.73	$+0.43$

Table 14: **Effect of exocentric reference timing and content.** We evaluate EgoForge with no exocentric input, a reference from the start or end of the same clip, or an image from an unrelated scene. A clip-start reference improves all four metrics over no reference, while the clip-end reference used in the main results provides further gains. An unrelated reference performs worse than omitting the auxiliary view.

Exocentric source	DINO↑	CLIP↑	FVD↓	Flow MSE↓
No exo (Ego + Text)	58.73	38.72	204.62	3.19
Exo at clip start	60.40	38.90	189.30	2.94
Exo at clip end (main tables)	**61.25**	**39.30**	**182.25**	**2.83**
Random exo from another scene	57.20	37.89	237.44	4.13

58.92). These results show that the refined model retains an advantage over Stage-1 on most metrics even when the auxiliary view is unavailable at inference.

H.6 Effect of Exocentric Reference Timestamp

The exocentric reference used in the main results is taken from the end of the clip and may therefore reveal the final object state. To evaluate whether an auxiliary view remains useful without providing a future frame, Table 14 compares clip-start and clip-end references with two controls: no exocentric input and an exocentric image from an unrelated scene. A clip-start reference improves all four reported metrics over the no-reference setting, reducing FVD from 204.62 to 189.30 and Flow MSE from 3.19 to 2.94. Using a clip-end reference further reduces these values to 182.25 and 2.83, respectively. In contrast, supplying an image from an unrelated scene degrades all four metrics relative to omitting the reference, indicating that the generated rollout is sensitive to the reference content. These results show that the auxiliary view provides a benefit even when it depicts only the initial scene. The clip-end setting achieves the strongest results, but differs from the no-reference and clip-start settings by providing future visual information at inference.

H.7 Component Ablation with a Clip-Start Reference

To evaluate the contribution of each component without providing future exocentric frames at inference, we repeat the component ablation using an exocentric reference from the start of each clip. All variants are evaluated on the same test examples with identical random seeds and generation settings. Table 15 shows that adding geometry weak supervision to denoising fine-tuning improves all seven metrics, and GRAFT provides further gains across every metric. Relative to FT + GWS, GRAFT

Table 15: **Component ablation with a clip-start exocentric reference.** FT denotes denoising fine-tuning, and GWS denotes geometry weak supervision. All variants receive the same conditioning inputs, including an exocentric image from the start of the clip, and use identical test examples, random seeds, and generation settings. Adding GWS and then GRAFT improves all seven metrics without access to future exocentric frames at inference.

Variant	DINO↑	CLIP↑	SSIM↑	LPIPS↓	FVD↓	Flow MSE↓	PSNR↑
FT only	55.96	36.70	0.73	0.22	267.94	4.93	21.67
FT + GWS	58.07	37.65	0.75	0.19	225.77	4.03	22.62
FT + GWS + GRAFT	**60.40**	**38.90**	**0.78**	**0.16**	**189.30**	**2.94**	**23.83**

Table 16: **Effect of reward components in GRAFT.** Each reward term contributes to improved semantic alignment and temporal coherence, while the full reward composition yields the strongest performance across all metrics.

Rewards	DINO-Score↑	CLIP-Score↑	SSIM ↑	LPIPS ↓	FVD ↓	flow MSE ↓	PSNR ↑
✗ $\mathcal{R}_{\text{goal}}$	59.62	38.49	0.78	0.16	205.96	3.48	23.48
✗ $\mathcal{R}_{\text{env}}$	60.67	39.05	0.78	0.16	200.49	3.43	23.60
✗ $\mathcal{R}_{\text{temp}}$	60.78	39.11	0.78	0.16	213.25	3.70	23.72
✗ $\mathcal{R}_{\text{per}}$	60.32	38.80	0.77	0.18	204.13	3.48	23.17
EgoForge (Ours)	**61.25**	**39.30**	**0.79**	**0.15**	**182.25**	**2.83**	**24.08**

increases DINO-Score by 2.33 and CLIP-Score by 1.25, while reducing FVD by 36.47 and Flow MSE by 1.09. Visual fidelity also improves, with SSIM and PSNR increasing by 0.03 and 1.21, respectively, and LPIPS decreasing by 0.03. These results show that the benefits of geometry supervision and GRAFT persist when the exocentric reference depicts only the initial scene rather than the final state.

