Seeing the World through Any Eyes

Yang Fu[1], **Jianqin Wang**[1], **Xiangtai Li**[2], **Henghui Ding**[1]*
[1]Institute of Big Data, College of Computer Science and Artificial Intelligence,
Fudan University, China
[2]ByteDance
https://henghuiding.com/EgoEye/



Figure 1: Examples generated by **EgoEye**. Given a single exocentric video, **EgoEye** synthesizes realistic egocentric views from the target's perspective, enabling immersive exploration of the scene.

Abstract

Egocentric video generation aims to synthesize first-person visual experiences, enabling applications in filmmaking, virtual reality, and embodied AI. Generating egocentric videos from exocentric observations is particularly challenging, as it requires reasoning across large viewpoint changes, limited visual overlap, and substantially different camera motions. To address these challenges, we propose ***EgoEye***, a framework that generates egocentric videos from a single exocentric input video. EgoEye integrates reward-guided egocentric context reasoning to infer unseen first-person content, multi-stage motion alignment to enforce cross-view temporal consistency, and egocentric pretraining to improve first-person realism. To support training and evaluation under diverse and distortion-free exo-ego settings, we further introduce ***EgoScape***, a large-scale dataset comprising 19.5K time-aligned exo-ego video pairs and 24.2K in-the-wild egocentric videos. EgoScape provides paired cross-view supervision while enriching the first-person visual priors needed for realistic egocentric synthesis. Extensive experiments

*Corresponding author (henghui.ding@gmail.com).

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

demonstrate that the proposed EgoEye generates realistic and temporally consistent egocentric videos across diverse scenarios.

1 Introduction

Imagine experiencing the world from perspectives beyond your own, such as seeing through the eyes of a boxer or working from a repairman's viewpoint, as shown in Figure 1. Egocentric video generation enables such immersive experiences by synthesizing realistic first-person videos from exocentric observations, with broad applications in filmmaking [36], virtual reality [28], and world modeling [44]. However, generating egocentric videos from exocentric inputs is highly challenging, as it requires reasoning across large viewpoint changes, limited cross-view visual overlap, and substantially different camera motions. Faithful generation must not only infer first-person content that is unobserved in the exocentric view, but also preserve spatio-temporal consistency across views. A natural solution is to adapt existing video generation paradigms, such as video translation [18, 48, 53, 54], image-conditioned generation [14, 46, 49], and camera-controlled generation [2, 3, 38]. However, these paradigms are not directly suited to exo-to-ego generation. Video translation methods tend to preserve exocentric visual characteristics and struggle with large viewpoint shifts. Image-conditioned methods often require additional egocentric inputs and lack explicit modeling of cross-view transformations. Camera-controlled methods rely on accurate camera trajectories, which are difficult to obtain in real-world scenarios.

To mitigate these challenges, existing methods introduce additional constraints or inputs, e.g., auxiliary egocentric signals [51], multi-view observations [27, 32], or user-specified camera trajectories [19]. While these strategies improve controllability, they introduce strong assumptions and limit general applicability. Recent work has relaxed some of these requirements, but still depends on fixed camera configurations [43], limiting their ability to generalize to diverse real-world settings.

Beyond method limitations, this direction is further constrained by the lack of large-scale, diverse, and high-quality ego–exo paired datasets. Existing exo-ego datasets are primarily designed for action understanding [24, 25, 37, 39] and are often collected in constrained environments, resulting in limited diversity in scenes, subjects, and activities. Moreover, egocentric videos captured with wearable devices often suffer from geometric distortion and color inconsistency [13], which can hinder both model training and reliable evaluation. Recent datasets such as EgoExo-8K [43] expand the scale of paired data, but still rely on constrained camera setups.

Egocentric View Generation. Unlike most previous methods that rely on additional priors or inputs, we propose ***EgoEye***, a framework for egocentric video generation from a single exocentric input. EgoEye integrates reward-guided egocentric context reasoning to infer unobserved first-person content from exocentric observations, optimizing the reasoning process with mixed egocentric rewards that encourage viewpoint consistency, spatial coherence, and semantic completeness. It further incorporates multi-stage motion alignment to improve cross-view temporal consistency by aligning motion features across transformer layers and stabilizing motion evolution under large viewpoint changes. In addition, we perform egocentric pretraining on diverse first-person videos to enhance first-person priors and improve generation quality.

EgoScape Dataset. To better support the training and evaluation of egocentric video generation, we introduce ***EgoScape***, a large-scale dataset with 19.5K time-aligned exo-ego video pairs and 24.2K in-the-wild egocentric videos. EgoScape provides rich cross-view supervision and first-person priors across diverse scenes, subjects, and camera settings. We further establish three benchmarks: *EgoEval* for controlled evaluation with ground truth, *EgoCross* for cross-dataset generalization, and *EgoWild* for in-the-wild evaluation. To comprehensively assess performance, we adopt a unified evaluation protocol across four perspectives: visual fidelity, egocentric perception quality, viewpoint transformation consistency, and human preference.

Our main contributions are as follows. 1) We propose EgoEye, a framework that integrates reward-guided egocentric context reasoning trained with designed egocentric rewards and multi-stage motion alignment for generation. 2) We introduce EgoScape, a large-scale exo-ego dataset with time-aligned paired videos and diverse egocentric data, along with three benchmarks and a unified evaluation protocol. 3) Extensive experiments show that EgoEye consistently achieves strong performance across all benchmarks, producing realistic and temporally coherent egocentric videos.

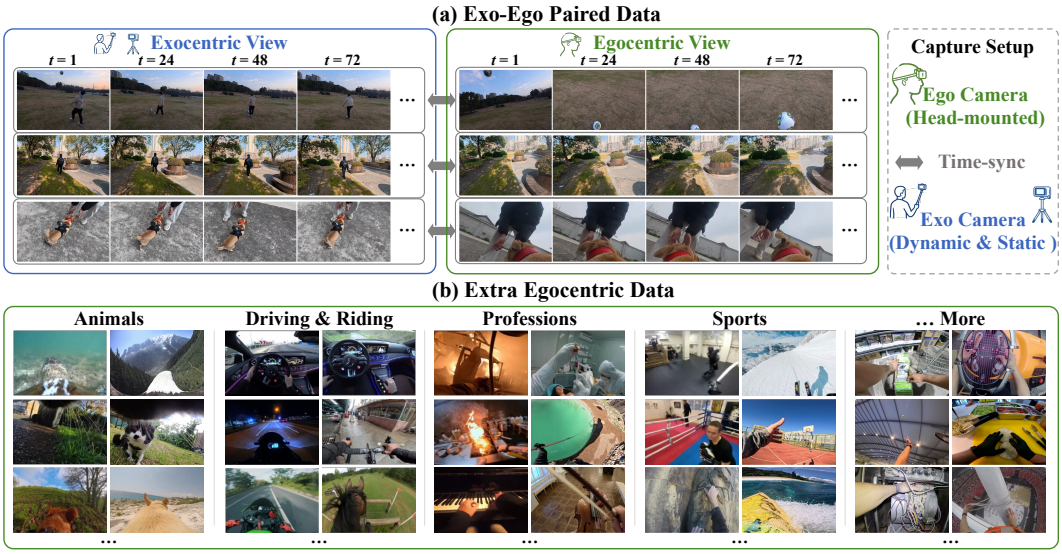

Figure 2: **Overview of the EgoScape Dataset.** The dataset consists of (1) high-quality, time-synced exo–ego paired data with both dynamic and static camera setups, covering diverse scenes and subjects, and (2) large-scale in-the-wild egocentric videos spanning a wide range of real-world scenarios.

2 Related Work

Controlled Video Generation. Cross-view generation can be naturally formulated as controlled video generation, where existing methods mainly rely on video editing, image guidance, or explicit camera control. Video editing methods [18, 22, 48] perform cross-view transformation within generative editing frameworks via instruction, but tend to preserve exocentric characteristics. Image-to-video methods [14, 46, 49] generate videos from a given first frame, but rely on additional egocentric visual inputs and lack explicit modeling of viewpoint transformation. Camera-controlled generation methods [2, 3, 15, 23, 38] inject camera trajectories or geometric cues to guide viewpoint and motion, but depend on accurate camera trajectories that are difficult to obtain.

Exo–Ego View Generation. Prior work on egocentric view generation often introduces additional priors to facilitate cross-view mapping from exocentric inputs. Early approaches [7, 10, 26, 33] focus on image-level synthesis, leveraging geometric cues, camera poses, or 3D representations to generate egocentric images. Recent works extend this paradigm to video generation [19, 27, 29, 30, 32, 43, 51], incorporating multi-view observations, motion cues, or interaction priors to improve controllability. Despite this progress, most methods rely on additional priors, underexploit visual cues from exocentric inputs, and lack explicit cross-view temporal consistency modeling. In parallel, existing exo-ego paired datasets [24, 25, 37, 39] are primarily designed for action understanding, leading to limited scene diversity. Although Ego-Exo4D [13] expands scene coverage, device inconsistency and fisheye distortion degrade cross-view alignment. Recent datasets such as EgoExo-8K [43] adopt specialized capture setups, but remain restricted in viewpoint diversity and introduce geometric distortions.

3 EgoScape Dataset Construction

Overview. As shown in Figure 2, EgoScape is a hybrid dataset comprising two complementary components that play distinct roles in learning. Exo–Ego Paired Data provides temporally aligned cross-view supervision for learning exocentric-to-egocentric transformations, while Extra Egocentric Data introduces diverse in-the-wild first-person observations to further enrich egocentric priors.

3.1 Exo–Ego Paired Data.

(1) Consistent Recording without Distortion. Egocentric and exocentric videos are captured using a unified recording setup to ensure high-fidelity cross-view consistency. All views are recorded using the same device (DJI Action 6) with identical camera parameters and a standard lens, eliminating

Table 1: **Comparison of existing egocentric datasets.** EgoScape is distinguished by high-resolution, distortion-free, time-aligned pairs with dynamic camera setups, while covering diverse subjects.

Dataset	View Pair	Ego Size	Ego Device	Camera Mode	Distortion Free	Consistent Device	Subject Type	Ego Hour	Paired Hour
Charades-Ego [41]	✗	1080P	Phone	Fix.	✓	✗	Pers.	34.7	–
H2O [24]	✓	720P	Azure Kinect	Fix.	✓	✓	Pers.	–	5.3
Ego4D [12]	✗	1080P	GoPro	Fix.	✓	✗	Pers.	3670.0	–
Assembly101 [39]	✓	640 × 480	Monochrome	Fix.	✗	✗	Pers.	–	42.8
EgoBody [56]	✓	1080P	HoloLens2	Fix.	✓	✗	Pers.	–	2.5
EgoHumans [20]	✓	1408 × 1408	Meta Aria	Fix.	✗	✗	Pers.	–	1.2
EgoExo-Fitness [25]	✓	456 × 256	GoPro	Fix.	✓	✗	Pers.	–	32.0
EgoExoLearn [17]	✗	1088 × 1080	Pupil Invisible	Fix.	✓	✗	Pers.	96.5	–
Ego-Exo4D [13]	✓	1408 × 1408	Meta Aria	Fix.	✗	✗	Pers.	–	1286.0
EgoPet [6]	✗	480P	Unknown	Fix.	✓	✗	Anim.	84.0	–
Nymeria [31]	✗	1408 × 1408	Meta Aria	Fix.	✗	✗	Pers.	300.0	–
HD-EPIC [35]	✗	1408 × 1408	Meta Aria	Fix.	✗	✗	Pers.	41.3	–
EgoLife [52]	✓	1408 × 1408	Meta Aria	Fix.	✗	✗	Pers.	–	266.0
HOT3D [5]	✗	1408 × 1408	Meta Aria	Fix.	✗	✗	Pers.	13.9	–
EgoExo-8K [43]	✓	1080P	DJI Action	Fix.	✗	✗	Pers.	–	Unknown
EgoScape (Ours)	✓	1080P	DJI Action	Fix. & Dyn.	✓	✓	Pers. & Anim.	33.6	27.1

color discrepancies and minimizing lens-induced distortion. Synchronized capture is achieved via a GPS-enabled Bluetooth remote controller, providing precise temporal alignment across views. **(2) Dynamic Camera Mode.** Unlike prior datasets relying on fixed exocentric viewpoints [13, 43], our dataset supports both static and dynamic camera settings. In addition to tripod-based recordings, handheld cameras operated by camera operators actively follow subjects from multiple viewpoints, capturing realistic camera motion patterns commonly observed in real-world scenarios. **(3) Diverse Subjects.** Beyond human subjects, we include animal-centered video pairs in which animals are equipped with wearable cameras. **(4) Diverse Lighting Conditions.** Data are collected throughout the day under varying lighting conditions, enhancing environmental diversity.

3.2 Extra Egocentric Data.

To enrich egocentric priors beyond controlled paired settings, we collect a large-scale dataset of in-the-wild egocentric videos with diverse, structured scenario coverage. As shown in Figure 2, the data spans multiple categories, including animals, driving, professions, sports, and activities involving complex hand-object interactions. These categories capture a wide range of first-person experiences, from human activities (e.g., professional tasks and sports) to non-human perspectives (e.g., animals), as well as dynamic camera motions in driving and riding scenarios. Such diversity introduces varied motion patterns, interaction types, and scene layouts that are difficult to obtain in controlled environments, improving the robustness and generalization of egocentric modeling.

3.3 Dataset Preprocessing and Statistics.

We annotate the appearance of the camera wearer as subject prompts to enable precise egocentric grounding and controllable generation. All captured videos are temporally segmented into 5-second fixed-length clips at 30 FPS. To facilitate downstream tasks, we provide clip-level captions generated by Qwen3-VL [4], capturing both scene content and egocentric motion dynamics.

Compared with existing datasets in Table 1, EgoScape uniquely offers high-quality, time-synchronized exo–ego pairs captured with consistent devices and distortion-free imaging, while supporting both static and dynamic camera modes. It further covers both human and animal subjects, resulting in a comprehensive dataset with 33.6 hours of egocentric videos and 27.1 hours of paired recordings. Detailed comparisons and analyses of prior dataset limitations are provided in Supp. Section B.1.

