A Appendix

A.1 Implementation Details

A.1.1 Egocentric View Reconstruction

To train the egocentric view reconstruction, we fine-tune a pre-trained LDM inpainting model (Rombach et al., 2022). Based on the PyTorch Lightning framework (Falcon & The PyTorch Lightning team, 2019), we set the training settings included a batch size of 3, a learning rate of 1×10^{-5}, and the AdamW optimizer (Loshchilov & Hutter, 2017), for a total of 5 epochs. All experiments are conducted on a single NVIDIA RTX 4090 GPU.

A.1.2 3D Egocentric Hand Pose Estimator

To train a 3D egocentric hand pose estimator from exocentric inputs, we adopt a backbone as ViT-224 (Dosovitskiy et al., 2020) and a regressor as MLP, which consists of two linear layers and one ReLU (Nair & Hinton, 2010) between linear layers. The input and output feature dimensions of the first linear layer are 768 and 512, and those of the last linear layer are 512 and 126. Based on the PyTorch framework (Paszke et al., 2019), we set the training settings included a batch size of 64, a learning rate of 1×10^{-4}, a criterion of MSE loss, and the Adam optimizer (Kingma & Ba, 2015), for a total of 100 epochs. All experiments were conducted on a single NVIDIA RTX 4090 GPU.

A.2 More Results

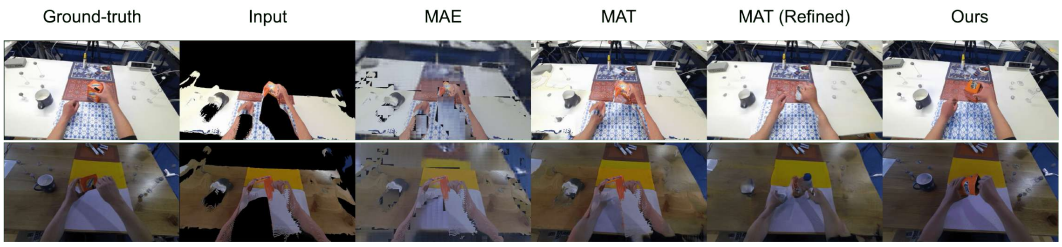

Figure A: **Results for backbones of egocentric view reconstruction.** Compared to backbones (*i.e.,* MAE (He et al., 2022) and MAT (Li et al., 2022)), LDM (Rombach et al., 2022) outperforms with respect to hand-object interaction and background regions for all cases.

A.2.1 Backbones of Egocentric View Reconstruction

Since egocentric view reconstruction closely resembles the image completion task, we compare our method with state-of-the-art image completion backbones, such as MAE (He et al., 2022), MAT (Li et al., 2022), and LDM (Rombach et al., 2022). Specifically, MAE specializes in mask-based image encoding, making it effective for filling missing pixel regions. MAT, a transformer-based model, excels at restoring large missing areas through long-range context modeling. LDM, serving as the baseline for *EgoWorld*, differs from the others in its ability to condition on diverse modalities such as text and pose. As shown in Fig. A, our LDM-based method reconstructs egocentric view images in a more natural and high-quality manner compared to other methods. Although the vanilla MAT model performs well in filling missing areas, it often struggles to maintain consistency with the surrounding content. For example, subtle differences in table color are noticeable. To address this, we develop a refined version of MAT that uses random

Table A: **Results for backbones of egocentric view reconstruction.** Compared to backbones (*i.e.,* MAE (He et al., 2022) and MAT (Li et al., 2022)), LDM (Rombach et al., 2022) outperforms in all metrics.

Backbones	FID↓	PSNR↑	SSIM↑	LPIPS↓
MAE (He et al., 2022)	169.91	24.623	0.4148	0.5041
MAT (Li et al., 2022)	89.933	28.922	0.4370	0.4758
MAT (Refined) (Li et al., 2022)	68.628	29.750	0.4731	0.4506
LDM (Rombach et al., 2022)	**41.334**	**31.171**	**0.4814**	**0.3476**