H.8 Reward Component Ablation

We ablate each reward term in GRAFT while retaining the remaining terms. Table 16 reports the results. The full reward composition achieves the best performance across all seven metrics. Removing $\mathcal{R}_{\text{per}}$ causes the largest deterioration in the visual-quality metrics, reducing SSIM from 0.79 to 0.77 and PSNR from 24.08 to 23.17, while increasing LPIPS from 0.15 to 0.18. Removing $\mathcal{R}_{\text{temp}}$ causes the largest decrease in temporal fidelity, increasing FVD from 182.25 to 213.25 and flow MSE from 2.83 to 3.70. Removing $\mathcal{R}_{\text{goal}}$ causes the largest decreases in semantic alignment, lowering DINO-Score from 61.25 to 59.62 and CLIP-Score from 39.30 to 38.49. Omitting $\mathcal{R}_{\text{env}}$ also degrades all seven metrics, consistent with the value of preserving scene context. These results support the complementary roles of the four reward terms. Appendix J separately examines reward-judge agreement, agreement with human raters, robustness to the choice of judge, and a shortcut probe.

I Refinement across Video Backbones

We evaluate whether the benefits of GRAFT extend beyond the Wan2.2 backbone used in EgoForge. We equip Cosmos and HunyuanVideo with our structured conditioning mechanism and geometry weak supervision, using the same inputs: an initial egocentric frame, a goal instruction, and an auxiliary exocentric image. Table 17 compares each backbone before and after GRAFT refinement, with the conditioning unchanged. GRAFT reduces FVD by 15.9% on Cosmos, 16.1% on HunyuanVideo, and 16.7% on Wan2.2, while improving all six remaining metrics on each backbone. These consistent gains show that the refinement method benefits all three evaluated backbones. The Wan2.2-based EgoForge achieves the best results across all seven metrics. Wan2.2+CI is our Stage-1 model, so its comparison with EgoForge measures the additional contribution of GRAFT. This differs from the comparison with off-the-shelf Wan2.2, which measures the combined benefit of conditioning, geometry supervision, and refinement.

J Reliability of the VLM Reward Judge

We evaluate the consistency of the VLM reward scores used by GRAFT by comparing GPT-5.5, our training-time judge, with Gemini 3.1 Pro and Qwen3-VL 235B. All three judges score the same

Table 17: **Effect of GRAFT across video backbones.** All three backbones use our structured conditioning mechanism and geometry weak supervision (+CI), with the same egocentric frame, goal instruction, and exocentric image as inputs. Each backbone is evaluated before and after GRAFT refinement, which improves all seven metrics. Wan2.2+CI is our Stage-1 model; adding GRAFT yields the full EgoForge model.

Model	DINO↑	CLIP↑	SSIM↑	LPIPS↓	FVD↓	Flow MSE↓	PSNR↑
Cosmos +CI	54.65	32.13	0.74	0.21	338.50	4.91	20.48
Cosmos +CI +GRAFT	56.77	33.21	0.76	0.18	284.66	3.62	21.59
HunyuanVideo +CI	57.44	31.64	0.75	0.20	276.36	4.42	21.18
HunyuanVideo +CI +GRAFT	59.37	32.75	0.77	0.17	231.89	3.28	22.20
Wan2.2 +CI (Stage-1)	58.92	38.05	0.76	0.18	218.72	3.92	22.87
Wan2.2 +CI +GRAFT (EgoForge)	**61.25**	**39.30**	**0.79**	**0.15**	**182.25**	**2.83**	**24.08**

Table 18: **Correlation between VLM reward scores.** We report Spearman rank correlation (ρ) and Pearson correlation (r) between each pair of judges on 300 held-out generated videos (50 conditions, six rollouts per condition). All judges evaluate the same sampled frames using identical prompts. GPT-5.5 is the judge used during GRAFT refinement.