4 EgoEye

Overview. As shown in Figure 3, the model encodes paired inputs into latent representations using a pretrained VAE [21], yielding $\mathbf{x}_{exo}$ and $\mathbf{x}_{ego}$. Gaussian noise is applied to $\mathbf{x}_{ego}$, which is concatenated with $\mathbf{x}_{exo}$ to form $\mathbf{x}_{in}$. Shared positional embeddings are applied with $\mathcal{P}_{ego}$ derived

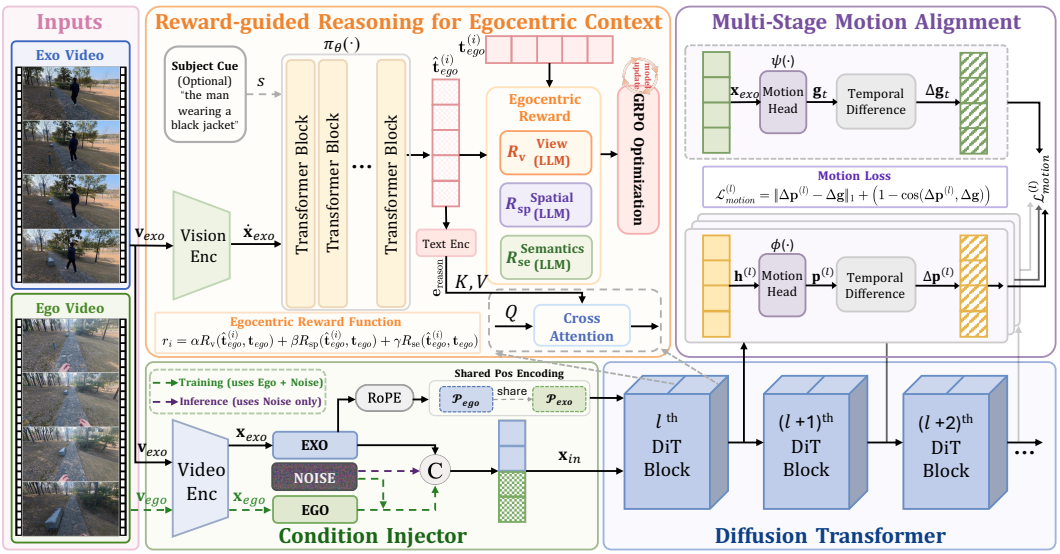

Figure 3: **The framework of EgoEye.** During training, we concatenate the noised egocentric latent with exocentric video features and encode them with shared positional embeddings, followed by a diffusion transformer. To ensure motion consistency, multi-stage motion alignment is applied across DiT layers to explicitly supervise temporal dynamics. Furthermore, a reward-guided egocentric context reasoning module infers first-person context from exocentric input and is optimized with mixed reward signals. At inference, the model generates egocentric video from exocentric video only.

from $\mathcal{P}_{exo}$, enabling consistent temporal encoding. The resulting tokens $\mathbf{x}_{in}$ are then processed by a diffusion transformer [34] for joint spatio-temporal modeling. On this backbone, our EgoEye incorporates two key components: (1) Reward-guided Reasoning for Egocentric Context, which infers first-person context from exocentric inputs and is optimized with mixed reward signals; and (2) Multi-Stage Motion Alignment, which aligns motion representations across DiT layers to ensure temporal consistency.

4.1 Reward-guided Reasoning for Egocentric Context

Exo-to-ego generation requires reasoning about the egocentric context that is not directly observable from the refered view, including spatial layout and scene content. Given an exocentric video $\mathbf{v}_{exo}$ and a subject prompt s, we model this process with a reasoning policy π_θ, which generates an egocentric description $\hat{\mathbf{t}}_{ego}$ as high-level guidance for video generation. To supervise this reasoning process, we derive the ground-truth egocentric caption $\mathbf{t}_{ego}$ from the paired egocentric video and compare it with $\hat{\mathbf{t}}_{ego}$ using an LLM-based judge. The resulting reward evaluates viewpoint consistency, spatial coherence, and semantic completeness, and is used to optimize the reasoning policy via GRPO [40].

Egocentric Reward Design. We design the reward to address three dominant failure modes in exo-to-ego generation: (i) viewpoint inconsistency, (ii) incorrect spatial relations, and (iii) loss of scene-specific semantics. Specifically, R_{v} evaluates whether the generated description follows a consistent first-person viewpoint, avoiding third-person bias and preserving egocentric perspective; R_{sp} enforces correct spatial relations between the subject and surrounding objects, particularly left-right consistency under viewpoint transformation; and R_{se} measures semantic completeness, encouraging accurate descriptions of object attributes such as appearance, color, and shape, as well as interactions. The total reward for the i-th sampled description is defined as:

$$r_i = \alpha R_{\text{v}}(\hat{\mathbf{t}}^{(i)}_{ego}, \mathbf{t}_{ego}) + \beta R_{\text{sp}}(\hat{\mathbf{t}}^{(i)}_{ego}, \mathbf{t}_{ego}) + \gamma R_{\text{se}}(\hat{\mathbf{t}}^{(i)}_{ego}, \mathbf{t}_{ego}), \tag{1}$$

where α, β, and γ are weighting coefficients.

GRPO Optimization. For each input $(\mathbf{v}_{exo}, s)$, the old policy $\pi_{\theta_{\text{old}}}$ samples a group of G candidate descriptions $\{\hat{\mathbf{t}}^{(i)}_{ego}\}_{i=1}^{G}$. We normalize rewards within the group to obtain the relative advantage:

$$A_i = \frac{r_i - \mu_r}{\sigma_r + \epsilon_a}, \tag{2}$$

where $\mu_r = \frac{1}{G}\sum_{i=1}^{G} r_i$ and $\sigma_r = \sqrt{\frac{1}{G}\sum_{i=1}^{G}(r_i - \mu_r)^2}$, and ϵ_a is a small constant for numerical stability. We minimize the GRPO loss:

$$\mathcal{L}_{\mathrm{GRPO}}(\theta) = -\mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\min(\rho_i A_i,\ \mathrm{clip}(\rho_i, 1-\epsilon_c, 1+\epsilon_c)A_i)\right] + \lambda D_{\mathrm{KL}}(\pi_\theta \,\|\, \pi_{\mathrm{ref}})\,, \quad (3)$$

where

$$\rho_i = \frac{\pi_\theta\left(\hat{\mathbf{t}}_{ego}^{(i)} \mid \mathbf{v}_{exo}, s\right)}{\pi_{\theta_{\mathrm{old}}}\left(\hat{\mathbf{t}}_{ego}^{(i)} \mid \mathbf{v}_{exo}, s\right)}. \quad (4)$$

Here, π_θ is the current policy being optimized, $\pi_{\theta_{\mathrm{old}}}$ is the behavior policy used to sample candidate descriptions, and π_{ref} is a frozen reference policy for KL regularization. ρ_i is the sequence-level likelihood ratio, ϵ_c is the clipping threshold, and λ controls the KL divergence penalty strength.

4.2 Multi-Stage Motion Alignment.

Maintaining temporally consistent motion under large viewpoint transformations is challenging, as motion cues in exocentric videos are often incomplete or ambiguous for egocentric perception. While diffusion models capture temporal dynamics through denoising, they lack explicit constraints on motion evolution, often leading to unstable or drifting patterns.

To address this issue, we propose a multi-stage motion alignment that explicitly supervises motion representations across multiple layers of the diffusion transformer. We extract intermediate token features from different layers and align their temporal dynamics with motion derived from egocentric latents. Formally, given token features $\mathbf{h}^{(l)} \in \mathbb{R}^{T\times D}$ at layer l and egocentric latent features $\mathbf{x}_{ego} \in \mathbb{R}^{T\times C}$, we project them into a shared motion embedding space:

$$\mathbf{p}_t^{(l)} = \phi(\mathbf{h}_t^{(l)}), \quad \mathbf{g}_t = \psi(\mathbf{x}_{ego,t}), \quad (5)$$

where $\phi(\cdot)$ and $\psi(\cdot)$ are learnable heads. We define motion as temporal differences between frames:

$$\Delta\mathbf{p}_t^{(l)} = \mathbf{p}_{t+1}^{(l)} - \mathbf{p}_t^{(l)}, \quad \Delta\mathbf{g}_t = \mathbf{g}_{t+1} - \mathbf{g}_t. \quad (6)$$

We then align both motion magnitude and direction using:

$$\mathcal{L}_{motion}^{(l)} = \|\Delta\mathbf{p}^{(l)} - \Delta\mathbf{g}\|_1 + \left(1 - \cos(\Delta\mathbf{p}^{(l)}, \Delta\mathbf{g})\right), \quad (7)$$

The final loss is obtained by averaging $\mathcal{L}_{motion}^{(l)}$ across multiple transformer layers.

4.3 Egocentric Pretraining.

Most existing video generation models are trained on generic video corpora and lack explicit exposure to egocentric data, limiting their ability to model first-person camera characteristics and object distributions. Consequently, they often fail to produce realistic egocentric views and may exhibit distorted perspectives (e.g., fisheye-like artifacts), as shown in Figure 4. To address this limitation, we introduce an egocentric pretraining stage using large-scale egocentric videos from EgoScape. Specifically, we perform text-to-video pretraining to inject egocentric priors into the model, enabling it to better capture first-person camera properties, scene layouts, and interaction patterns, and to improve alignment between textual instructions and generated outputs.

5 Experiments

5.1 Experimental Settings

Evaluation Benchmarks. To enable comprehensive evaluation, we construct three benchmarks: **(1) EgoEval** is built from 50 paired videos sampled from EgoScape for controlled in-domain evaluation with egocentric targets; **(2) EgoCross** is formed by randomly sampling 40 paired videos from existing datasets [24, 56] to assess cross-dataset generalization; and **(3) EgoWild** is constructed from 200 in-the-wild exocentric videos collected to evaluate performance in unconstrained settings.

Table 2: **Quantitative Comparison on EgoEval.** The best result is shown in **bold** and the second-best result is shown in underline.

Method	Common Fidelity Criteria				Egocentric Metrics						User Study
	PSNR ↑	SSIM ↑	LPIPS ↓	FVD ↓	VA ↑	IP ↑	PC ↑	Mean ↑	S_g ↑	S_{gx} ↑	User ↑
+ Egocentric Camere Trajectories											
ReCameraMaster [3]	10.801	0.273	0.645	1793.415	0.066	0.061	0.060	0.062	0.447	0.264	1.8
EgoX [19]	9.239	0.217	0.738	2284.654	0.482	0.431	0.367	0.427	0.468	0.590	4.9
+ Egocentric First Image											
Wan2.2 [46]	9.697	0.236	0.645	1717.161	0.468	0.432	0.350	0.417	0.504	0.443	5.1
LTX-2 [14]	10.675	0.280	0.616	1105.611	0.714	0.653	0.659	0.675	0.660	0.576	7.9
Hunyuan [49]	10.230	0.231	0.642	1448.783	0.576	0.549	0.504	0.543	0.540	0.466	6.2
+ None											
VACE [18]	10.556	0.276	0.648	1919.772	0.085	0.093	0.153	0.110	0.462	0.259	1.1
UniVideo [48]	10.608	0.253	0.641	2006.733	0.084	0.075	0.069	0.076	0.460	0.273	1.4
Helios [54]	7.473	0.297	0.713	1693.351	0.684	0.632	0.532	0.616	0.460	0.557	6.5
CogVideoX [53]	10.873	0.290	0.652	1654.489	0.202	0.174	0.182	0.186	0.506	0.513	2.6
EgoEye (Ours)	**13.046**	**0.356**	**0.504**	**917.326**	**0.884**	**0.769**	**0.767**	**0.807**	**0.821**	**0.682**	**9.2**

Evaluation Metrics. Egocentric videos are highly sensitive to camera placement and intrinsic parameters, where even slight variations can lead to substantial pixel-level mismatches between generated and ground-truth videos. As a result, conventional fidelity metrics alone are insufficient for reliable evaluation. We therefore evaluate generation quality from four complementary perspectives: **(1) Common Fidelity Criteria.** We report PSNR [16], SSIM [47], LPIPS [55], and FVD [45]. These metrics capture reconstruction quality and temporal coherence, but do not reflect egocentric realism under viewpoint shifts. **(2) Egocentric Perception Quality.** We evaluate first-person plausibility using *Viewpoint Authenticity* (VA), *Interaction Plausibility* (IP), and *Prompt Consistency* (PC), assessing viewpoint realism, physical interactions, and semantic alignment with the intended description. We also report their average and implement these metrics using a structured perception model [4]. **(3) Viewpoint Transformation Consistency.** We measure viewpoint transformation in a visual feature space, as existing methods [18, 48] often retain exocentric characteristics. Specifically, we extract frame-level features $\{\mathbf{f}_t\}_{t=1}^{T}$ using DINOv3 [42] and compute:

$$\mathbf{f} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{f}_t, \quad \tilde{\mathbf{f}} = \frac{\mathbf{f}}{\|\mathbf{f}\|_2}. \tag{8}$$

$$S_g = \cos(\tilde{\mathbf{f}}_{pred}, \tilde{\mathbf{f}}_{ego}), \quad S_{gx} = \frac{\cos(\tilde{\mathbf{f}}_{pred}, \tilde{\mathbf{f}}_{ego}) - \cos(\tilde{\mathbf{f}}_{pred}, \tilde{\mathbf{f}}_{exo}) + 1}{2}. \tag{9}$$

S_g measures alignment with the egocentric target, while S_{gx} evaluates whether the generated video moves away from the exocentric input and toward the egocentric target. **(4) Human Evaluation.** We conduct a user study with 20 annotators rating VA, IP, and PC on a 0–10 scale and report the average.

Implementation Details. During training, we resize all paired exo-ego videos to a resolution of 1280×704 and sample 49 frames with a temporal stride of 2. We adopt Wan2.2 [46] as the backbone model and train it with LoRA adaptation of rank 80. The model is optimized for $6{,}000$ steps with a batch size of 4 on eight H100 GPUs, using a learning rate of 5×10^{-5}. In the first stage, we pretrain the model on text-video paired data. We then initialize from the pretrained LoRA weights and continue training for video-to-video generation under the same configuration. The detailed experimental configurations and baseline reproductions are in the supplement Section E.

