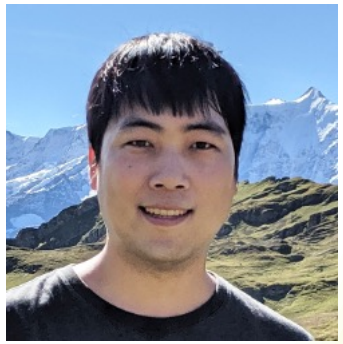
EgoWorld: Translating Exocentric View to Egocentric View using Rich Exocentric Observations





Junho Park[1], Andrew Sangwoo Ye[2], Taein Kwon[3†]

([†]Corresponding author)

[1]AI Lab, LG Electronics, [2]KAIST, [3]Visual Geometry Group, University of Oxford

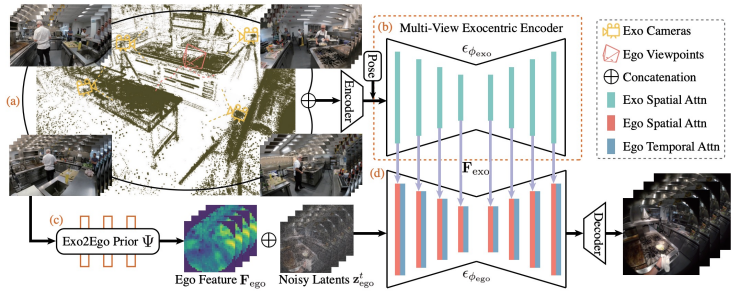
Motivation

Third-Person View

First-Person View

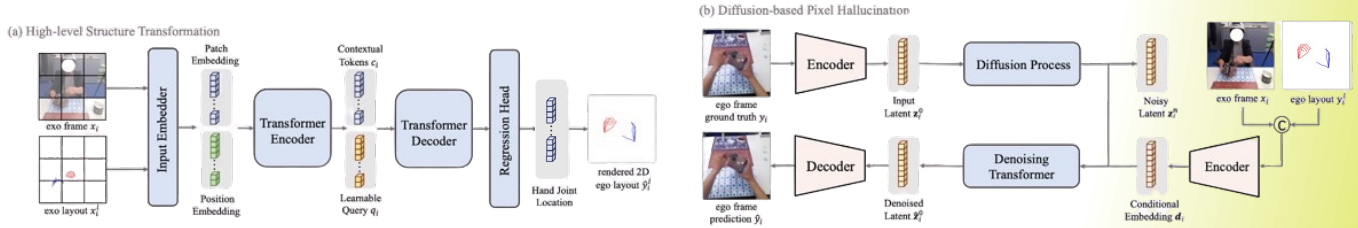
- **First-person views** are particularly valuable for capturing detailed hand-object interactions, which are essential in **skill-intensive tasks such as cooking, assembling, or playing instruments.**
- However, most existing resources are recorded from **third-person views**, primarily due to **the limited availability of head-mounted cameras and wearable recording devices.**

Previous Works



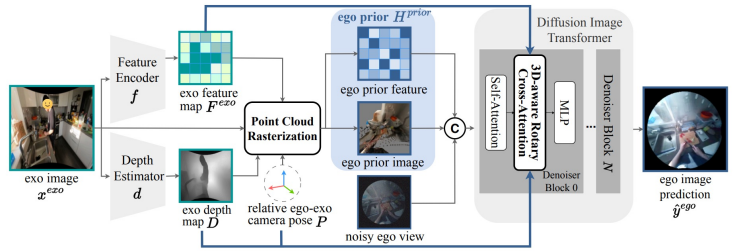
Multi-view Setting

Exo2Ego-V (NeurIPS 2024)



Pure 2D-based Generation

Exo2Ego (ECCV 2024)