Judge pair	**Spearman ρ**	**Pearson r**
GPT-5.5 and Gemini 3.1 Pro	0.83	0.81
GPT-5.5 and Qwen3-VL 235B	0.78	0.76
Gemini 3.1 Pro and Qwen3-VL 235B	0.76	0.74

300 held-out generated videos, comprising six rollouts for each of 50 conditions, using identical prompts and sampled frames. Table 18 reports pairwise correlations between their scores. GPT-5.5 has Spearman correlations of 0.83 with Gemini 3.1 Pro and 0.78 with Qwen3-VL 235B, indicating that the training judge and both alternatives tend to rank the evaluated rollouts similarly. Table 19 additionally reports intraclass correlations and agreement on the highest-reward rollout within each condition.

J.1 Agreement across judges

Table 19 reports an ICC(2,1) of 0.60 for a single judge and an ICC(2,k) of 0.82 for the mean of all three judges ($k = 3$). The single-judge measure corresponds to our training setup, which uses GPT-5.5 rather than an ensemble. Before computing the refinement objective, GRAFT normalizes rewards within each condition (Eq. (5)). The training signal therefore reflects how each rollout scores relative to the other candidates for the same condition, rather than its absolute score. Table 19 also reports top-1 agreement, which measures whether the judges select the same highest-reward rollout within each condition.

J.2 Robustness to the choice of judge

We repeat GRAFT refinement using Qwen3-VL as the reward judge in place of GPT-5.5. The resulting model achieves an FVD of 187.9 and a DINO-Score of 60.8, compared with 182.25 and 61.25, respectively, when using GPT-5.5. Although performance is slightly lower with Qwen3-VL, both variants improve over Stage-1, which achieves an FVD of 218.72 and a DINO-Score of 58.92. Thus, GRAFT retains its advantage over Stage-1 when the training judge is replaced.

J.3 Agreement with human raters

We compare the reward scores with the independently collected ratings from the blind human study in Appendix G. Table 20 pairs each reward component with the corresponding human rating dimension. Spearman correlations range from 0.71 to 0.79 across the four components. The total reward also correlates with the overall mean human rating ($\rho = 0.80$, $\tau = 0.66$), indicating that higher-reward videos tend to receive higher ratings from human evaluators.

Table 19: **Reliability of VLM reward scores across judges.** ICC(2,1) measures reliability for a single judge, corresponding to our training setup; ICC(2,k), with $k = 3$, measures reliability when averaging scores from all three judges. Top-1 agreement measures how often all three judges select the same highest-reward rollout among the six candidates for a condition.

Reliability measure	Value
ICC(2,1), a single judge, the configuration used for training	0.60
ICC(2,k), mean of the three judges	0.82
Within-condition top-1 agreement, $K = 6$ rollouts	81.4%

Table 20: **Correlation between reward scores and human ratings.** We report Spearman rank correlation (ρ) and Kendall rank correlation (τ) between each reward component and the corresponding rating dimension from the blind human study in Appendix G. The final row compares the total reward R_{total} with the overall mean human rating.

Reward term	Human dimension	Spearman ρ	Kendall τ
$\mathcal{R}_{\text{goal}}$	Alignment	0.79	0.64
$\mathcal{R}_{\text{env}}$	Smooth Environment	0.73	0.58
$\mathcal{R}_{\text{temp}}$	Smooth Motion	0.71	0.56
$\mathcal{R}_{\text{per}}$	Fidelity	0.76	0.61
$\mathcal{R}_{\text{total}}$	Overall mean	0.80	0.66

J.4 Reward hacking

We first examine whether GRAFT improves evaluation measures that are not used to compute the training reward. Under the clip-start evaluation in Table 15, adding GRAFT to FT + GWS increases DINO-Score from 58.07 to 60.40 and CLIP-Score from 37.65 to 38.90, reduces Flow MSE by 27.0% (from 4.03 to 2.94), and increases SSIM from 0.75 to 0.78. None of these four metrics contributes to the reward. The independently collected blind human ratings in Table 7 also improve after refinement. Together, these results provide evidence that the benefits of GRAFT extend beyond the scores it directly optimizes. We do not use improvements in PSNR, FVD, or LPIPS as additional evidence here, since these metrics contribute to $\mathcal{R}_{\text{per}}$.

