Generating Egocentric View from Exocentric View via Multimodal Observations

Junho Park
AI Lab, LG Electronics
junho18.park@gmail.com

Andrew Sangwoo Ye
KAIST
andyye@kaist.ac.kr

Taein Kwon†
VGG, University of Oxford
taein@robots.ox.ac.uk

Abstract

Egocentric vision is essential for both human and machine visual understanding, particularly in capturing the detailed hand-object interactions needed for manipulation tasks. Translating third-person views into first-person views significantly benefits augmented reality (AR), virtual reality (VR) and robotics applications. However, current exocentric-to-egocentric translation methods are limited by their dependence on 2D cues, synchronized multi-view settings, and unrealistic assumptions such as necessity of initial egocentric frame and relative camera poses during inference. To overcome these challenges, we introduce GenEgo, a novel two-stage framework that generates an egocentric view from multimodal exocentric observations, including projected point clouds, 3D hand poses, and textual descriptions. Our approach reconstructs a point cloud from estimated exocentric depth maps, reprojects it into the egocentric perspective, and then applies diffusion-based inpainting to produce dense, semantically coherent egocentric images. Evaluated on the H2O and TACO datasets, GenEgo achieves state-of-the-art performance and demonstrates robust generalization to unseen objects, actions, scenes, and subjects. Moreover, GenEgo shows promising results even on unlabeled real-world examples.

1. Introduction

Egocentric vision is crucial for understanding fine-grained hand-object interactions, which are central to skill-intensive tasks like cooking or assembling. Yet, most videos are recorded from third-person views due to the scarcity of head-mounted cameras, making it challenging to provide intuitive first-person guidance. Translating exocentric inputs into egocentric views would benefit AR/VR applications, instructional content, and robotics, where perception is inherently egocentric. For example, converting third-person instructional videos into first-person perspectives allows clearer visualization of finger movements and enables user-centered world models for real-time interaction.

Despite its potential, exocentric-to-egocentric translation is difficult due to large visual and geometric gaps between the two views. Egocentric views emphasize hands and manipulated objects, while exocentric views capture broader context. Occlusions, field-of-view restrictions, and unseen regions make geometric alignment challenging. Existing generative approaches often rely on restrictive assumptions such as multi-view inputs, pre-defined camera poses, or reference egocentric frames [2, 8, 20], limiting their applicability. A recent work, Exo2Ego [11], generates egocentric views from a single image, but depends heavily on accurate 2D hand layouts, making it unreliable under occlusion or clutter and prone to overfitting.

Therefore, we introduce *GenEgo*, a framework that generates egocentric view from exocentric view by leveraging multimodal exocentric observations. Our two-stage pipeline (1) extracts projected point clouds, 3D hand poses, and text descriptions from exocentric view, and (2) reconstructs dense egocentric view via diffusion models with exocentric observations. Specifically, we construct a point cloud by combining the input exocentric RGB image with a scale-aligned estimated exocentric depth map, using the 3D exocentric hand pose for spatial calibration. This point cloud is then transformed into the egocentric view using a translation matrix computed from the predicted 3D hand poses in both views. After the projection of the point cloud, a sparse egocentric image is obtained and it is subsequently reconstructed into a dense, high-quality egocentric image with semantic and structural guidance.

We evaluate the effectiveness of *GenEgo* through extensive experiments conducted on the H2O [7] and TACO [9] datasets, which provides well-annotated exocentric and egocentric video pairs. Our method achieves state-of-the-art performance on this benchmark. As a result, thanks to its end-to-end design, *GenEgo* demonstrates strong generalization across various scenarios, including unseen objects, actions, scenes, and subjects. Furthermore, we conduct on unlabeled real-world examples, and *GenEgo* shows powerful in-the-wild generalization, which implies *GenEgo* can extend to real-world use cases.

†Corresponding author.