5.2 Comparison with the State-of-the-Art Methods

Quantitative Evaluation. We compare EgoEye with representative video generation methods under three settings on EgoEval and EgoCross, with results reported in Table 2 and Table 3 for both in-domain and out-of-domain evaluation: (1) methods requiring egocentric camera trajectories (ReCameraMaster [3], EgoX [19]), (2) methods conditioned on the first-person image(Wan2.2 [46], Huanyuan [49], LTX-2 [14]), and (3) recent video generation or editing models without egocentric

Table 3: **Quantitative Comparison on EgoCross.** The best result is shown in **bold** and the second-best result is shown in underline.

Method	Common Fidelity Criteria				Egocentric Metrics						User Study
	PSNR ↑	SSIM ↑	LPIPS ↓	FVD ↓	VA ↑	IP ↑	PC ↑	Mean ↑	S_g ↑	S_{gx} ↑	User ↑
+ Egocentric Camera Trajectories											
ReCameraMaster [3]	9.496	0.241	0.670	1970.137	0.125	0.138	0.218	0.160	0.524	0.300	2.8
EgoX [19]	8.026	0.209	0.735	2362.537	0.533	0.431	0.380	0.448	0.398	0.509	7.4
+ Egocentric First Image											
Wan2.2 [46]	9.704	0.237	0.675	1851.583	0.463	0.446	0.398	0.435	0.486	0.407	7.1
LTX-2 [14]	8.812	0.217	0.679	2000.426	0.463	0.435	0.415	0.438	0.539	0.422	7.2
Hunyuan [49]	8.868	0.201	0.670	2099.204	0.438	0.421	0.398	0.419	0.467	0.391	6.8
+ None											
VACE [18]	9.539	0.234	0.675	2022.343	0.085	0.093	0.153	0.110	0.503	0.280	1.9
UniVideo [48]	9.387	0.223	0.669	2228.679	0.133	0.139	0.200	0.157	0.523	0.289	2.6
Helios [54]	7.878	0.268	0.703	1909.696	0.275	0.233	0.255	0.254	0.460	0.436	4.9
CogVideoX [53]	9.469	0.242	0.676	2023.574	0.103	0.104	0.110	0.105	0.448	0.400	1.7
EgoEye (Ours)	**10.014**	**0.281**	**0.668**	**1795.866**	**0.640**	**0.531**	**0.480**	**0.550**	**0.568**	**0.509**	**8.5**

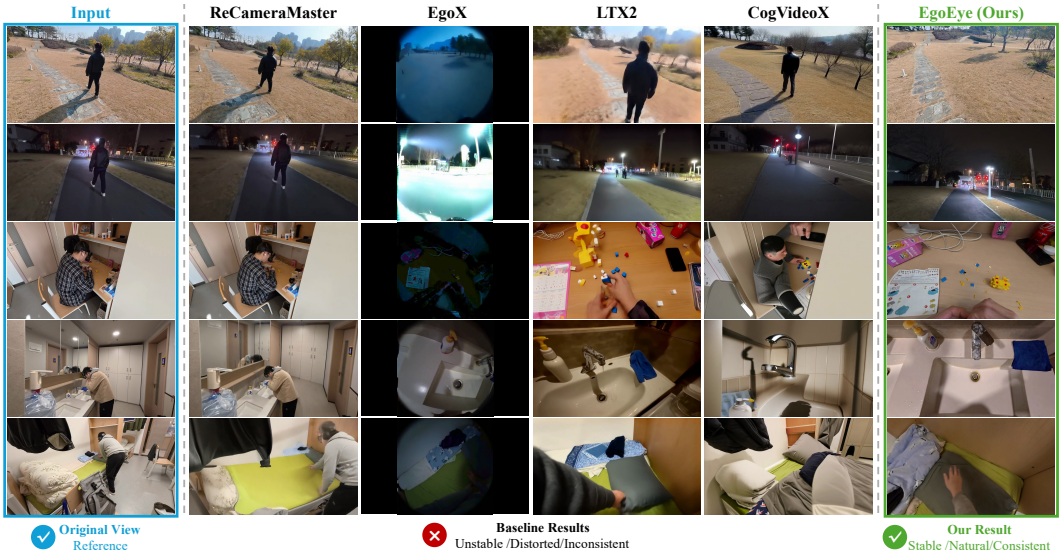

Figure 4: **Qualitative results on EgoEval.** Trajectory-based methods [3, 19] often produce unstable and inconsistent egocentric views, with EgoX [19] showing noticeable geometric distortion. Image-to-video models [14] suffer from motion inconsistency, while video-to-video models [53] tend to generate unrealistic viewpoints. EgoEye generates stable, natural, and consistent egocentric videos.

priors(VACE [18], UniVideo [48], Helios [54], CogVideoX [53]). Trajectory-based methods improve viewpoint transformation but fail to produce realistic egocentric perception. First-image-conditioned methods achieve better perceptual quality but rely on additional egocentric inputs. In contrast, prior-free methods tend to preserve exocentric characteristics and struggle to perform effective viewpoint transformation. EgoEye consistently outperforms all baselines across both in-domain and out-of-domain settings, achieving strong egocentric perception and accurate viewpoint transformation while maintaining competitive fidelity without requiring additional priors.

Qualitative Evaluation. For qualitative evaluation, we compare our method with representative trajectory-based methods (ReCameraMaster [3], EgoX [19]), image-conditioned models (LTX-2 [14]), and prior-free methods (CogVideoX [53]), as shown in Figure 4 and Figure 5. Trajectory-based methods fail to produce correct egocentric viewpoints, with EgoX exhibiting noticeable geometric distortion, while image-conditioned models suffer from motion inconsistency. Prior-free methods tend to preserve exocentric characteristics and struggle with viewpoint transformation. In contrast, EgoEye generates stable, natural, and consistent egocentric videos across diverse scenarios.

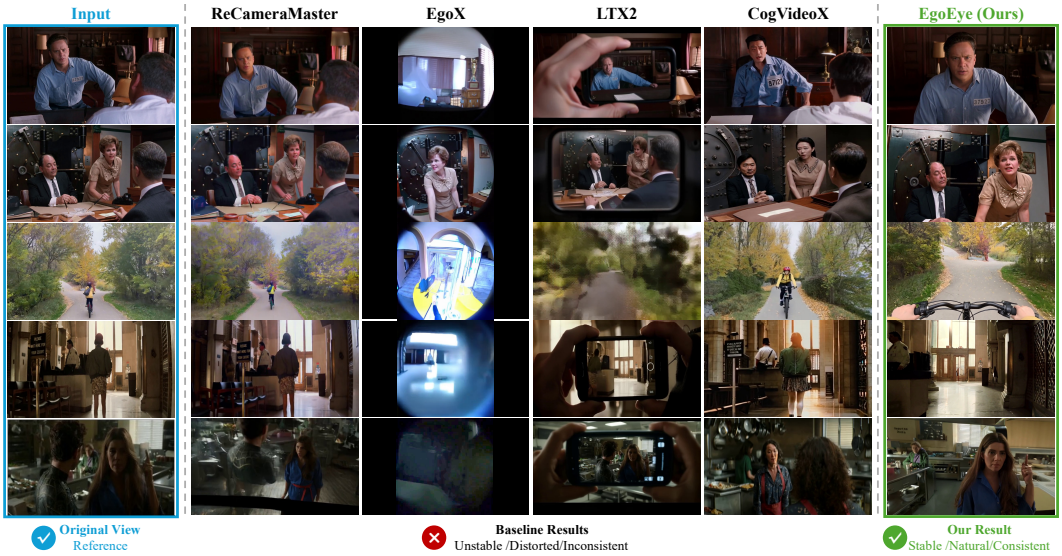

Figure 5: **Qualitative results on EgoWild.** EgoEye remains robust across diverse in-the-wild scenarios, including multi-person interactions, fast-moving sports, and low-light conditions.

Table 4: **Ablation study of EgoEye.** † denotes the WorldWander model [43] trained on the EgoScape dataset. Based on EgoScape, the egocentric video transformation performance consistently improves with the addition of motion alignment, egocentric pretraining, and reasoning modules.

Method	Common Visual Metrics				Egocentric Metrics						User Study
	PSNR↑	SSIM↑	LPIPS↓	FVD↓	VA↑	IP↑	PC↑	Mean↑	S_g↑	S_{gx}↑	User ↑
Base	11.756	0.313	0.590	1157.443	0.828	0.751	0.592	0.724	0.675	0.602	7.4
Base + Motion	12.422	0.344	0.553	1021.748	0.842	0.764	0.627	0.744	0.725	0.645	7.8
Base + Pretrain	12.688	0.341	0.546	954.778	0.856	0.763	0.674	0.764	0.734	0.639	8.1
Base + Reason	13.034	0.354	0.548	936.542	0.873	**0.769**	0.706	0.783	0.813	0.670	8.8
Base + All (EgoEye)	**13.046**	**0.356**	**0.504**	**917.326**	**0.884**	**0.769**	**0.767**	**0.807**	**0.821**	**0.682**	**9.2**
WorldWander† [43]	12.396	0.332	0.557	1066.968	0.882	0.765	0.655	0.767	0.751	0.637	8.0

5.3 Ablation Study

We perform ablation studies to evaluate the effectiveness of our designs on EgoEval, as shown in Table 4, where all variants are trained under the same experimental settings. For the model components, we observe that motion alignment, egocentric pretraining, and reasoning each consistently improve performance over the base model, while their combination achieves the best results. Our base model, built on Wan2.2 [46] and trained on EgoScape, outperforms the original Wan2.2, demonstrating the dataset's effectiveness. Applying the same training setting to WorldWander [43] also yields clear improvements, while EgoEye consistently outperforms it, indicating that the performance gains also stem from the proposed modules.

6 Conclusion

In this work, we study egocentric video generation from exocentric inputs, a challenging setting involving large viewpoint changes, limited visual overlap, and inconsistent camera motion. To address these challenges, we propose ***EgoEye***, a framework that integrates GRPO-based egocentric context reasoning, multi-stage motion alignment, and egocentric pretraining to infer unseen first-person content and improve temporal consistency. To support training and evaluation, we further introduce ***EgoScape***, a large-scale dataset comprising time-aligned exo-ego video pairs for cross-view supervision and diverse in-the-wild egocentric videos to enrich first-person priors. Extensive experiments show that EgoEye generates realistic egocentric videos while maintaining temporal coherence and robustness across diverse scenarios.

Limitations. EgoEye generates videos under a canonical camera setting, and extending it to support user-specified camera parameters (e.g., camera type and focal length) is a direction for future work.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant No. 62472104 and the Science and Technology Commission of Shanghai Municipality under Grant No. 25511103600.

References

[1] Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

[2] Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B Lindell, and Sergey Tulyakov. Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. In *CVPR*, pages 22875–22889, 2025.

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

[4] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[5] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, et al. Hot3d: Hand and object tracking in 3d from egocentric multi-view videos. In *CVPR*, pages 7061–7071, 2025.

[6] Amir Bar, Arya Bakhtiar, Danny Tran, Antonio Loquercio, Jathushan Rajasegaran, Yann LeCun, Amir Globerson, and Trevor Darrell. Egopet: Egomotion and interaction data from an animal's perspective. In *ECCV*, pages 377–394. Springer, 2024.

[7] 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 *ECCV*, pages 409–427. Springer, 2024.

[8] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, Philip HS Torr, and Song Bai. MOSE: A new dataset for video object segmentation in complex scenes. In *ICCV*, pages 20224–20234, 2023.

[9] Henghui Ding, Kaining Ying, Chang Liu, Shuting He, Xudong Jiang, Yu-Gang Jiang, Philip HS Torr, and Song Bai. MOSEv2: A more challenging dataset for video object segmentation in complex scenes. *arXiv preprint arXiv:2508.05630*, 2025.

[10] Mohamed Elfeki, Krishna Regmi, Shervin Ardeshir, and Ali Borji. From third person to first person: Dataset and baselines for synthesis and retrieval. *arXiv preprint arXiv:1812.00104*, 2018.

[11] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, et al. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023.

[12] 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 *CVPR*, pages 18995–19012, 2022.

[13] 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 *CVPR*, pages 19383–19400, 2024.

[14] Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, et al. Ltx-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

[15] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. In *ICLR*, 2025.

[16] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *ICPR*, 2010.

[17] Yifei Huang, Guo Chen, Jilan Xu, Mingfang Zhang, Lijin Yang, Baoqi Pei, Hongjie Zhang, Lu Dong, Yali Wang, Limin Wang, et al. Egoexolearn: A dataset for bridging asynchronous ego-and exo-centric view of procedural activities in real world. In *CVPR*, pages 22072–22086, 2024.

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

[19] Taewoong Kang, Kinam Kim, Dohyeon Kim, Minho Park, Junha Hyung, and Jaegul Choo. Egox: Egocentric video generation from a single exocentric video. In *CVPR*, 2026.

[20] Rawal Khirodkar, Aayush Bansal, Lingni Ma, Richard Newcombe, Minh Vo, and Kris Kitani. Ego-humans: An ego-centric 3d multi-human benchmark. In *ICCV*, pages 19807–19819, 2023.

[21] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[22] Max Ku, Cong Wei, Weiming Ren, Huan Yang, and Wenhu Chen. Anyv2v: A tuning-free framework for any video-to-video editing tasks. *Transactions on Machine Learning Research*, 2024.

[23] Zhengfei Kuang, Shengqu Cai, Hao He, Yinghao Xu, Hongsheng Li, Leonidas J Guibas, and Gordon Wetzstein. Collaborative video diffusion: Consistent multi-video generation with camera control. In *NeurIPS*, pages 16240–16271, 2024.

[24] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *ICCV*, pages 10138–10148, 2021.

[25] Yuan-Ming Li, Wei-Jin Huang, An-Lan Wang, Ling-An Zeng, Jing-Ke Meng, and Wei-Shi Zheng. Egoexo-fitness: Towards egocentric and exocentric full-body action understanding. In *ECCV*, pages 363–382. Springer, 2024.