We also test whether the reward favors a shortcut that depicts the target state without performing the required actions. We construct a static video by repeating the target-state image throughout the clip. Although this video receives a visual-alignment score of 0.82 within $\mathcal{R}_{\text{goal}}$, it scores only 0.21 on temporal causality ($\mathcal{R}_{\text{temp}}$) and 0.30 on environment consistency ($\mathcal{R}_{\text{env}}$). Its composite reward is 0.44, below the median of 0.68 for generated rollouts. Thus, in this test, matching the target image without generating the required action sequence does not yield a high overall reward.

K Additional Qualitative Examples

Across diverse activities and environments, the generated rollouts follow the specified instructions while maintaining stable scene structure and consistent egocentric viewpoint motion. Importantly, complex multi-step behaviors (*e.g.*, manipulating objects, interacting with appliances, or performing physical activities) unfold naturally over time, demonstrating that EgoForge captures both the procedural structure of actions and the surrounding scene context required for goal-directed egocentric simulation. In more dynamic scenarios like rock climbing or shooting a basketball, the rollouts exhibit clear temporal progression toward visually identifiable goals while maintaining consistent environment geometry, including climbing holds and the basketball hoop. These examples highlight the ability of EgoForge to generate temporally coherent, goal-directed egocentric sequences that sustain interaction dynamics and scene consistency over long durations.

L Device for Real World Experiments

For our real-world experiments, we used the DigiLens ARGO smartglasses shown in Figure 9 to capture egocentric video data. The ARGO smartglasses employ Augmented Reality (AR) technology,

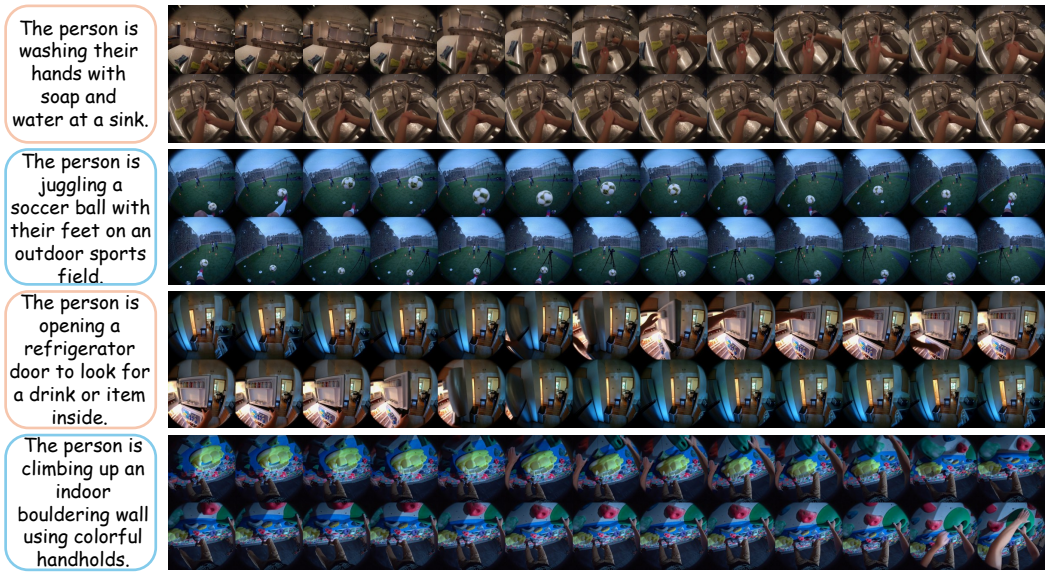


Figure 6: **Qualitative egocentric video rollouts.** EgoForge generates temporally coherent first-person video trajectories that follow the intended activity while preserving scene structure and realistic hand–object interactions across diverse environments.