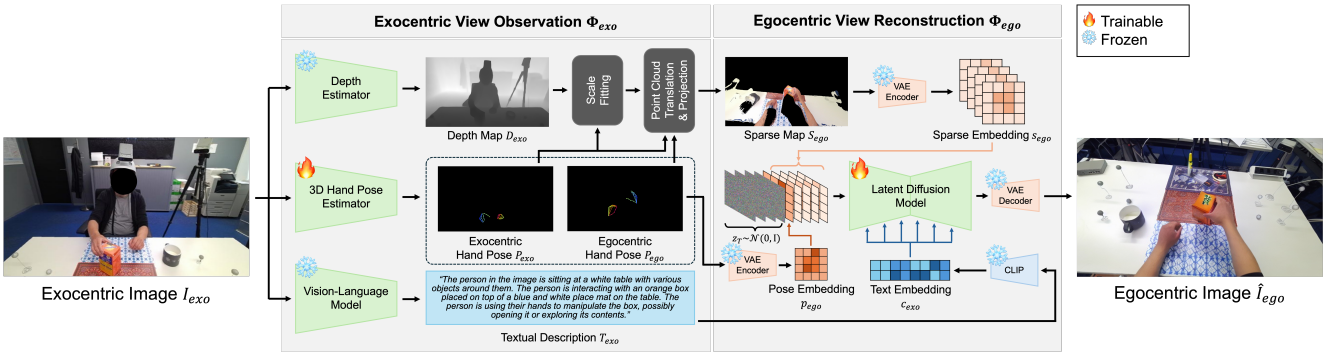

Figure 1. **Overall framework of *GenEgo*.** *GenEgo* is a two-stage pipeline : (1) Exocentric view observation Φ_{exo}, which extracts diverse observations from the exocentric view, including projected point clouds, 3D hand poses, and textual descriptions; and (2) egocentric view reconstruction Φ_{ego}, which reconstructs the egocentric view based on the exocentric view observation.

2. Method

2.1. Exocentric View Observation

As shown in Fig. 1, exocentric view observation Φ_{exo} takes various real-world observations, such as sparse egocentric RGB map S_{ego}, 3D egocentric hand pose P_{ego}, and textual description T_{exo}, from a single exocentric image I_{exo}.

First, with an off-the-shelf depth estimator [17], an exocentric depth map D_{exo} is extracted from I_{exo}. Next, a 3D exocentric hand pose P_{exo} is extracted from I_{exo} with an off-the-shelf hand pose estimator [22]. As D_{exo} provides only relative depth and is inherently affected by scale ambiguity, it is crucial to leverage P_{exo} for reasonable scale fitting. Thus, we extract a metrically-scaled P_{exo} and an exocentric hand depth map D_{hand} from the estimated MANO[14]-based mesh of P_{exo}. We define a hand region Ω_{hand}, which is a pixel-level valid area determined by D_{hand}, and compute a global scale factor s^* by comparing it with D_{exo}. Applying s^* yields a metrically-calibrated exocentric depth map $D'_{exo} = s^* D_{exo}$. Therefore, with I_{exo} and an exocentric camera intrinsic parameter K_{exo}, which is estimated from the off-the-shelf depth estimator, D'_{exo} is utilized to obtain a point cloud C_{exo}.

To project C_{exo} in the egocentric view, we need an exocentric-to-egocentric view translation matrix X, which can be computed through a transformation between P_{exo} and P_{ego}. To obtain P_{ego}, we build a 3D egocentric hand pose estimator ϕ_{ego}, which is designed with a simple architecture consisting of a ViT[3]-based backbone $\phi_{backbone}$ and an MLP-based regressor ϕ_{reg}. We optimize ϕ_{ego} with an L2 loss function. From P_{exo} and P_{ego}, we calculate X between them with the Umeyama algorithm [16], which estimates a transformation matrix. Therefore, we translate C_{exo} with X into C_{ego}, project it into egocentric view with an egocentric camera intrinsic parameters K_{ego}, and obtain the sparse egocentric RGB map S_{ego}.