[26] Gaowen Liu, Hugo Latapie, Ozkan Kilic, and Adam Lawrence. Parallel generative adversarial network for third-person to first-person image generation. In *CVPRW*, pages 1917–1923, 2022.

[27] Jia-Wei Liu, Weijia Mao, Zhongcong Xu, Jussi Keppo, and Mike Zheng Shou. Exocentric-to-egocentric video generation. In *NeurIPS*, pages 136149–136172, 2024.

[28] Tao Lu, Qian Zhu, Tiffany Ma, Wong Kam-Kwai, Anlan Xie, Alex Endert, and Yalong Yang. Ego vs. exo and active vs. passive: Investigating the individual and combined effects of viewpoint and navigation on spatial immersion and understanding in immersive storytelling. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 2025.

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

[30] Mi Luo, Zihui Xue, Alex Dimakis, and Kristen Grauman. Put myself in your shoes: Lifting the egocentric perspective from exocentric videos. In *ECCV*, pages 407–425. Springer, 2024.

[31] 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 *ECCV*, pages 445–465. Springer, 2024.

[32] Mohammad Mahdi, Yuqian Fu, Nedko Savov, Jiancheng Pan, Danda Pani Paudel, and Luc Van Gool. Exo2egosyn: Unlocking foundation video generation models for exocentric-to-egocentric video synthesis. *arXiv preprint arXiv:2511.20186*, 2025.

[33] 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.

[34] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pages 4195–4205, 2023.

[35] Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha, Omar Emara, Sam Pollard, Kranti Kumar Parida, Kaiting Liu, Prajwal Gatti, Siddhant Bansal, Kevin Flanagan, et al. Hd-epic: A highly-detailed egocentric video dataset. In *CVPR*, pages 23901–23913, 2025.

[36] Chiara Plizzari, Gabriele Goletto, Antonino Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Dima Damen, and Tatiana Tommasi. An outlook into the future of egocentric vision. *IJCV*, 132 (11):4880–4936, 2024.

[37] Nishant Rai, Haofeng Chen, Jingwei Ji, Rishi Desai, Kazuki Kozuka, Shun Ishizaka, Ehsan Adeli, and Juan Carlos Niebles. Home action genome: Cooperative compositional action understanding. In *CVPR*, pages 11184–11193, 2021.

[38] Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *CVPR*, pages 6121–6132, 2025.

[39] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *CVPR*, pages 21096–21106, 2022.

[40] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[41] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Actor and observer: Joint modeling of first and third-person videos. In *CVPR*, pages 7396–7404, 2018.

[42] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

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

[44] Yuanpeng Tu, Hao Luo, Xi Chen, Xiang Bai, Fan Wang, and Hengshuang Zhao. Playerone: Egocentric world simulator. In *NeurIPS*, 2025.

[45] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. In *ICLR Workshop*, 2019.

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

[47] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4), 2004.

[48] Cong Wei, Quande Liu, Zixuan Ye, Qiulin Wang, Xintao Wang, Pengfei Wan, Kun Gai, and Wenhu Chen. Univideo: Unified understanding, generation, and editing for videos. In *ICLR*, 2026.

[49] Bing Wu, Chang Zou, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Jack Peng, Jianbing Wu, Jiangfeng Xiong, Jie Jiang, et al. Hunyuanvideo 1.5 technical report. *arXiv preprint arXiv:2511.18870*, 2025.

[50] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

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

[52] Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, et al. Egolife: Towards egocentric life assistant. In *CVPR*, pages 28885–28900, 2025.

[53] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *ICLR*, 2025.

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

[55] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[56] Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. Egobody: Human body shape and motion of interacting people from head-mounted devices. In *ECCV*, pages 180–200. Springer, 2022.

Seeing the World through Any Eyes

Supplementary Material

Outline

Below is the outline of the supplementary material:

- **Section A** presents additional experimental results across all benchmarks.
- **Section B** provides detailed analysis and construction of the EgoScape dataset.
- **Section C** introduces the benchmark protocols, including EgoEval, EgoCross, and EgoWild.
- **Section D** describes the evaluation metrics for assessing egocentric generation quality.
- **Section E.1** presents the implementation details of EgoEye and training strategies.
- **Section E.2** details the implementation of compared baseline methods.
- **Section F** provides runtime and efficiency analysis.
- **Section G** discusses reproducibility and ethical considerations.
- **Section H** presents possible future work.
- **Section I** discusses broader impacts.

A Additional Experimental Results

Figure 6, Figure 7, and Figure 8 present qualitative comparisons with all baseline methods on the three benchmarks, with additional in-the-wild results shown in Figure 9. The results show that EgoEye produces realistic and temporally coherent egocentric videos, demonstrating strong generalization and effective viewpoint transformation. Additional results are available at `https://henghuiding.com/EgoEye/`.

B EgoScape Dataset Details

B.1 Limitations of Existing Datasets

We analyze the limitations of representative egocentric video datasets. As shown in Figure 10, existing datasets fall short of supporting high-quality exocentric-to-egocentric video generation and evaluation.

Detailed weaknesses in prior datasets. Many existing datasets suffer from one or more of the following issues:

(1) **Lack of time-synchronized paired views.** Some datasets provide only egocentric videos without corresponding exocentric observations, such as Ego4D [12] and Nymeria [31]. Others include both views but lack temporal alignment. For instance, Charades-Ego [41] records the two views separately, while EgoExoLearn [17] provides asynchronous pairs that are not strictly aligned.

(2) **Low resolution.** Most existing time-aligned paired datasets are captured at relatively low resolutions. For example, H2O [24] provides 720p videos, Assembly101 [39] provides 640×480, and EgoExo-Fitness [25] provides 456×256. Such low resolution limits high-fidelity exocentric-to-egocentric generation and fine-grained visual evaluation.

(3) **Restricted camera settings.** Exocentric views in many datasets are captured under constrained camera setups. For example, Ego-Exo4D [13] and EgoLife [52] use tripod-mounted cameras, limiting camera motion. EgoExo-8K [43] allows camera movement by attaching the camera behind the subject, but the viewpoint remains restricted to a fixed following angle. Such setups fail to capture the diverse viewpoints present in real-world video.

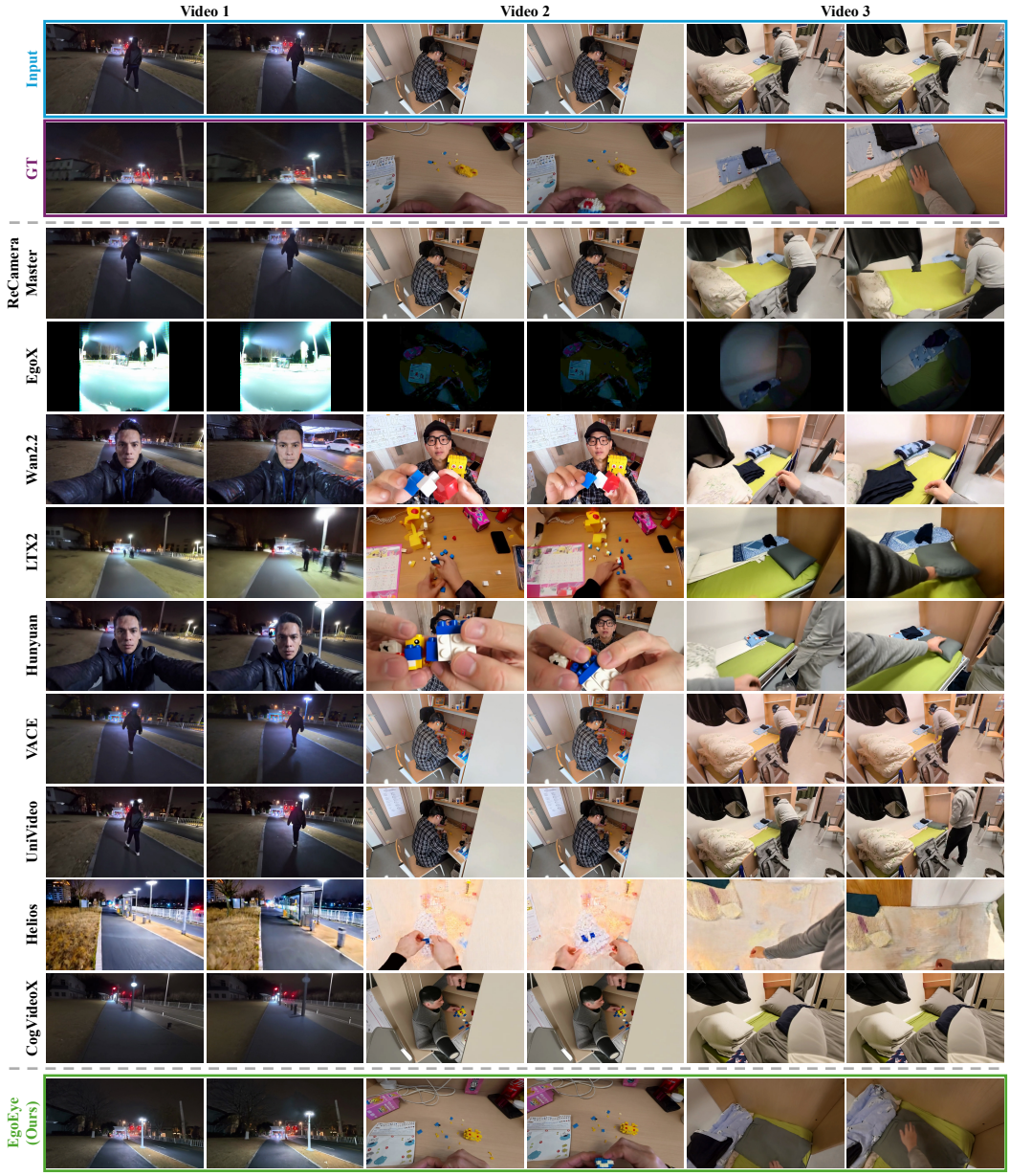

Figure 6: **Qualitative results on EgoEval.** Compared with recent methods, EgoEye generates stable, natural, and consistent egocentric videos.

(4) **Distorted imaging.** Many datasets use fisheye or wide-angle lenses, leading to geometric distortion. For example, Ego-Exo4D [13], Nymeria [31], HD-EPIC [35], and EgoLife [52] rely on Meta Aria devices [11], while EgoExo-8K [43] and EgoPet [6] adopt wide-angle cameras. Such distortion deviates from standard first-person views captured with normal lenses, reducing visual fidelity and immersion in exocentric-to-egocentric generation.

(5) **Inconsistent capture devices.** Most datasets use different devices for egocentric and exocentric views, leading to cross-view appearance inconsistencies. For instance, EgoBody [56] uses HoloLens 2 for egocentric capture and Azure Kinect for exocentric views, while Ego-Exo4D [13], EgoLife [52], and EgoExo-8K [43] adopt different camera devices for egocentric and exocentric capture. This introduces discrepancies in color, field of view, and imaging characteristics, hindering reliable exocentric-to-egocentric transformation.

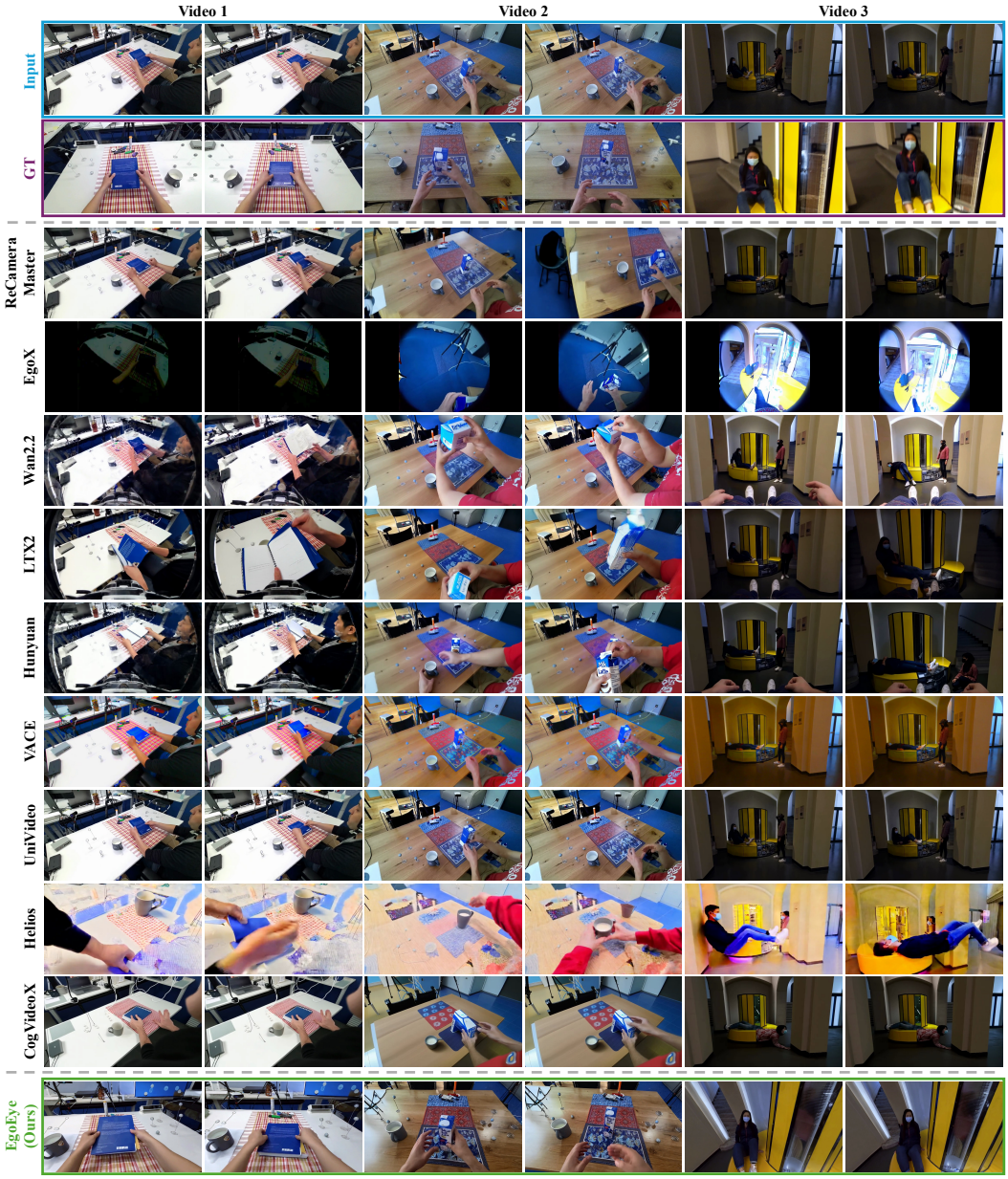

Figure 7: **Qualitative results on EgoCross.** EgoEye generalizes better to out-of-domain scenarios than recent methods, producing realistic and coherent egocentric videos.