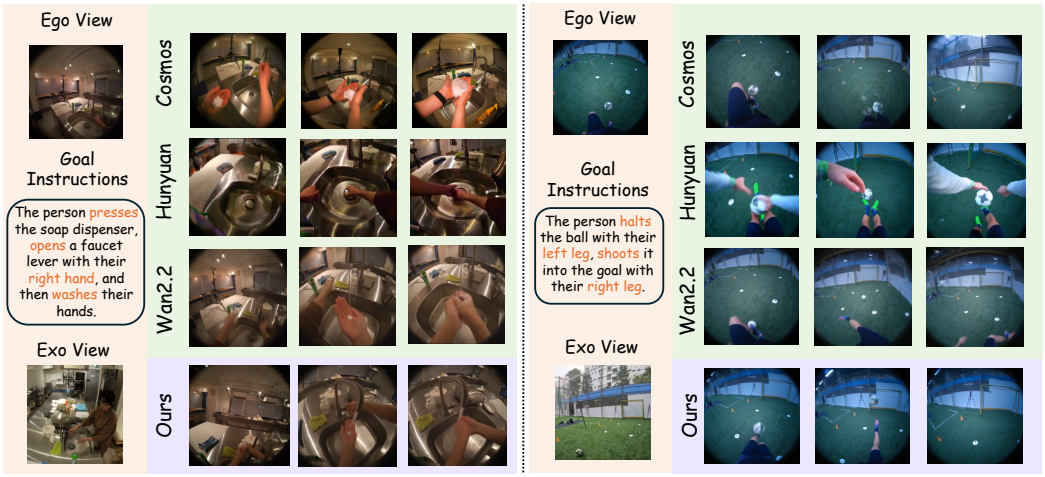


Figure 7: **Qualitative Comparison between EgoForge and baselines**. Sample frames from generated videos illustrating two scenarios. Top row: In the hand-washing task, baselines struggle with object consistency (*e.g.*, Cosmos hallucinates the soap source) or ignore scene context (*e.g.*, Wan2.2 bbypasses the soap on the table), while EgoForge successfully executes the action. using existing objects. Bottom row: In the soccer task, baselines exhibit severe artifacts like ghosting (Cosmos) or fail to follow precise instructions regarding motion and goals (Hunyuan, Wan2.2). EgoForge accurately executes the complex command: trapping with the left leg and shooting with the right.

which creates a visual overlay on the real world seen through the transparent display. This device is an enterprise-grade standalone system based on DigiLens' core waveguide technology and is designed for industrial and enterprise use cases. The ARGO provides several key technical advantages for high-quality data capture, including a 48MP camera with autofocus, optical image stabilization (OIS), and electronic image stabilization (EIS), supporting 4×4 pixel binning and enhanced low-light performance. It also features a sophisticated five-microphone beamforming array designed to pick up the wearer's voice in noisy environments and provides integrated stereo spatial audio recordings. The device runs on the Qualcomm Snapdragon XR2 Platform, providing onboard mobile computing capabilities. The ARGO smartglasses are critical to our evaluation because they allow us to test EgoForge's robustness in real-world, out-of-domain (OOD) settings, addressing a limitation of previous egocen-