Finally, T_{exo} is extracted with an off-the-shelf vision-language model (VLM) [1]. For example, when I_{exo} and user-provided question (*i.e., "Describe in detail about the scene and the object that the person is interacting with using their hands."*) are given, VLM outputs the corresponding answer T_{exo}.

2.2. Egocentric View Reconstruction

As shown in Fig. 1, egocentric view reconstruction Φ_{ego} reconstructs a dense and reliable egocentric image $\hat{I}_{ego}$ using exocentric observations S_{ego}, P_{ego}, and T_{exo} by leveraging the powerful latent diffusion model (LDM) [13]. Following the LDM, S_{ego} and P_{ego} are encoded into the latent embedding s_{ego} and p_{ego} using a frozen VAE encoder [4], random noises are added and denoised to s_{ego} and p_{ego}, and the denoised latent embedding is decoded into $\hat{I}_{ego}$ using the frozen VAE decoder.

During training, the ground-truth egocentric image I_{ego} is encoded to a clean latent z_0 through the VAE encoder, and the noise ϵ_t is added to z_0 to make a noisy embedding z_t with timestep t. By concatenating s_{ego}, p_{ego}, and z_t, we obtain 9-channel latent embedding z'_t, which is fed into the input of a pre-trained U-Net [15]. Simultaneously, a textual description T_{exo} is passed through CLIP [12] to obtain a text embedding c_{exo}, which serves as guidance for the U-Net of LDM. In this manner, the forward and reverse processes for denoising network ϵ_θ are carried out to predict ϵ_t with the L2 objective.

During sampling, we start the denoising process from a random Gaussian noise $z_T \sim \mathcal{N}(0, \mathbf{I})$ with well-trained ϵ_θ. We concatenate z_T with s_{ego} and p_{ego}, and feed to ϵ_θ to obtain the predicted latent $\hat{z}_0$ by reversing the schedule at each timestep t. We adopt classifier-free guidance (CFG) [6] to strengthen textual guidance. To the end, the final generated egocentric image $\hat{I}_{ego}$ is obtained from $\hat{z}_0$ by passing the VAE decoder.

Table 1. **Comparisons with state-of-the-arts on unseen scenarios (*i.e.,* objects, actions, scenes, and subjects) in H2O [7].** Compared to state-of-the-arts (*i.e.,* pix2pixHD [18], pixelNeRF [21], and CFLD [10]), *GenEgo* outperforms for all unseen scenarios in all metrics (*i.e.,* FID, PSNR, SSIM, and LPIPS).

Scenarios	Unseen Objects				Unseen Actions			
Methods	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓
pix2pixHD [18]	436.25	25.012	0.2993	0.6057	211.10	24.420	0.2854	0.6127
pixelNeRF [21]	498.23	26.557	0.3887	0.5372	251.76	27.061	0.3950	0.8159
CFLD [10]	59.615	25.922	0.4307	0.4539	50.953	28.529	0.4324	0.4593
GenEgo (Ours)	**41.334**	**31.171**	**0.4814**	**0.3476**	**33.284**	**31.620**	**0.4566**	**0.3780**
Scenarios	Unseen Scenes				Unseen Subjects			
Methods	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓
pix2pixHD [18]	490.32	18.567	0.2425	0.7290	452.13	18.172	0.3310	0.7234
pixelNeRF [21]	489.13	26.537	0.2574	0.7143	493.13	22.636	0.4135	0.6838
CFLD [10]	118.10	29.030	0.3696	0.6841	129.30	21.050	0.4001	0.6269
GenEgo (Ours)	**90.893**	**31.004**	**0.4096**	**0.6519**	**96.429**	**24.851**	**0.4605**	**0.6188**