(6) **Limited subject diversity.** Most existing datasets focus on daily human activities, with limited coverage of specialized or complex scenarios. For example, Ego-Exo4D [13] and Ego4D [12] primarily capture daily actions, with limited data for activities such as driving, horseback riding, or boxing. EgoPet [6] includes animal-centric data but lacks paired exocentric views. This restricts the generalization of models to diverse subjects and real-world scenarios.

Detailed reports. Below, we highlight representative issues in recent datasets:

- **Charades-Ego (2018)** [41]: Captures both exocentric and egocentric videos of scripted activities. However, the two views are recorded in separate sessions, resulting in a lack of strict temporal alignment.

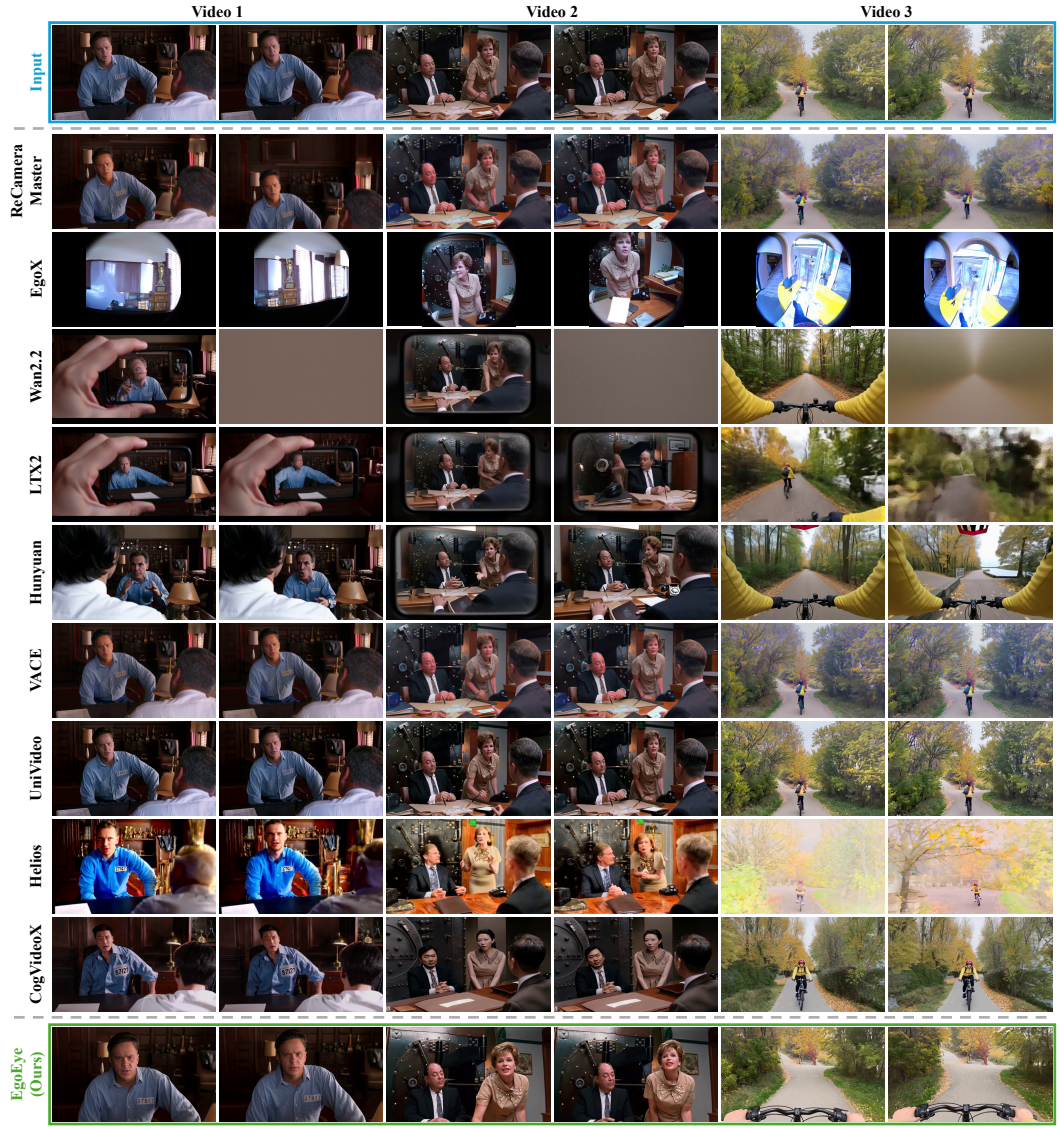

Figure 8: **Qualitative results on EgoWild.** EgoEye produces more realistic and visually competitive results than recent methods across diverse scenarios.

- **H2O (2021)** [24]: Provides synchronized multi-view recordings using Azure Kinect. However, the resolution is limited to 720p, the scenes are restricted to indoor hand-object interactions, and the total duration is only 5.3 hours.
- **Ego4D (2022)** [12]: Collects large-scale egocentric videos using GoPro cameras. However, it does not provide paired exocentric views and primarily focuses on daily human activities.
- **Assembly101 (2022)** [39]: Provides synchronized multi-view recordings for assembly tasks. However, the egocentric views are captured by monochrome cameras at 640×480, while the exocentric views are captured by RGB cameras at 1920×1080, leading to significant cross-view inconsistencies in color, resolution, and imaging characteristics.
- **EgoBody (2022)** [56]: Uses HoloLens 2 for egocentric capture and Azure Kinect for exocentric views. This heterogeneous setup introduces appearance inconsistencies, and the paired data is limited to 2.5 hours.
- **EgoHumans (2023)** [20]: Captures egocentric views with Meta Aria [11] and exocentric views with stationary cameras. The use of heterogeneous devices and fisheye lenses

introduces cross-view inconsistencies and geometric distortion. Moreover, the paired video duration is limited to only 1.2 hours.

- **EgoExo-Fitness (2024)** [25]: Provides paired data captured with GoPro cameras. However, frames are resized to 456×256, and the dataset is limited to indoor fitness scenarios.
- **EgoExoLearn (2024)** [17]: Includes both egocentric and exocentric videos. However, the views are not temporally aligned, as egocentric videos are recorded by following exocentric demonstrations.
- **Ego-Exo4D (2024)** [13]: Provides large-scale paired exo-ego videos. However, the use of heterogeneous devices and fisheye-based egocentric capture introduces distortion and cross-view inconsistencies.
- **EgoPet (2024)** [6]: Extends egocentric data to animal subjects. However, it lacks paired exocentric views and is limited to 480p resolution.
- **Nymeria (2024), HOT3D (2025), and HD-EPIC (2025)** [5, 31, 35]: Use Meta Aria [11] for egocentric capture, introducing noticeable fisheye distortion, and do not provide paired exocentric views.
- **EgoLife (2025)** [52]: Uses Meta Aria [11] for egocentric capture and GoPro for exocentric views. Device inconsistency and fisheye distortion lead to visual discrepancies.
- **EgoExo-8K (2025)** [43]: Uses a head-mounted action camera for egocentric capture and a panoramic DJI camera for exocentric views, introducing visual inconsistencies across views. The viewpoint is further restricted to a fixed following angle, and the inclusion of synthetic GTA-5 data limits realism. As detailed statistics are not publicly available, we report the paired length as unknown in Table 1.

B.2 Dataset Preprocessing

Our preprocessing pipeline consists of three steps. (1) **Subject specification.** For each paired video, we manually annotate a concise appearance description of the camera wearer, which serves as the subject condition for viewpoint transformation in multi-person scenes. The description includes only discriminative cues, such as clothing color. (2) **Synchronized clipping.** We split each raw paired video into synchronized 5-second clips while preserving the original resolution, ensuring temporal alignment between the egocentric and exocentric views. (3) **Egocentric captioning.** We generate captions for each egocentric clip using Qwen3-VL-32B-Instruct [4] with the prompt shown in Figure 11. The prompt follows a viewpoint-grounded egocentric captioning protocol that produces dense first-person descriptions, capturing scene state, spatial layout, visible self-appearance, camera motion, temporal evolution, and first-person interactions. It also enforces strict observability, grounding descriptions only in visible evidence to reduce hallucinations.

B.3 Dataset Statistics

As shown in Figure 13, EgoScape spans diverse subjects, actions, and scenes, with all paired videos captured at 1080p. The paired exo-ego subset contains 19.5K clips (27.1 hours), covering 25 subject categories, 1367 actions, and 90 scenes. The additional egocentric subset includes 24.2K clips (33.6 hours), covering 64 subject categories, 1094 actions, and 255 scenes. We obtain the action and scene statistics using Qwen3-VL-32B-Instruct [4] with the prompt shown in Figure 12. Figure 14 further reports the duration ratio between egocentric-only and paired exo-ego videos. Overall, EgoScape combines high-resolution paired supervision with broad diversity in first-person videos, supporting egocentric video understanding and generation.

C Benchmark Protocol

C.1 EgoEval

We construct an evaluation benchmark, termed **EgoEval**, by randomly splitting 50 videos from EgoScape. The selected videos are disjoint from the training set and cover diverse scenes and actions. We will release this benchmark to facilitate future evaluation and comparison.

C.2 EgoCross

We construct a cross-domain benchmark, termed **EgoCross**, from H2O [24] and EgoBody [56], two datasets that provide paired-view videos with distortion-free imaging. H2O is captured by Azure Kinect at 720p, while EgoBody uses the HoloLens-2 at 1080p. Since both datasets are released as image sequences, we convert them into videos for consistent evaluation. We randomly sample 20 videos from each dataset, resulting in 40 cross-domain videos. As shown in Figure 15, we list the original filenames for transparency and reproducibility. EgoCross evaluates cross-domain generalization across datasets, scenes, subjects, and actions, as well as robustness to device variation.

C.3 EgoWild

We construct **EgoWild** to evaluate real-world generalization of exocentric-to-egocentric video generation. We randomly sample 200 publicly available online videos covering diverse subjects, scenes, and motion patterns. Unlike EgoEval and EgoCross, EgoWild does not provide paired egocentric ground truth, and instead contains only exocentric inputs for open-world evaluation. This setting evaluates whether a method can generate plausible egocentric videos under unconstrained real-world conditions. All videos are used for academic research only, and we release the source links rather than redistributing the original content.

D Evaluation Metrics

Egocentric Perception Quality. A minor change in the egocentric camera position or orientation can lead to substantial pixel-level variations. As a result, common fidelity metrics such as FVD [45], PSNR [16], SSIM [47], and LPIPS [55] do not reliably reflect perceptual quality in egocentric video generation. As shown in Tabs. 2 and 3, although EgoEye produces visually plausible results, its performance under these metrics does not always align with qualitative observations, and all methods further degrade under cross-domain evaluation due to misalignment between exocentric inputs and egocentric ground truth. This issue is exacerbated by cross-view discrepancies in camera devices, viewpoints, and imaging conditions, as well as low-resolution data. To address this limitation, we introduce *Egocentric Perception Quality*, which evaluates high-level first-person perception rather than pixel-level similarity. This metric measures three complementary aspects: Viewpoint Authenticity (VA), Interaction Plausibility (IP), and Prompt Consistency (PC), capturing viewpoint realism, physical interaction plausibility, and semantic alignment, respectively. As shown in Figure 16, we use Qwen3-VL-32B-Instruct [4] to score generated videos given video frames and ground-truth egocentric captions. We will release the evaluation code to support reproducible and standardized evaluation.

Viewpoint Transformation Consistency. As shown in Figure 4, some methods largely preserve exocentric content rather than performing effective viewpoint transformation. To assess this ability, we introduce *Viewpoint Transformation Consistency*, which measures whether the generated video is closer to the egocentric ground truth than to the exocentric input in a feature space. We use DINOv3 [42] to extract video-level representations for the generated, egocentric, and exocentric videos, treating the egocentric video as a positive reference and the exocentric video as a negative reference. We report two scores: S_g, the cosine similarity between the generated and egocentric videos, and S_{gx}, which further encourages closeness to the egocentric reference while pushing away from the exocentric input. A higher S_{gx} indicates stronger viewpoint transformation consistency.

Human Evaluation. We conduct a user study with 20 annotators, who rate VA, IP, and PC on a 0-10 scale, and report the average scores. As shown in Figure 17, we develop a dedicated evaluation interface that allows multiple annotators to submit ratings in a standardized manner. The interface synchronously plays the input, generated, and ground-truth videos to facilitate temporal consistency comparison, and also provides the corresponding prompt for evaluating prompt consistency.

Table 5: Training hyperparameters of DiT and motion alignment in EgoEye.

Parameter	Value
Model	
Base model	Wan2.2-TI2V-5B-Diffusers
Drop text	True (drop ratio = 0.1)
LoRA rank	80
Precision	BF16 mixed precision
Gradient checkpointing	True
Optimization	
Batch size	4
Gradient accumulation	1
Max training steps	6000
Warmup steps	10
Learning rate	5×10^{-5}
Weight decay	1×10^{-2}
Motion module	
Motion weight	0.1
Motion dimension	256
Motion layers	{9, 19, 29}
Data	
Input resolution	704×1280
Number of frames	49
Frame stride	2
Num workers	8
Shuffle	True
Drop last	True
Pin memory	True

E Implementation Details

E.1 EgoEye Implementation

As shown in Table 5, we summarize the training configuration of EgoEye, including key hyperparameters and optimization settings. Training follows a two-stage pipeline. We first perform text-to-video egocentric pretraining on large-scale egocentric videos to enhance egocentric visual priors. The pretrained model is then used as the initialization for video-to-video training for exocentric-to-egocentric transformation. Due to the latency of LLM-based reward evaluation and its decoupling from diffusion training, we optimize the reasoning module separately from the video generation module. During inference, we use 50 denoising steps for video generation. We detail egocentric pretraining, reward-guided context reasoning, and multi-stage motion alignment in the following subsections.