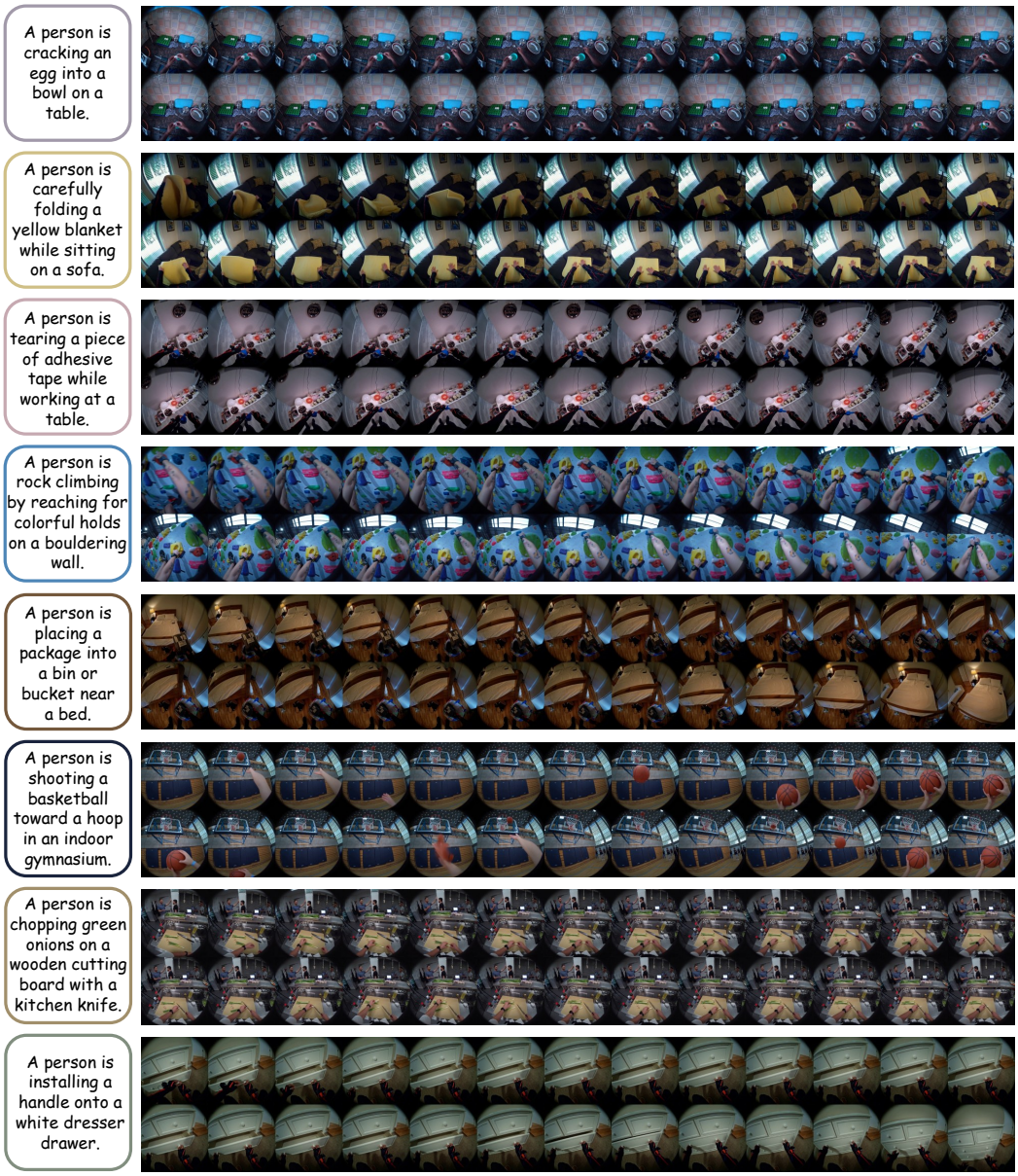


Figure 8: **Dense visualization of long-duration sequences.** For each EgoForge-generated video, we show 26 uniformly sampled frames to illustrate how actions unfold over time. Across diverse first-person tasks, EgoForge preserves scene structure, maintains a stable egocentric viewpoint, and produces smooth, goal-consistent hand–object interactions.

tric world models. The high-fidelity data captured by the ARGO's advanced sensor suite provides the essential grounding needed for EgoForge to reliably transfer visual cues and follow high-level semantic intent, producing coherent and controllable rollouts despite significant real-world variability.