Pre-defined Camera Pose

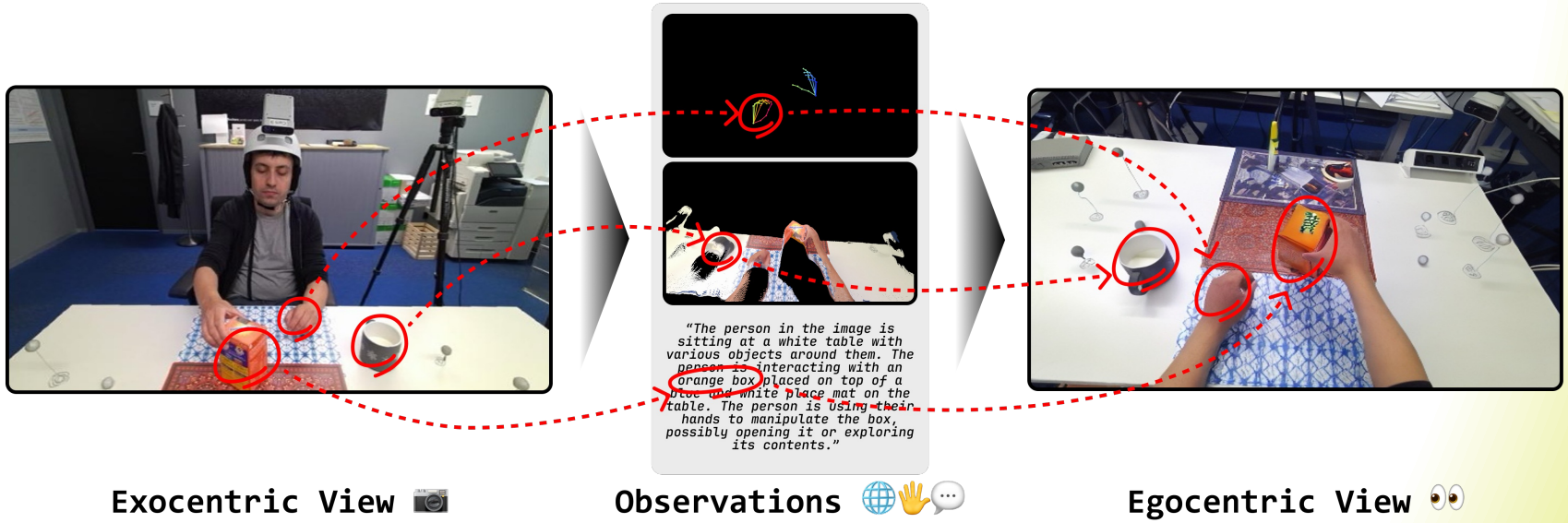
4Diff (ECCV 2024)

Given First Egocentric Frame

EgoExo-Gen (ICLR 2025)

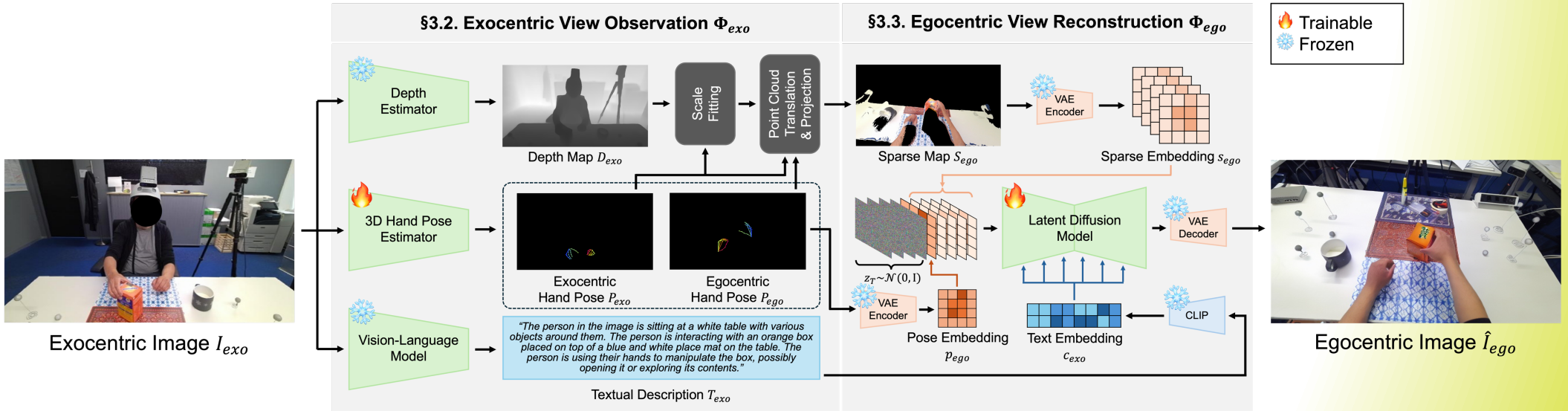
- Many existing approaches rely on input conditions.
  - **Multi-view setting, pre-defined camera pose, pure 2D-based generation, or given first egocentric frame.**
- These assumptions make them **impractical** for scenarios where **only single exocentric image is available.**