Egocentric Pretraining Recent video generation models are primarily trained on large-scale exocentric data, limiting their ability to capture egocentric-specific characteristics such as first-person motion and viewpoint-dependent scene structure. To address this limitation, we perform LoRA-based fine-tuning on Wan2.2 [46] to inject egocentric priors. The training configuration is identical to that of the second stage (Table 5), except that motion alignment and reasoning modules are not applied in this stage.

Reward-guided Reasoning for Egocentric Context To better infer egocentric context from exocentric videos, we optimize the reasoning process via GROP [40] with three complementary rewards.

(1) **Viewpoint Reward** R_v enforces a consistent first-person perspective, avoiding third-person bias.

(2) **Spatial Reward** R_{sp} enforces correct spatial relations, particularly left–right consistency under viewpoint transformation.

(3) **Semantic Reward** R_{se} encourages semantic completeness, including accurate object attributes and interactions.

Table 6: Training hyperparameters for Reward-guided egocentric context reasoning.

Parameter	**Symbol**	**Value**
Data and Input		
Video FPS	–	2.0
Max prompt length	–	4096
Max response length	–	256
Rollout batch size	–	16
Min pixels	–	65536
Max pixels	–	602112
Random seed	–	1
GRPO Objective		
Advantage estimator	–	GRPO
KL loss	–	Enabled
KL penalty type	–	low-var KL
KL coefficient	λ	1×10^{-2}
Reward filter key	–	overall
Reward filter range	–	$[0.01,\ 0.99]$
Policy Model and Optimization		
Policy model	–	Qwen3-VL-32B-Instruct
Learning rate	–	5×10^{-6}
Weight decay	–	1×10^{-2}
Warmup ratio	–	0.03
Max grad norm	–	1.0
Optimizer	–	AdamW (BF16)
Gradient checkpointing	–	True
Freeze vision tower	–	True
LoRA rank	–	32
LoRA alpha	–	32
LoRA target modules	–	all-linear
LoRA excluded modules	–	.*visual.*
Rollout Sampling		
Group size	G	2
Sampling temperature	–	1.0
Top-p	–	1.0
Tensor parallel size	–	4
GPU memory utilization	–	0.70
Max batched tokens	–	16384
Reward Composition		
R_{v} weight	α	0.4
R_{sp} weight	β	0.3
R_{se} weight	γ	0.3
Reward workers	–	8 CPUs / 8 workers
Training Setup		
Total epochs	–	3
GPUs per node	–	8
Global batch size	–	16
Micro-batch size (update)	–	2
Micro-batch size (experience)	–	4

We use GPT-OSS-120B [1] as a reward model to score generated egocentric descriptions against ground-truth captions, using the evaluation prompt shown in Figure 19. The reasoning model is optimized using these rewards, with detailed hyperparameters provided in Table 6.

Multi-Stage Motion Alignment The module enforces motion consistency across multiple stages of the DiT model. We extract intermediate features from layers $\{9, 19, 29\}$ to align motion representations at early, middle, and late stages. A lightweight motion aligner maps both DiT token features and ground-truth VAE latent features into a shared motion embedding space. Spatially pooled token features are passed through a token projection head, while pooled VAE features are processed by a

VAE projection head, each consisting of LayerNorm followed by a bias-free linear layer. Motion is defined as frame-to-frame differences in the embedding space and supervised with ℓ_1 and cosine losses to align predicted motion with ground-truth egocentric motion. Detailed hyperparameters are provided in Table 5.

E.2 Baseline Implementation

We describe the implementation of baseline methods, grouped into three categories based on their conditioning signals. For fair comparison, all methods use the same reasoned egocentric context prompts as EgoEye during evaluation, rather than simple viewpoint transformation commands.

Egocentric Camera Trajectory-Based Methods We follow the EgoX [19] camera trajectory annotation pipeline, as illustrated in Figure 18. For each video, we manually annotate three benchmark trajectories, requiring approximately 10 days of effort for all 290 videos. The process is challenging, as it involves aligning head orientation using 3D point clouds, and the egocentric view is often blurry, further increasing annotation difficulty.

Egocentric First-Image-Based Methods To obtain the first-frame condition, we use Qwen-Image-Edit [50] to generate a first-person view from the initial frame with a transformation prompt. The prompt enforces a realistic first-person perspective while preserving the original scene and actions. As shown in Figure 20, the generated first frame may already contain artifacts or inaccuracies, which can accumulate during subsequent video generation, leading to degraded temporal consistency and visual quality.

Other Methods For fair comparison, all methods use identical prompts, and we reproduce representative video-to-video baselines, including VACE [18], UniVideo [48], Helios [54], and CogVideoX [53].

F Runtime and Efficiency Analysis

We analyze the time cost of existing methods, including additional overhead from human annotation and context description. For a fair comparison, we omit prompt construction time, since all methods use the same prompts as EgoEye. All methods are evaluated under the same inference setting, using 50 denoising steps to generate 49 frames at a resolution of 832×480 on a B200 GPU.

As shown in Figure 21, EgoX [19] is the most time-consuming due to heavy preprocessing, including manual annotation and point cloud rendering. Our method achieves competitive efficiency. Compared to Wan2.2 [46], our approach introduces no additional inference overhead, as motion alignment is applied only during training. It also removes the need for manual egocentric context description by enabling automatic reasoning from exocentric inputs.

G Additional Discussions

Reproducibility. We will release our dataset and code to support reproducibility. Moreover, as shown in Table 4, training WorldWander on our dataset yields significant improvements over its base model, Wan2.2 [46].

Ethics and Data Governance During data collection, participants were clearly informed of the intended usage scope, and explicit consent and authorization were obtained from all individuals appearing in the videos, with compensation provided above the local minimum wage. EgoWild is constructed from publicly available online videos, and we will provide the corresponding source links for transparency. The dataset is intended strictly for research use, and users will be required to adhere to a usage agreement upon release.

H Future Work

Although our method performs robustly across diverse scenes, it can fail when the exocentric input is ambiguous. As shown in Figure Figure 22, generated videos remain coherent and artifact-free

but may contain semantically incorrect details. For example, when a person is facing away from the camera, and the hand-held object is heavily occluded [8, 9], the reasoning module can infer the hand motion and partially reconstruct plausible interactions from temporal context. However, if the object never appears throughout the entire video, the observation becomes inherently ambiguous, and the model cannot recover the unique ground-truth instance. Instead, it produces a reasonable hypothesis consistent with the observed scene and motion. Another limitation is the lack of camera customization. EgoEye operates in a canonical camera setting with implicitly learned parameters and does not support user-specified configurations, such as lens characteristics. Nevertheless, EgoEye demonstrates the feasibility of generating realistic egocentric videos from a single exocentric input. Future work will explore explicit camera conditioning to enable more flexible control.

I Broader impacts

This work enables egocentric video generation from exocentric observations, benefiting applications such as filmmaking, virtual reality, embodied AI, robotics, and immersive content creation. The proposed framework may support more efficient first-person video production and facilitate future research on cross-view video understanding and generation. Potential negative impacts mainly arise from inappropriate use of realistic video generation technologies, including synthetic first-person content that may be misleading in certain scenarios. However, our work focuses on viewpoint transformation rather than identity manipulation or unrestricted generation. To reduce potential risks, the released resources will be limited to research purposes under a usage agreement, and EgoWild only provides publicly accessible source links rather than redistributing original videos.

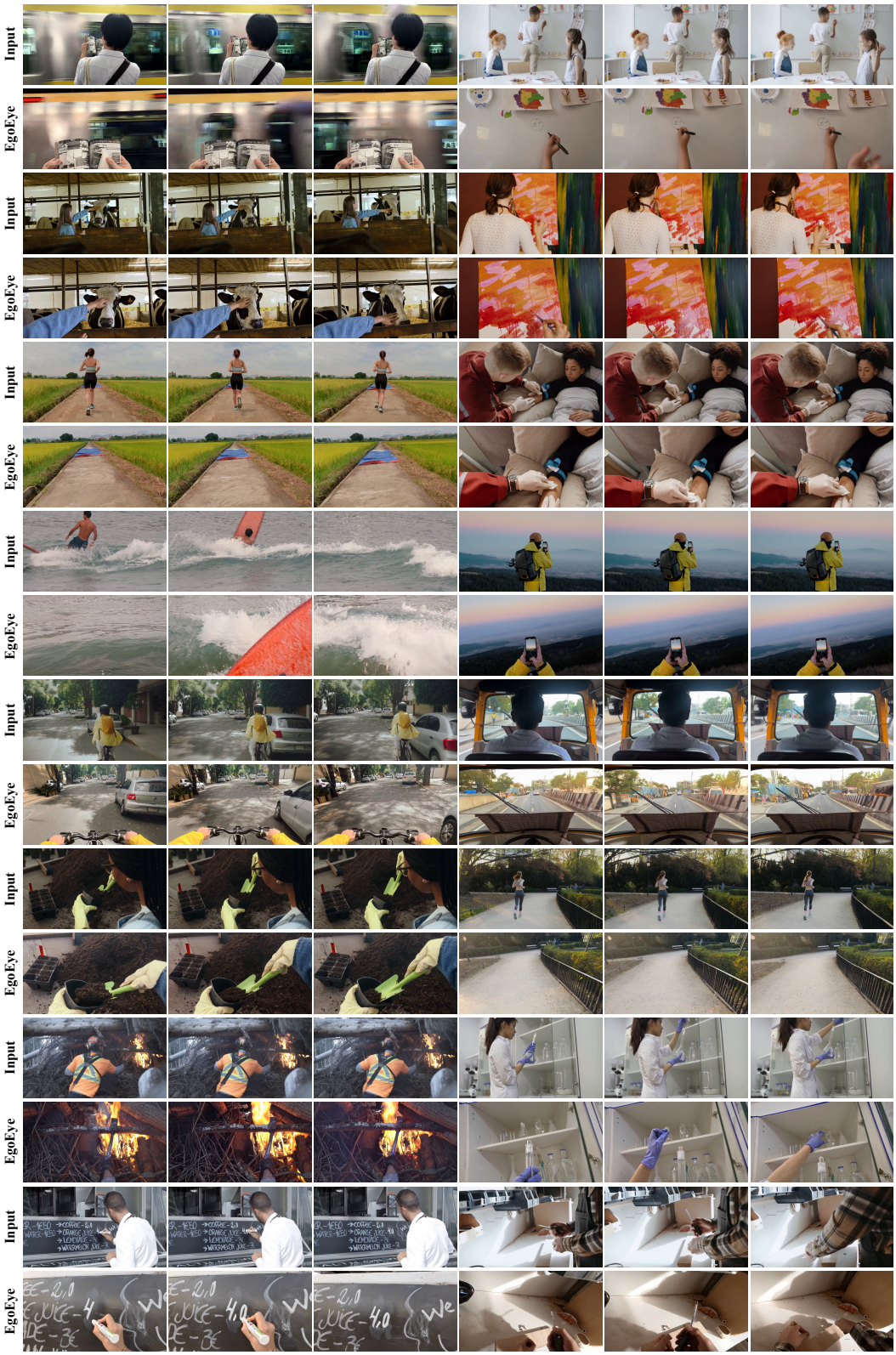

Figure 9: **More visualization of EgoEye on EgoWild.**

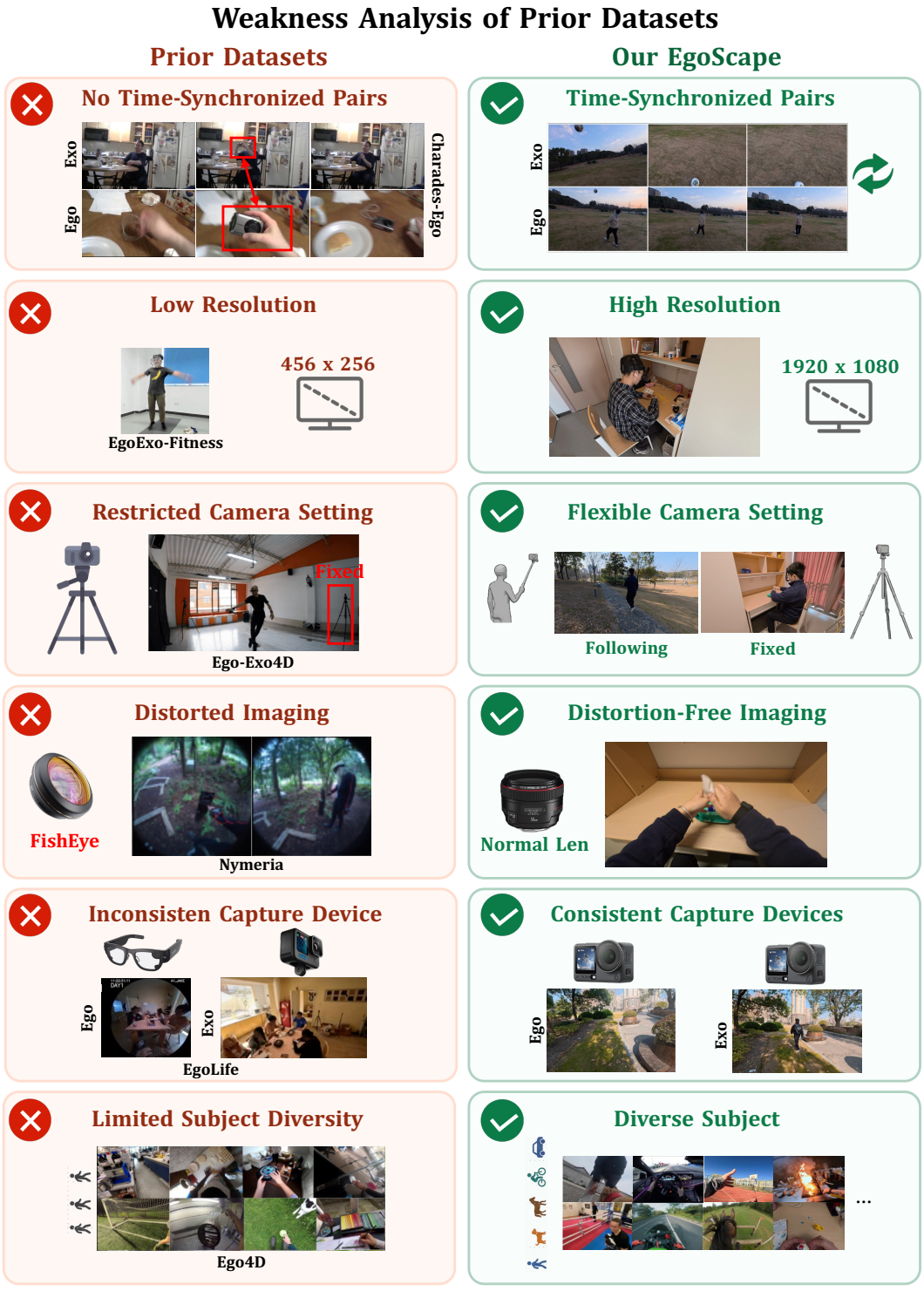

Figure 10: **Weakness Analysis of Prior Datasets.** Compared with existing egocentric datasets, EgoScape provides time-synchronized exo-ego video pairs, high-resolution videos, flexible camera settings, distortion-free imaging, consistent capture devices, and diverse subjects.

