EgoEye: Seeing the World through Any Eyes

Yang Fu[1], Jianqin Wang[1], Xiangtai Li[2], Henghui Ding[1]*
[1] Fudan University, China [2] ByteDance
* Corresponding author: henghui.ding@gmail.com

NeurIPS 2026

Project & videos

Task: Exocentric-to-Egocentric Video Generation

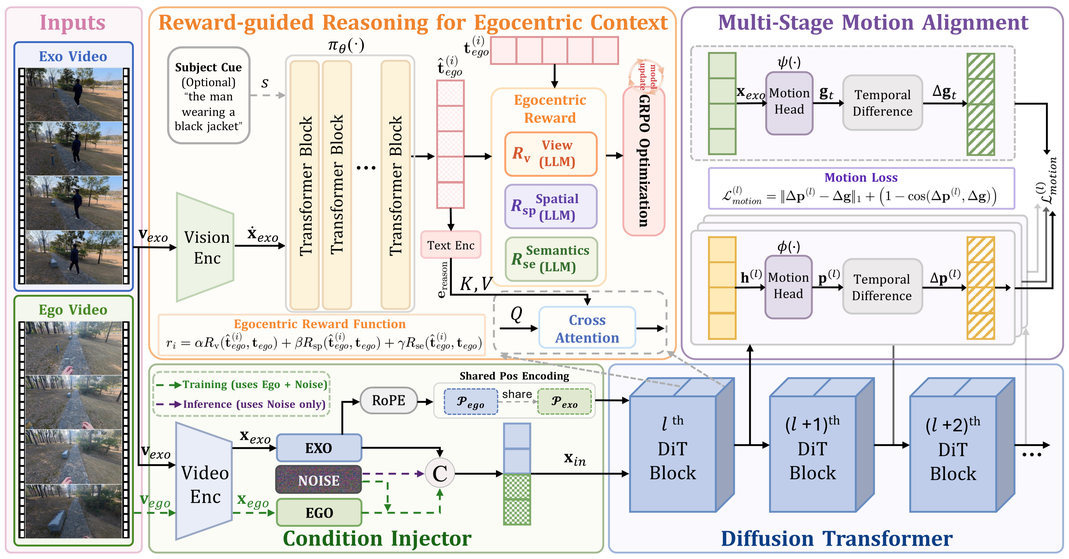
Goal: Generate the target subject's first-person view from one exocentric video.
Challenges: Large viewpoint shifts, unseen content, and cross-view motion.



Optional cue: "the boxer". No ego first frame or camera trajectory is required.

Method: Reward-Guided Reasoning + Motion Alignment

Reasoning: Optimize first-person context with viewpoint, spatial and semantic rewards.
Motion: Align temporal differences across DiT layers; pretrain on egocentric videos.



Ablation (Mean): Base 0.724 | +Motion 0.744 | +Pretrain 0.764 | +Reason 0.783 | All 0.807

Dataset: EgoScape

19.5K time-aligned exo–ego pairs (27.1 h)
24.2K extra egocentric videos (33.6 h)

Consistent devices, distortion-free imaging, static and dynamic cameras.
Human and animal subjects; 5-second clips at 30 fps.

Experimental Results

Quantitative comparison: EgoEval (50 pairs) and EgoCross (40 unseen pairs).

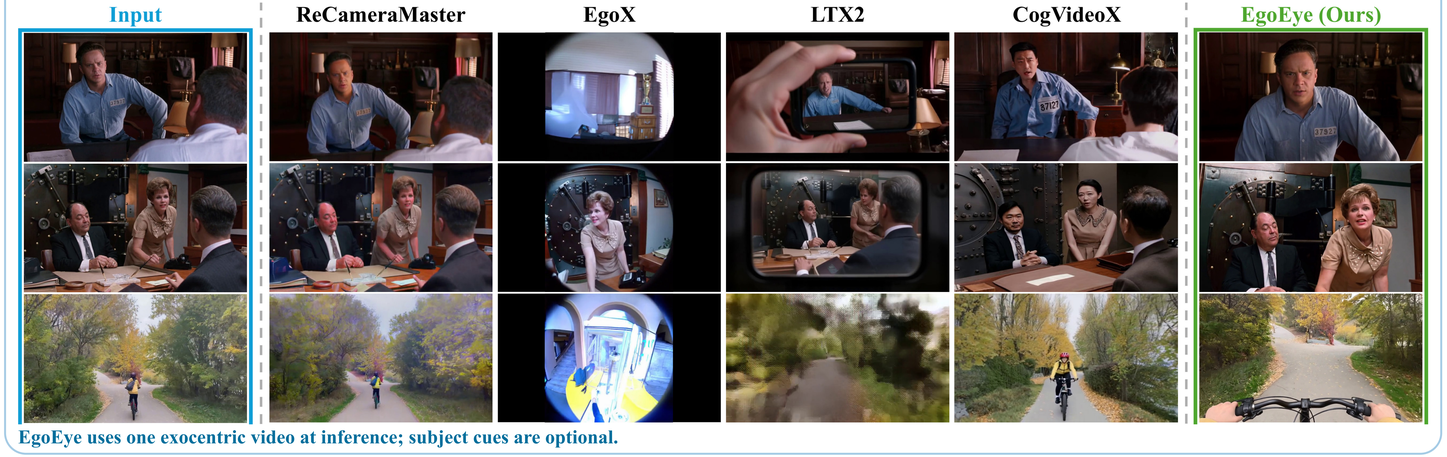
Method	Extra input	EgoEval			EgoCross		
		Mean ↑	FVD ↓	User ↑	Mean ↑	FVD ↓	User ↑
ReCameraMaster	T	0.062	1793.415	1.8	0.160	1970.137	2.8
EgoX	T	0.427	2284.654	4.9	0.448	2362.537	7.4
Wan2.2	F	0.417	1717.161	5.1	0.435	1851.583	7.1
LTX-2	F	0.675	1105.611	7.9	0.438	2000.426	7.2
Hunyuan	F	0.543	1448.783	6.2	0.419	2099.204	6.8
VACE	–	0.110	1919.772	1.1	0.110	2022.343	1.9
UniVideo	–	0.076	2006.733	1.4	0.157	2228.679	2.6
Helios	–	0.616	1693.351	6.5	0.254	1909.696	4.9
CogVideoX	–	0.186	1654.489	2.6	0.105	2023.574	1.7
EgoEye (Ours)	–	**0.807**	**917.326**	**9.2**	**0.550**	**1795.866**	**8.5**

Extra inputs: T = ego trajectory; F = ego first frame; – = none.
Mean = average of VA, IP and PC; User = human rating on a 0–10 scale.

EgoEval: Selected in-domain examples across viewpoint shifts and low-light scenes.



EgoWild: Selected in-the-wild examples with interactions and dynamic motion.



EgoEye uses one exocentric video at inference; subject cues are optional.