3. Experiments

3.1. Setup

We evaluate *GenEgo* on H2O [7] and TACO [9], which provide synchronized multi-view data with 3D hand poses and depth. H2O contains diverse indoor videos with egocentric and exocentric views. Following [11], we assess generalization on four splits: (1) unseen objects, (2) unseen actions, (3) unseen scenes, and (4) unseen subjects. To further verify scalability, we test on TACO, which involves 196 objects and 15 actions, adopting an unseen actions setting for evaluation. In addition, we adopt common image quality metrics [8, 11]: FID [5], PSNR, SSIM [19], and LPIPS [23], covering perceptual fidelity, pixel-level accuracy, and human-perceived similarity.

3.2. Results

We compare with pix2pixHD [18], pixelNeRF [21], and CFLD [10]. On H2O across all four unseen splits, *GenEgo* achieves the best results on every metric as shown in Tab. 1. pix2pixHD and pixelNeRF yield noisy or blurry results due to their design assumptions (label-to-image or multi-view synthesis). CFLD shows stronger hand reconstructions but struggles with unseen objects and backgrounds. In contrast, *GenEgo* integrates diverse cues, *i.e.,* sparse maps, 3D poses, and text, to generate coherent and natural egocentric views, yielding large improvements (up to 30–35% in FID and consistent gains in PSNR, SSIM, LPIPS). As shown in Fig. 2, visual comparisons confirm that our method restores both hand-object details and background regions more faithfully. On TACO, which is more challenging than H2O due to its diversity of objects and actions, *GenEgo* maintains strong generalization, outperforming CFLD by reconstructing not only hands but also interacting objects and scene context as shown in Fig. 3.

To test in-the-wild generalization, we apply *GenEgo* to single exocentric images captured by a smartphone. As illustrated in Fig. 4, our approach generates realistic egocentric views, while CFLD produces unnatural outputs biased toward training data. Therefore, *GenEgo* effectively leverages sparse cues to generalize beyond curated datasets, demonstrating strong potential for real-world applications.

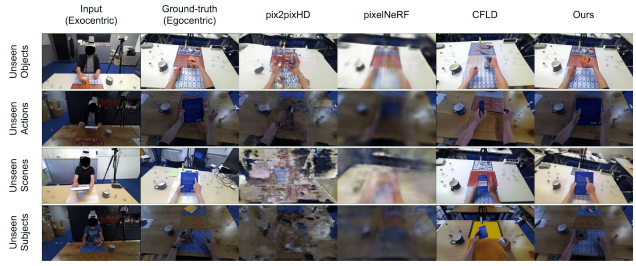

Figure 2. **Comparisons with state-of-the-arts on unseen scenarios (*i.e.,* objects, actions, scenes, and subjects) in H2O [7].** Compared to state-of-the-arts (*i.e.,* pix2pixHD [18], pixelNeRF [21], and CFLD [10]), *GenEgo* outperforms the reconstruction quality with respect to hand-object interaction and background regions for all unseen scenarios.

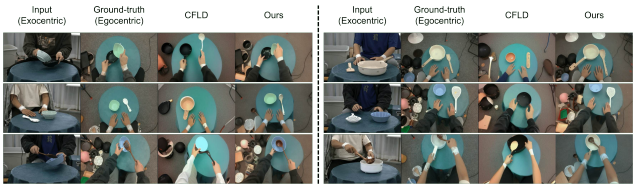

Figure 3. **Comparisons with state-of-the-art on unseen actions scenario in TACO [9].** Compared to state-of-the-art (*i.e.,* CFLD [10]), *GenEgo* outperforms the reconstruction quality with respect to hand-object interaction and background regions even on more challenging scenarios than H2O [7].