Method



- We introduce ***EgoWorld***, a novel end-to-end framework that **reconstructs high-fidelity egocentric views from a single exocentric image by leveraging rich multi-modal cues**, including point clouds, 3D hand poses, and textual descriptions.
- Our two-stage pipeline integrates **geometric reasoning and diffusion-based inpainting model** that significantly enhances **hand-object and scene generalization** and **semantic alignment** for generating egocentric images.

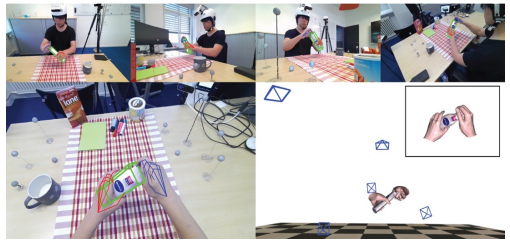
Method



- Stage 1: We construct a **point cloud** by combining the input exocentric RGB image with a scale-aligned exocentric depth map, using the **3D exocentric hand pose for spatial calibration**.
  - This point cloud is then **transformed into the egocentric view** using a transformation matrix computed from the **exo-ego 3D hand poses**.
- Stage 2: After the projection of the point cloud, a **egocentric sparse map** is obtained and it is subsequently **reconstructed into a dense, high-quality egocentric image** using a latent diffusion model.
  - To further enhance the semantic and geometric alignment of the hand-object reconstruction, we incorporate the **text description and estimated egocentric hand pose** during the reconstruction process.

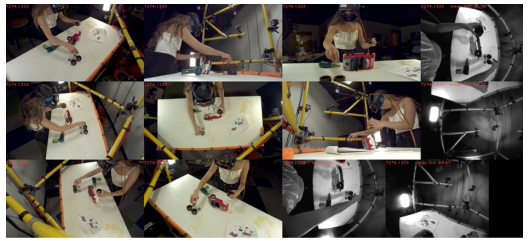
Experimental Setup

- Datasets



H2O

TACO



Assembly101

Ego-Exo4D

- Evaluation Metrics
  - **Frechet Inception Distance (FID)**: Distribution gap between generated and real images
  - **Peak Signal-to-Noise Ratio (PSNR)**: Pixel-wise reconstruction quality
  - **Structural Similarity Index Measure (SSIM)**: Perceptual structural similarity
  - **Learned Perceptual Image Patch Similarity (LPIPS)**: Learned perceptual difference
  - **Procrustes Analysis Mean Per Joint Position Error (PA-MPJPE)**: Aligned 3D joint error
  - **CLIPScore**: Image–text semantic alignment