You are an egocentric video captioner.

The video is recorded from a first-person (egocentric) viewpoint, and all descriptions must be grounded strictly in what is visually observable from the camera's perspective only.

The camera defines the viewpoint of the scene, and the camera-bearing agent must not be assigned any identity, role, intent, or internal state unless such information is directly visible.

Describe the video in one single coherent paragraph of natural language, without using lists, bullet points, timestamps, or any structured formatting.

Your description must implicitly contain two types of information: static visual state and dynamic evolution over time, woven into a natural egocentric narrative.

First, describe the static visual state of the scene, including the visible appearance of the camera-bearing agent itself as seen from first-person view (such as visible body parts, clothing, accessories, or items being carried), the detailed visual appearance of other visible people, animals, or objects in the scene, and the persistent environmental layout and spatial context, describing the overall geometry of the space, the relative spatial arrangement of major surfaces and boundaries (such as walls, floors, ceilings, roads, terrain, or open areas), dominant materials and textures, large immobile structures or landmarks, the relative depth and occlusion relationships between visible entities, and any visually perceivable lighting conditions, dominant light sources, brightness, shadows, color tone, exposure level, or optical effects.

Then describe the dynamic evolution over time, including how the camera moves through space (changes in direction, speed, height, rotation, or stability) and how this motion alters the visual field, how the camera-bearing agent's own visible actions change over time, how other entities move or change over time, how interactions between the agent and other entities unfold visually from the first-person perspective, and any visually perceivable changes in lighting, visibility, scale, or motion magnitude.

All events must be described in temporal order, emphasizing continuity and smooth transitions.

Do not infer emotions, intentions, causes, or unobservable mental states.

Do not introduce any information that is not directly visible.

If any aspect of the scene is ambiguous or unclear, explicitly express the uncertainty in the description.

The output must be exactly one natural paragraph and nothing else.

Figure 11: **Egocentric captioning prompt.** The prompt guides Qwen3-VL-32B-Instruct [4] to generate dense, viewpoint-grounded first-person descriptions that capture scene state, spatial layout, visible self-appearance, camera motion, temporal evolution, and interactions, while enforcing strict observability to reduce hallucinations.

You are an egocentric video labeling assistant.

The input is a first-person video. Your task is to assign three short labels:
1.subject
2.scene
3.action
Definitions:
•subject: the main subject category of the video.
Use a short word or short phrase only.
Prefer the dominant camera-wearer identity or the most central visible target inferred strictly from visual evidence.
Examples:
person, guitarist, pianist, fitness person, badminton player, firefighter, barber, skier, runner,
dog, cat, horse, bird
•scene: the main environment or location.
Use a short word or short phrase only.
Examples:
badminton court, grass field, road, dorm room, bedroom, classroom, gym, kitchen,
living room, street, sidewalk, park, forest, beach, bathroom, barbershop
•action: the main action happening in the video.
Use a short word or short phrase only.
Examples:
playing guitar, playing piano, working out, running, walking, jogging,
cutting hair, drawing, painting, extinguishing fire, playing badminton,
picking up shuttlecock, cooking, cleaning, riding bicycle
Rules:
•The video is first-person. Only use visually observable information.
•Keep each field short: one word or a short phrase.
•Do not write full sentences.
•If the subject is a human and there is a clear profession / role / identity cue, output the more specific label,
such as guitarist, pianist, firefighter, barber, fitness person, badminton player.
•If the subject is an animal, output the specific animal category, such as dog, cat, horse.
•If uncertain, output the most likely visible category.
•If very unclear, use "unknown".
•Output exactly one line in this format and nothing else:
subject,scene,action

Figure 12: **Dataset annotation prompt.** We use Qwen3-VL-32B-Instruct [4] with the designed prompt to obtain action, scene and subject statistics for EgoScape.

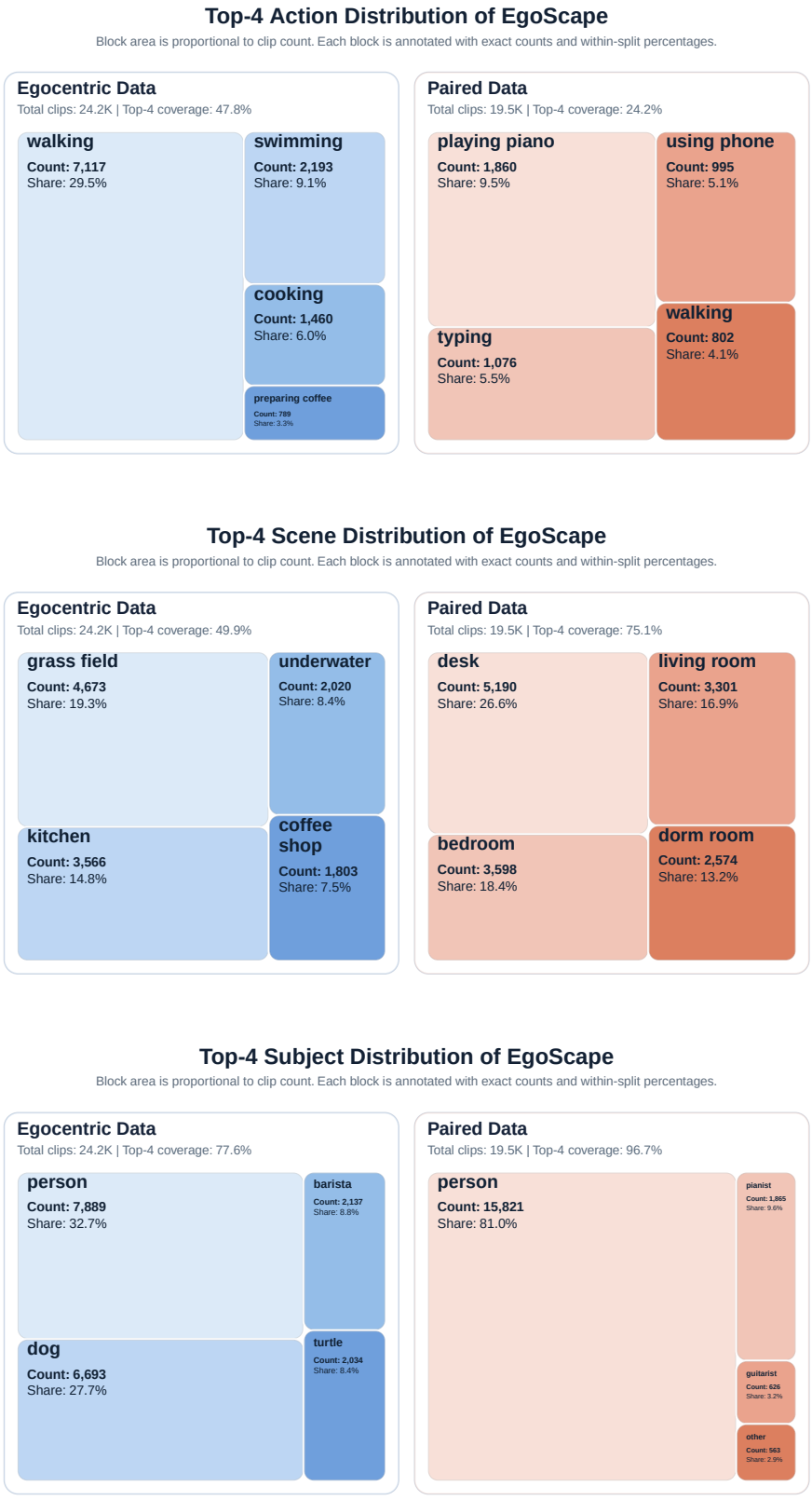

Figure 13: **EgoScape statistics.** EgoScape includes 19.5K paired exo-ego clips and 24.2K additional egocentric clips, covering diverse actions, scenes, and subjects.

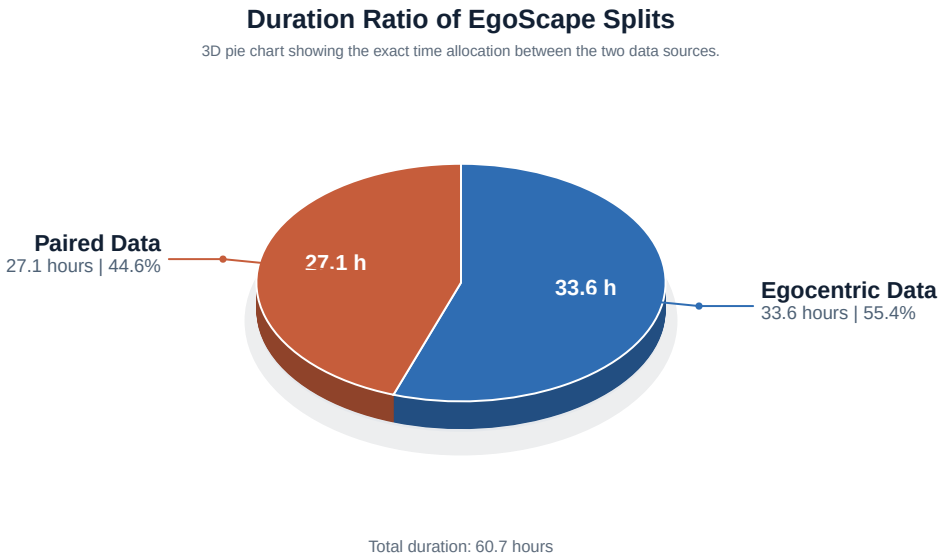

Figure 14: **EgoScape duration composition.** EgoScape contains 27.1 hours of paired exo-ego videos and 33.6 hours of additional egocentric videos

HOI	EgoBody
subject1_h2-2.mp4	recording_20210910_S05_S06_01.mp4
subject1_k1-5.mp4	recording_20210910_S06_S05_02.mp4
subject1_k1-7.mp4	recording_20210911_S06_S07_02.mp4
subject1_k2-0.mp4	recording_20210921_S05_S12_02.mp4
subject1_k2-5.mp4	recording_20210921_S10_S11_01.mp4
subject2_h1-0.mp4	recording_20210929_S15_S11_01.mp4
subject2_h1-1.mp4	recording_20210929_S16_S05_01.mp4
subject2_h2-1.mp4	recording_20211004_S19_S06_01.mp4
subject2_h2-2.mp4	recording_20211004_S20_S12_01.mp4
subject2_o2-6.mp4	recording_20211004_S20_S12_02.mp4
subject3_h2-2.mp4	recording_20220215_S21_S22_01.mp4
subject3_h2-6.mp4	recording_20220215_S22_S21_02.mp4
subject3_k1-1.mp4	recording_20220218_S23_S02_02.mp4
subject3_k1-2.mp4	recording_20220312_S28_S29_01.mp4
subject3_k1-6.mp4	recording_20220312_S28_S29_02.mp4
subject4_h1-5.mp4	recording_20220315_S30_S21_01.mp4
subject4_h2-5.mp4	recording_20220318_S31_S32_01.mp4
subject4_k2-5.mp4	recording_20220318_S32_S31_01.mp4
subject4_k2-6.mp4	recording_20220318_S33_S34_02.mp4
subject4_o1-6.mp4	recording_20220415_S36_S35_02.mp4

Figure 15: **EgoCross video selection.** Original filenames of the randomly selected videos from H2O [24] and EgoBody [56].

You are an expert evaluator for exo-to-ego video generation.

Your task is to evaluate whether the generated video matches the intended FIRST-PERSON (EGO) experience described below.

Ground-truth ego prompt:
\"\"\"{gt_prompt}\"\"\"

Evaluation rules:
- Focus only on visible evidence in the video.
- Do not hallucinate unseen details.
- Evaluate each metric independently.
- Ego videos may include natural motion shake, partial occlusion, and imperfect stability.
- Do NOT reward cinematic third-person framing if it weakens first-person realism.
- This task focuses on egocentric video generation. Videos that do not exhibit a first-person viewpoint should be considered inferior, even if they are visually high-quality.
- All metrics use the same direction: higher is better.

Scoring:
- Each metric is in [0, 1].
- 1.0 = the video strongly satisfies the metric.
- 0.0 = the video strongly violates the metric.

Evaluate the following 3 high-level metrics:

1. Viewpoint Authenticity
- Does the video present a genuinely egocentric (first-person) viewpoint consistent with a human-worn camera?
- Consider camera position, viewpoint behavior, and motion patterns.
- High: clear first-person perspective with body-attached motion, such as head movement or natural shake.
- Low: third-person or observer-like view, floating camera, or overly cinematic framing.
- IMPORTANT: Natural first-person motion, such as head movement and slight shake, should be considered evidence of realism.