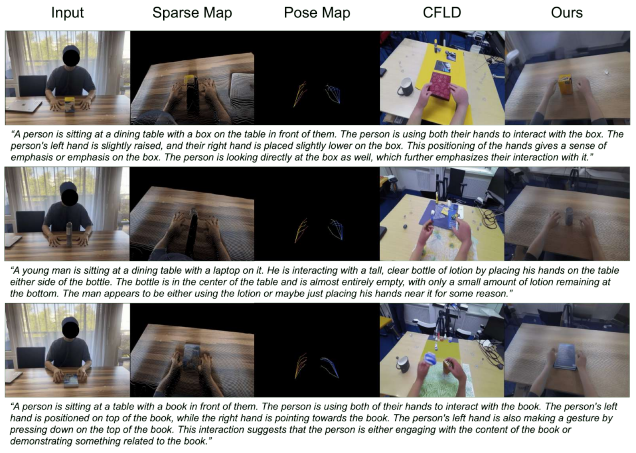

Figure 4. **Real-world comparisons with state-of-the-art.** Compared to state-of-the-art (*i.e.,* CFLD [10]), *GenEgo* significantly outperforms with respect to hand-object interaction and background regions for in-the-wild scenarios.

4. Conclusion

In this work, we propose *GenEgo*, a novel framework for translating exocentric observations into egocentric views using rich multimodal exocentric cues. Our two-stage approach first extracts exocentric observations, such as projected point clouds, 3D hand poses, and textual descriptions, and then generates a realistic egocentric image from a sparse egocentric map via a diffusion model conditioned on pose and text. Extensive experiments on the H2O and TACO benchmarks validate the effectiveness and superiority of *GenEgo*. Moreover, *GenEgo* shows powerful generalization ability on unlabeled real-world samples compared to state-of-the-art, and it implies *GenEgo* is enough to extend in-the-wild scenarios. These results demonstrate the potential of *GenEgo* as a robust and versatile solution for egocentric view generation from exocentric inputs, paving the way for future research in cross-view understanding and generation.

References

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 2

[2] 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 407–425, 2024. 1

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, G Heigold, S Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 2

[4] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, pages 12873–12883, 2021. 2

[5] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 30:6626–6637, 2017. 3

[6] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2

[7] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *CVPR*, pages 10138–10148, 2021. 1, 3

[8] Jia-Wei Liu, Weijia Mao, Zhongcong Xu, Jussi Keppo, and Mike Zheng Shou. Exocentric-to-egocentric video generation. *NeurIPS*, 37:136149–136172, 2024. 1, 3

[9] Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, and Li Yi. Taco: Benchmarking generalizable bimanual tool-action-object understanding. In *CVPR*, pages 21740–21751, 2024. 1, 3

[10] Yanzuo Lu, Manlin Zhang, Andy J Ma, Xiaohua Xie, and Jianhuang Lai. Coarse-to-fine latent diffusion for pose-guided person image synthesis. In *CVPR*, pages 6420–6429, 2024. 3

[11] 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, 2024. 1, 3

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Int. Conf. Mach. Learn.*, pages 8748–8763, 2021. 2

[13] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2

[14] Javier Romero, Dimitris Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM TOG*, 36(6), 2017. 2

[15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015. 2

[16] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE TPAMI*, 13 (04):376–380, 1991. 2

[17] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. *arXiv preprint arXiv:2503.11651*, 2025. 2

[18] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, pages 8798–8807, 2018. 3

[19] 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):600–612, 2004. 3

[20] Jilan Xu, Yifei Huang, Baoqi Pei, Junlin Hou, Qingqiu Li, Guo Chen, Yuejie Zhang, Rui Feng, and Weidi Xie. Egoexogen: Ego-centric video prediction by watching exo-centric videos. In *ICLR*, 2025. 1

[21] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *CVPR*, pages 4578–4587, 2021. 3

[22] Zhengdi Yu, Shaoli Huang, Fang Chen, Toby P. Breckon, and Jue Wang. Acr: Attention collaboration-based regressor for arbitrary two-hand reconstruction. In *CVPR*, 2023. 2

[23] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 3