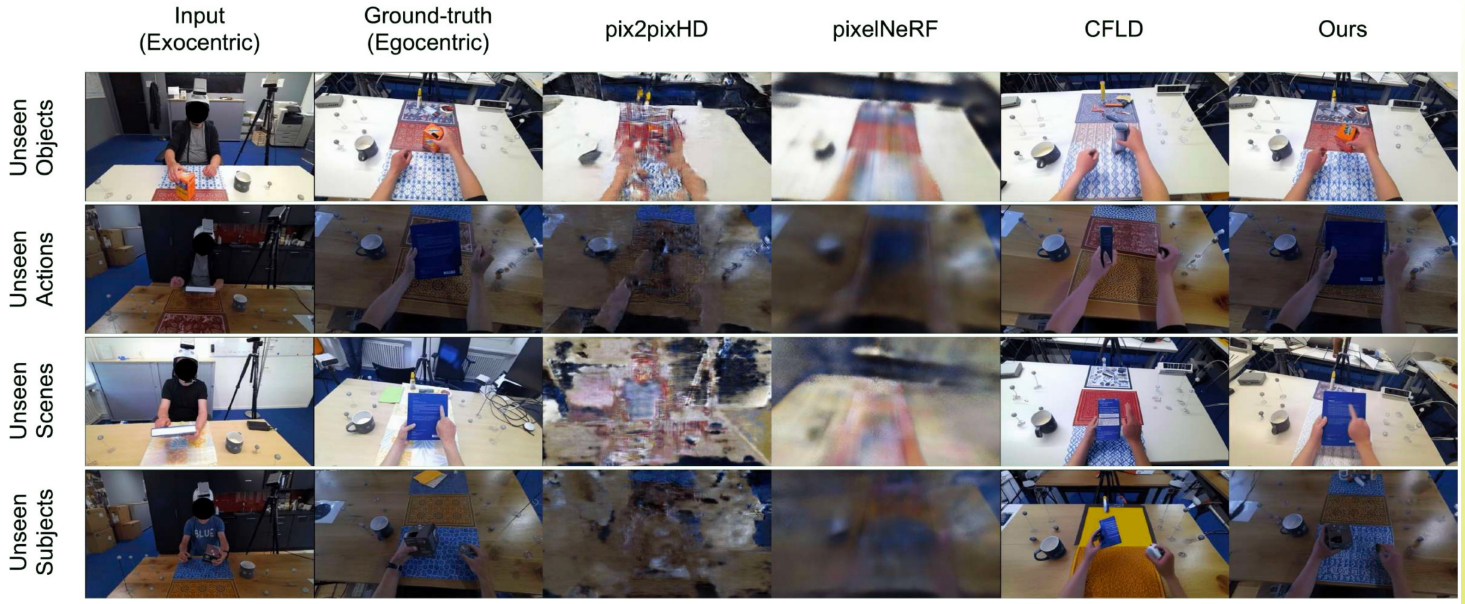
- Baselines
  - **pix2pixHD**: GAN-based image-to-image translation model
  - **pixelNeRF**: NeRF-based novel view synthesis model
  - **CFLD**: Pose-guided conditional latent diffusion model

Result #1: H2O w/ 4 Unseen Scenarios

ICLR
International Conference On Learning Representations

- Comparison on 4 Unseen Scenarios in H2O (i.e., unseen objects, actions, scenes, and subjects)

Scenarios	Unseen Objects						Unseen Actions					
Methods	FID↓	PSNR↑	SSIM↑	LPIPS↓	PA-MPJPE↓	CLIPScore↑	FID↓	PSNR↑	SSIM↑	LPIPS↓	PA-MPJPE↓	CLIPScore↑
pix2pixHD (Wang et al., 2018)	436.25	25.012	0.2993	0.6057	18.007	0.2302	211.10	24.420	0.2854	0.6127	17.754	0.2450
pixelNeRF (Yu et al., 2021)	498.23	26.557	0.3887	0.5372	15.746	0.2270	251.76	27.061	0.3950	0.8159	14.636	0.2315
CFLD (Lu et al., 2024)	59.615	25.922	0.4307	0.4539	7.9971	0.2656	50.953	28.529	0.4324	0.4593	8.1199	0.2699
EgoWorld (Ours)	**41.334**	**31.171**	**0.4814**	**0.3476**	**7.3178**	**0.2731**	**33.284**	**31.620**	**0.4566**	**0.3780**	**7.2602**	**0.2824**
Scenarios	Unseen Scenes						Unseen Subjects					
Methods	FID↓	PSNR↑	SSIM↑	LPIPS↓	PA-MPJPE↓	CLIPScore↑	FID↓	PSNR↑	SSIM↑	LPIPS↓	PA-MPJPE↓	CLIPScore↑
pix2pixHD (Wang et al., 2018)	490.32	18.567	0.2425	0.7290	20.229	0.2159	452.13	18.172	0.3310	0.7234	21.357	0.2311
pixelNeRF (Yu et al., 2021)	489.13	26.537	0.2574	0.7143	17.085	0.2097	493.13	22.636	0.4135	0.6838	18.131	0.2263
CFLD (Lu et al., 2024)	118.10	29.030	0.3696	0.6841	7.8766	0.2506	129.30	21.050	0.4001	0.6269	9.5606	0.2461
EgoWorld (Ours)	**90.893**	**31.004**	**0.4096**	**0.6519**	**7.4087**	**0.2585**	**96.429**	**24.851**	**0.4605**	**0.6188**	**8.1031**	**0.2582**