2. Interaction Plausibility
- Are interactions with objects and the environment physically plausible from a first-person perspective?
- Consider spatial alignment, hand/object interaction, and coordination between motion and viewpoint.
- High: interactions are physically grounded, body-driven, and spatially consistent.
- Low: misaligned interactions, floating or disconnected motion, or implausible object behavior.

3. Prompt Consistency
- Do the actions, objects, and scene details match the given prompt?
- Consider correctness of actions, presence and attributes of objects, and overall scene consistency.
- High: key actions, objects, and attributes are correctly represented.
- Low: missing or incorrect actions, objects, or scene elements.
- IMPORTANT: Do NOT consider viewpoint quality or visual artifacts here. Focus only on semantic correctness.

Return ONLY one JSON object with exactly this schema and no extra text:

{{
"Viewpoint_Authenticity": float,
"Interaction_Plausibility": float,
"Prompt_Consistency": float
}}

Figure 16: **Prompt for Egocentric Perception Quality.** We use this prompt to guide Qwen3-VL-32B-Instruct [4] in evaluating generated egocentric videos.

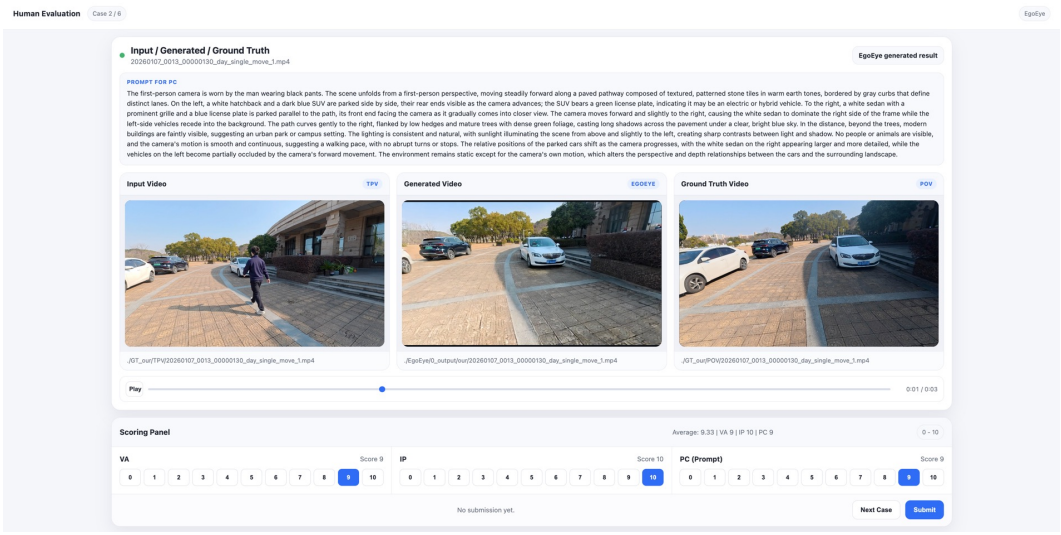

Figure 17: **Human Evaluation Interface.** The interface synchronously displays the input, generated, and ground-truth videos for temporal consistency comparison, while providing the corresponding prompt to assist annotators in evaluating viewpoint authenticity (VA), interaction plausibility (IP), and prompt consistency (PC).

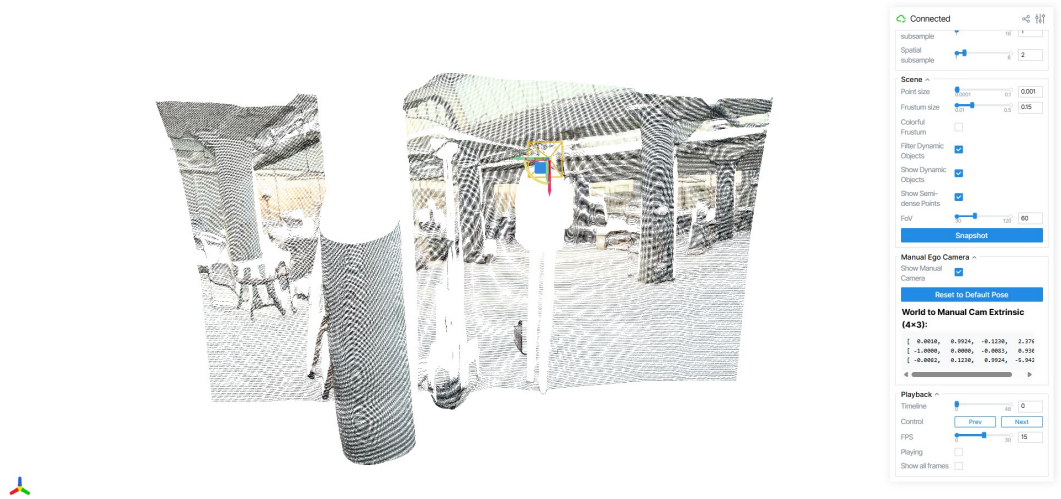

Figure 18: **Camera Trajectory Annotation Pipeline.** We follow the EgoX [19] pipeline to manually annotate benchmark camera trajectories by aligning head orientation with 3D point clouds, which is labor-intensive and challenging under blurry egocentric views.

R_v

_RUBRIC_VIEW = (
"Focus ONLY on left/right spatial references in the Candidate. "
"Score 10 if every left/right mention is consistent with the Reference "
"egocentric caption (i.e. uses the Subject's own handedness). "
"Score 5 if neither text uses left/right. "
"Score 0 if the Candidate clearly inverts left/right (mirror of reference)."
)

R_{sp}

_RUBRIC_SPATIAL = (
"Score a single shared value for caption style consistency, first-person faithfulness, "
"and viewpoint/subject compatibility. "
"The desired style matches the GT corpus: dense, objective, annotation-like first-person video description, "
"visually grounded, detailed, temporally ordered, and not conversational or chatty. "
"Describing OTHER people from outside is allowed; what must be penalised is narrating the Subject THEMSELVES "
"as if observed from outside, or implying body parts / clothing / held items that do not fit the Subject. "
"Use one overall 0-10 score for this merged aspect. "
"Score 10 if the Candidate strongly matches GT caption style, keeps a stable first-person viewpoint, "
"and all subject/viewpoint cues are compatible. "
"Score 7 if it mostly matches, with only minor style drift or weak viewpoint ambiguity. "
"Score 5 if style and viewpoint quality are mixed or only partially convincing. "
"Score 2 if it is clearly off-style, repeatedly slips into external narration of the Subject, "
"or shows several viewpoint/subject inconsistencies. "
"Score 0 if it is completely unlike the GT caption style or clearly cannot belong to the Subject's viewpoint."
)

R_{se}

_RUBRIC_SEMANTICS = (
"Score a single shared value for content fidelity, combining salient-object coverage and action/temporal consistency. "
"Judge whether the Candidate captures the main visible objects, landmarks, other people, and the key actions in the correct temporal order. "
"Use one overall 0-10 score for this merged aspect. "
"Score 10 if the Candidate preserves essentially all salient objects and the main action sequence in a temporally compatible order. "
"Score 7 if most salient objects and main actions are present, with only minor omissions or order drift. "
"Score 5 if some important objects/actions are present but coverage is partial or temporal structure is weakened. "
"Score 2 if only a small portion of the important objects/actions is preserved, or several major elements are wrong. "
"Score 0 if the Candidate misses most salient objects and major actions, or describes a clearly incompatible event."
)

Figure 19: **Reward Prompt for Egocentric Context Reasoning.**

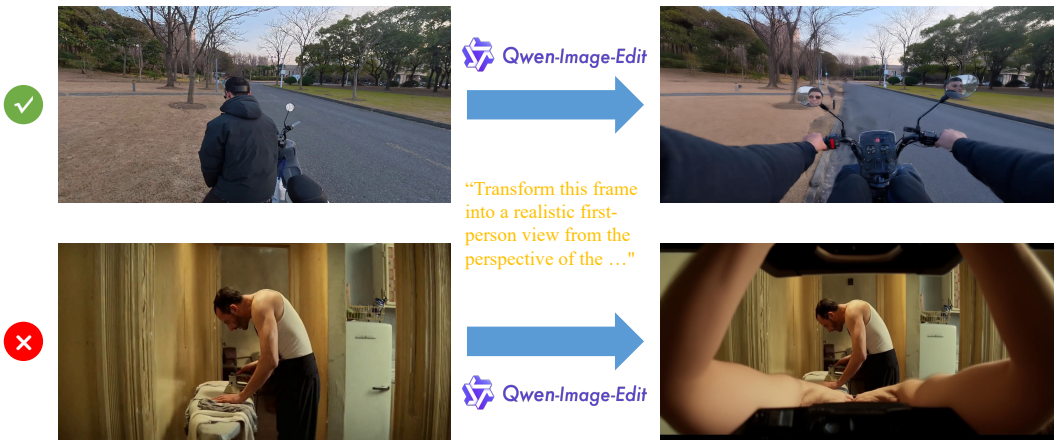

Figure 20: **Analysis of first-frame conditioning.** We use Qwen-Image-Edit [50] to transform the first exocentric frame into an egocentric first-frame condition. However, in some cases, the generated frame fails to produce the correct first-person viewpoint, which can affect subsequent video generation.

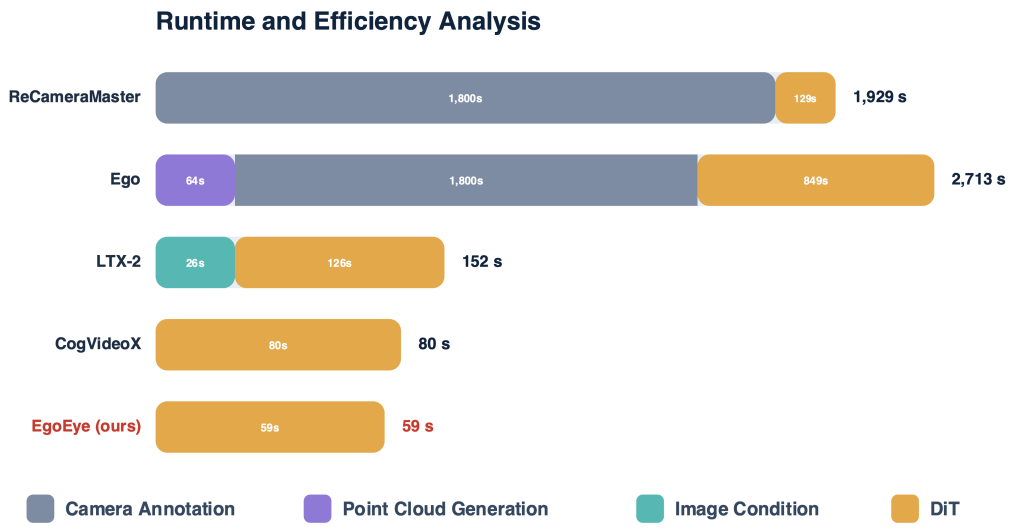

Figure 21: **Runtime Analysis.** Compared with other methods, EgoEye achieves stable and realistic generation results while maintaining the lowest runtime cost and highest efficiency.

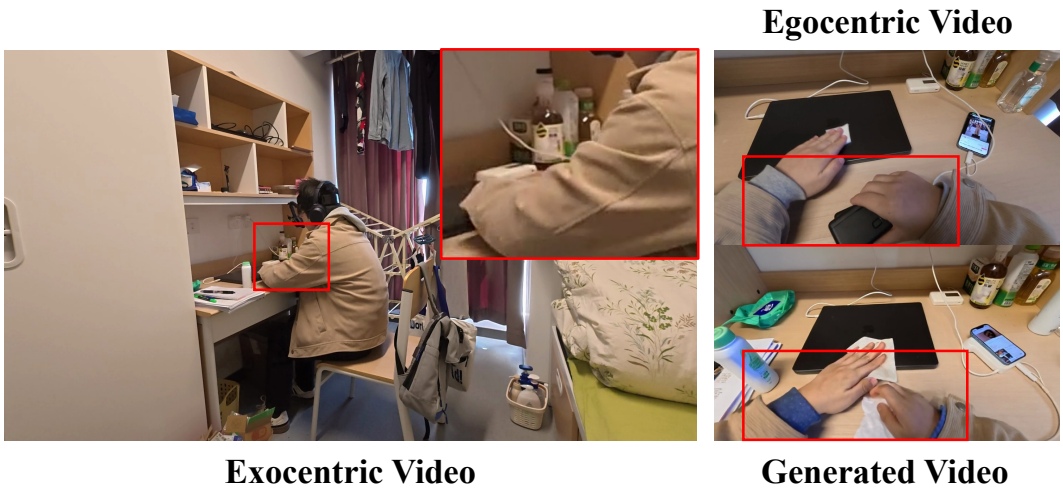

Figure 22: **Failure case under ambiguous observations.** When critical objects are fully occluded across the entire video, e.g., a hand-held object, the model can infer plausible hand motion from temporal context but cannot recover the unique ground-truth instance. As a result, it produces a reasonable yet semantically incorrect hypothesis consistent with the observed scene and motion

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: Our main contributions, including (1) introducing EgoEye as a solution framework for the exocentric-to-egocentric video generation task and (2) developing EgoScape with benchmarks and evaluation protocols, are accurately reported in the abstract and the introduction.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We discuss the limitations in the conclusion section.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [N/A]

Justification: Not applicable.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: We present the implementation details in the experiment section, with further details provided in Section E.

   Guidelines:

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

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

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

Answer: [Yes]

Justification: The proposed evaluation metric code and partial data are provided in the supplemental materials, and the complete data and code will be released upon acceptance.

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Details are provided in the experiment section and Supplemental Section E.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [No]

   Justification: Most of our experiments are not stochastic.

   Guidelines:

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

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Compute resource information are presented in the experiment part.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The research follows the rules.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: We discuss the potential impacts in the supplementary Section I.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: Not applicable.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: Those works are properly cited in the paper.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [N/A]

    Justification: Not applicable.

    Guidelines:

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

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

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

    Answer: [Yes]

    Justification: We provide the instructions and screenshots in the supplementary material. All data collectors were compensated above the local minimum wage.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: Not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: Not applicable.

Guidelines:

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