M Broader Impacts

EgoForge aims to advance egocentric world simulation from sparse wearable observations, enabling controllable first-person video rollouts that can support future research in XR, assistive AI, human–AI collaboration, and embodied learning. Potential positive impacts include safer and lower-cost prototyping of interactive AR/VR systems, synthetic data generation for egocentric perception, and simulation-based evaluation of task-oriented agents without requiring repeated real-world data



Figure 9: **DigiLens ARGO Smart glasses**

collection. At the same time, realistic egocentric video generation may introduce risks if misused, such as creating misleading first-person footage, simulating private environments, or enabling surveillance-like applications when combined with wearable sensors. The use of egocentric data also raises privacy considerations, since first-person views may contain bystanders, personal spaces, or sensitive objects. We therefore view EgoForge as a research tool for simulation and evaluation rather than for identification, surveillance, or decision-making about individuals. Responsible deployment should include consented and licensed data sources, privacy filtering or de-identification when appropriate, disclosure or watermarking of synthetic videos, restrictions against impersonation or surveillance use cases, and human oversight in safety-critical applications.

N Evaluation and Annotation Prompts

Prompt for GRAFT Reward

You are an expert evaluator assessing goal achievement in egocentric video generation. Your task is to evaluate whether a generated first-person video successfully accomplishes a specified manipulation task by analyzing three dimensions: Goal Alignment, Environment Preservation, and Temporal Causality. You will evaluate a process video alongside its initial state image, target state image, and task description, providing quantitative scores for each dimension based on the criteria specified below.

The first task: Goal Alignment Scoring (Total 0–2.0 points)
Assess goal achievement through task completion analysis and semantic alignment between the video final state and the target state:

- **Task Completion Score:** Assign scores from 0.0 to 1.0 based on state transition analysis from initial to final frames, where 0.7∼1.0 indicates all or most required state changes achieved, 0.4∼0.6 indicates partial achievement with some changes incomplete, and 0.0∼0.3 indicates minimal or no required changes with task not accomplished.
- **Visual Alignment Score:** Assign scores from 0.0 to 1.0 based on semantic correspondence between video final state and target image (accounting for first-person vs third-person viewpoint difference), where 0.7∼1.0 indicates all or most target elements and spatial relationships present, 0.4∼0.6 indicates some elements present with others missing, and 0.0∼0.3 indicates minimal or no target elements present.

Provide a detailed breakdown for task completion and visual alignment.
You need to give the score with the following format:
`{"score": [task completion score, visual alignment score]}`

The second task: Environment Preservation Scoring (Total 0–2.0 points)
Assess environment consistency with initial state and resistance to target image contamination:

- **Consistency Score:** Assign scores from 0.0 to 1.0 based on environment consistency between initial and video states (ignoring target image environment which may differ), where 0.7∼1.0 indicates all or most environmental features unchanged, 0.4∼0.6 indicates some features consistent with moderate drift, and 0.0∼0.3 indicates most features changed with major inconsistencies.
- **Contamination Score:** Assign scores from 0.0 to 1.0 based on freedom from target image environmental contamination, where 0.7∼1.0 indicates no or minimal contamination with video

Prompt for GRAFT Reward (continued)

maintaining original environment, 0.4∼0.6 indicates moderate contamination with some target features leaked, and 0.0∼0.3 indicates heavy contamination with video heavily influenced by target's different environment.

Provide a detailed breakdown for environmental consistency assessment.
You need to give the score with the following format:
`{"score": [consistency score, contamination score]}`

The third task: Temporal Causality Scoring (Total 0–2.0 points)
Assess whether the video demonstrates physically plausible and causally coherent progression from initial to final state:

- **Physics Plausibility Score:** Assign scores from 0.0 to 1.0 based on adherence to real-world physics, where 0.7∼1.0 indicates all or most actions are physically realistic with proper gravity and motion, 0.4∼0.6 indicates some actions are realistic with several implausibilities, and 0.0∼0.3 indicates most actions violate physics with impossible movements.
- **Causal Logic Score:** Assign scores from 0.0 to 1.0 based on clarity of cause-effect relationships, where 0.7∼1.0 indicates all or most actions have clear visible causes with logical sequences, 0.4∼0.6 indicates some actions have visible causes with a partial causal chain, and 0.0∼0.3 indicates most actions lack visible causes with broken logic.