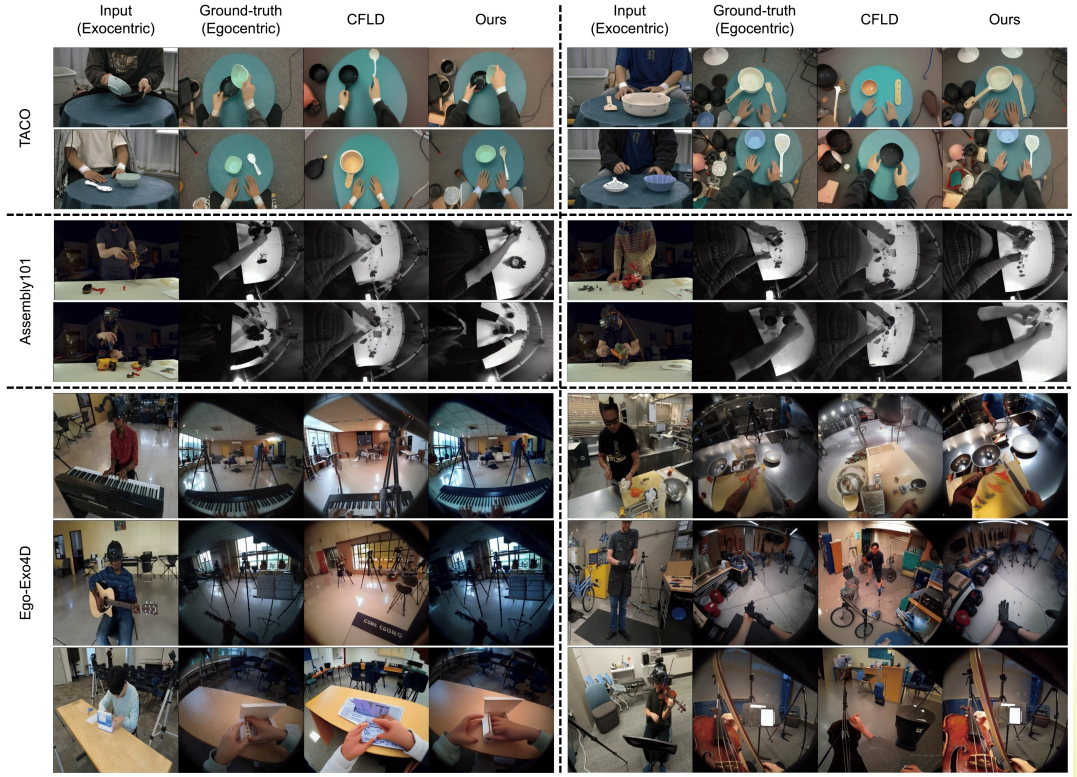


Result #2: TACO, Assembly101, Ego-Exo4D

ICLR
International Conference On Learning Representations

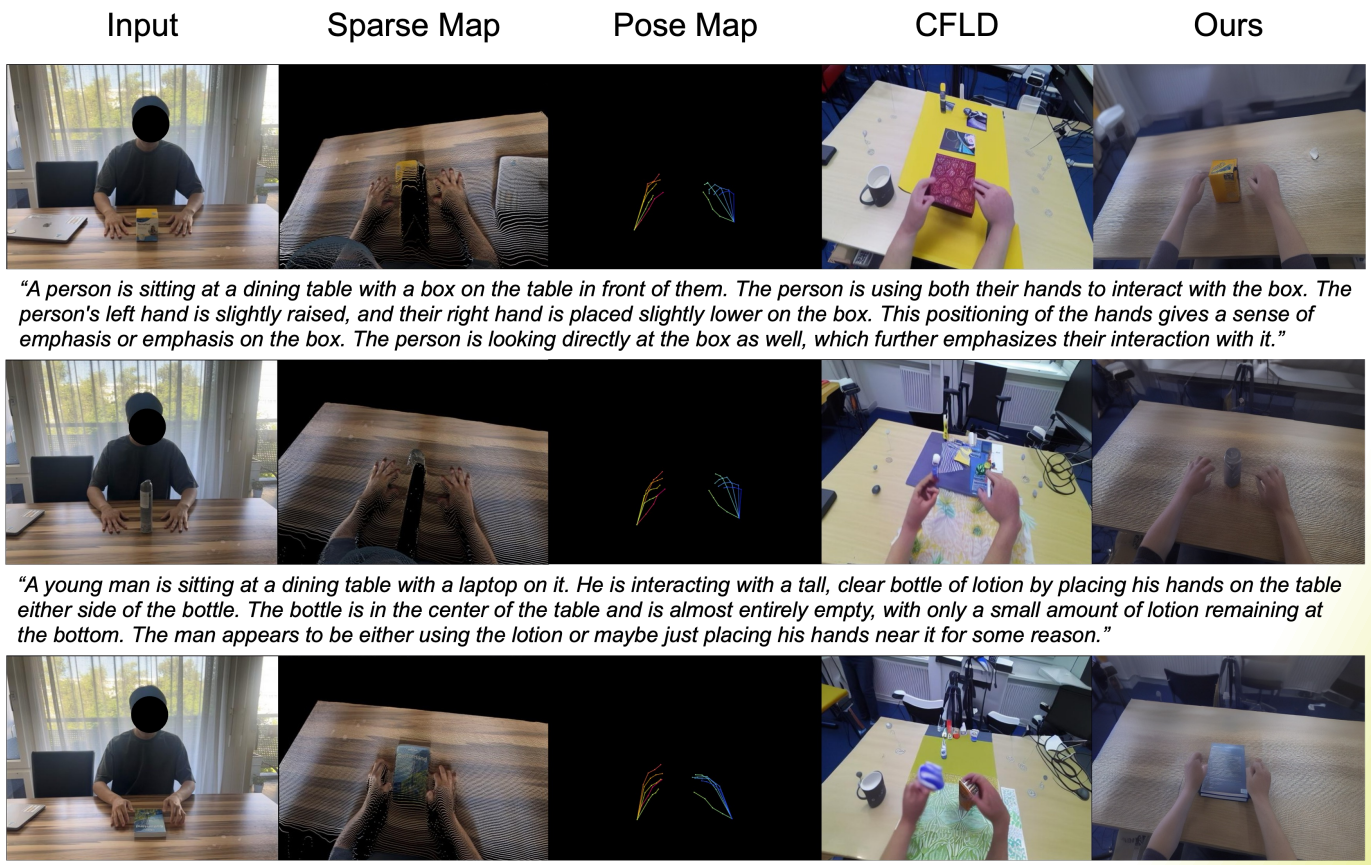
- Comparison on Challenging & Diverse Scenarios (i.e., TACO, Assembly101, and Ego-Exo4D)

Methods	TACO (Liu et al., 2024b)						Assembly101 (Sener et al., 2022)						Ego-Exo4D (Grauman et al., 2024)					
	FID↓	PSNR↑	SSIM↑	LPIPS↓	PA-MPJPE↓	CLIPScore↑	FID↓	PSNR↑	SSIM↑	LPIPS↓	PA-MPJPE↓	CLIPScore↑	FID↓	PSNR↑	SSIM↑	LPIPS↓	PA-MPJPE↓	CLIPScore↑
pix2pixHD (Wang et al., 2018)	227.87	25.875	0.2806	0.7037	19.054	0.2309	350.97	17.107	0.3587	0.6578	21.967	0.2114	401.48	14.792	0.3065	0.6899	25.082	0.2203
pixelNeRF (Yu et al., 2021)	302.19	26.661	0.3888	0.8543	16.137	0.2251	356.44	19.037	0.3761	0.6019	19.658	0.2070	367.39	17.347	0.3618	0.7134	23.793	0.2149
CFLD (Lu et al., 2024)	61.357	28.769	0.4009	0.5033	7.9078	0.2715	53.931	20.998	0.3988	0.5566	11.108	0.2458	70.476	21.578	0.3614	0.5975	15.010	0.2670
EgoWorld (Ours)	**37.191**	**30.155**	**0.4237**	**0.4025**	**7.3590**	**0.2828**	**50.232**	**25.365**	**0.4101**	**0.5142**	**10.561**	**0.2558**	**61.231**	**24.985**	**0.3986**	**0.5482**	**13.992**	**0.2862**



Result #3: Real-World Scenarios

- Comparison on Unconstrained Real-World Scenarios




ICLR
International Conference On
Learning Representations

Thank you☺