Provide a detailed breakdown for each dimension.
You need to give the score with the following format:
`{"score": [physics plausibility score, causal logic score]}`

Caption Refinement Prompt. The model takes both the original coarse caption and the video frames as input to generate a visually grounded fine-grained description.

You are an expert visual data annotator for ego-centric video datasets. Your task is to rewrite and refine a video text annotation by synthesizing the provided **Original Caption** with the actual **Video Content**.
Your Output Goal: A vivid, photorealistic description (3–4 sentences) that strictly aligns with the visual evidence in the video while maintaining the semantics of the original caption.
Instructions:

1. **Visual Grounding**: Analyze the `Video Input` first. Correct the `Original Caption` if it contradicts the visual evidence (e.g., wrong object color, wrong hand usage).
2. **Structure**:
   - **Sentence 1 (Setup)**: Describe the visible appearance of the hands and objects (e.g., material, texture) based on the video start frame.
   - **Sentence 2 (Action)**: Describe the fine-grained motion trajectory observed in the video clips.
   - **Sentence 3 (Reaction)**: Describe the object's physical response (deformation, displacement) shown in the video.
   - **Sentence 4 (Outcome)**: Conclude with the final state visible in the end frame.
3. **Constraints**: Do not hallucinate details not present in the video.

Example Input:

- **Original Caption**: "The person pours water."
- **Video Input**: [Sequence of frames showing a hand tilting a blue cup]

Example Output: "The person's hand, gripping a ribbed blue plastic cup, tilts it sharply over a white bowl. A clear liquid cascades from the cup... [omitted for brevity] ...leaving droplets on the rim."

Input Data:

Original Caption: `{input_caption}`

Video Input: `<video_frames>`

Video Segmentation Prompt for action localization

You are an expert temporal-action summarizer. You are given non-overlapping 10-second segments derived from a long video. Each segment includes:

- start and end: segment time in seconds (absolute timestamps from the video).
- raw: concatenated human annotations overlapping that 10-second window.
- filtered: the same annotations with any sentences containing disallowed categories removed, to highlight stationary/hand-object actions.

Your task:

1. Select 3 to 4 distinct stationary atomic actions across the timeline.
2. "Stationary" means the action is performed in place (hand/object manipulations are preferred).
3. The set of allowed stationary actions is NOT fixed. Infer candidate actions from the provided text (e.g., *shuffle, deal, draw, put, take, hold, open, close, align, sort, count, stack, fan, flip/turn over, organize, adjust, press, push, pull, slide, rotate an object, etc.*).
4. The following categories are strictly disallowed for selection as the main action: `{disallowed_list}`.
5. Avoid describing locomotion (walking/moving/stepping/running), speech (talking/speaking/chatting), or waiting/idle states.
6. Prefer segments where a single hand-on-object action clearly dominates the 10-second window. Use the "`filtered`" field primarily to judge this. If "`filtered`" is empty for a segment, skip it.
7. Each chosen output must represent a different action type (no duplicates). If the same action appears in multiple segments, choose only one representative segment.
8. For each selected segment, write exactly one concise English sentence that describes only the atomic action (single main verb phrase). Do not include walking, talking, or waiting. Avoid chaining multiple verbs. Keep it specific and objective (no speculation).
9. IMPORTANT: Use "the person" as the subject in all descriptions. Do NOT use "she", "he", "C", or any other pronouns or names.

Formatting requirements:

- Output ONLY a JSON array (no other text).
- Each array element MUST be an object with keys: "start", "end", "text".
- "start" and "end" MUST copy the segment's start and end values (integer timestamps in seconds).
- "text" MUST be a single English sentence describing only the stationary atomic action.
- Return exactly 5 to 8 objects. If fewer than 5 valid stationary actions exist, return as many as valid (possibly an empty array).

Segments JSON: `{segments_json}`