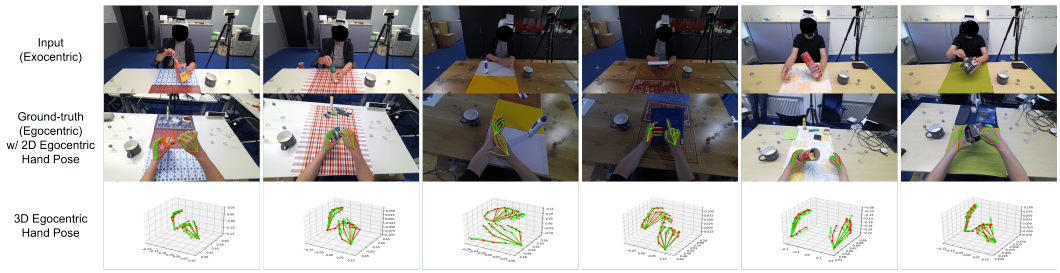

Figure B: **Results for 3D egocentric hand pose estimator.** Green and red poses indicate the ground-truth and estimated pose, respectively. Estimated poses are well-aligned with the ground-truth both in 2D and 3D spaces.

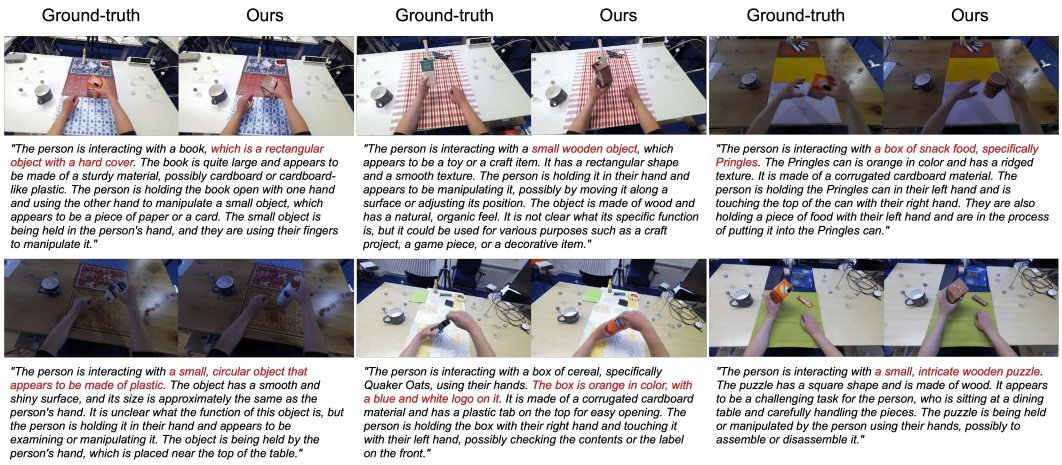

Figure C: **Results for incorrect textual description guidance of egocentric view reconstruction.** The red-colored texts represent incorrect descriptions, which are reflected as conditioning inputs for *EgoWorld* to generate egocentric images.

patch masking and recovery. However, this approach tends to fail in preserving detailed local interactions, such as hand-object interaction. In contrast, our LDM-based method, which operates by adding and removing noise in latent space, achieves coherent restoration not only in local regions but also in preserving consistency with existing areas. Moreover, as shown in Table A, our approach outperforms all other methods quantitatively across all evaluation metrics. Therefore, based on these results, we adopt LDM as the backbone for *EgoWorld*.

A.2.2 3D Egocentric Hand Pose Estimator of Exocentric View Observation

To validate the effectiveness of our newly proposed exocentric image-based 3D egocentric hand pose estimator, we conduct a qualitative analysis. As shown in Fig. B, given a single exocentric view image as input, our model predicts 3D hand poses that closely resemble the ground-truth. This demonstrates that the estimator is highly useful in the exocentric view observation stage for calculating the translation matrix, as well as in the egocentric view reconstruction stage for initializing the hand pose map.

A.2.3 Incorrect Textual Description Guidance of Egocentric View Reconstruction

To evaluate the effect of textual description guidance of the egocentric view reconstruction, we intentionally provide an incorrect textual description that does not match the exocentric image. As shown in Fig. C, the object in the egocentric view is generated to match the object described in the description. From this result, we observe two key insights: (1) the final egocentric image can vary depending on the output of the VLM, highlighting the importance of the VLM's performance; and

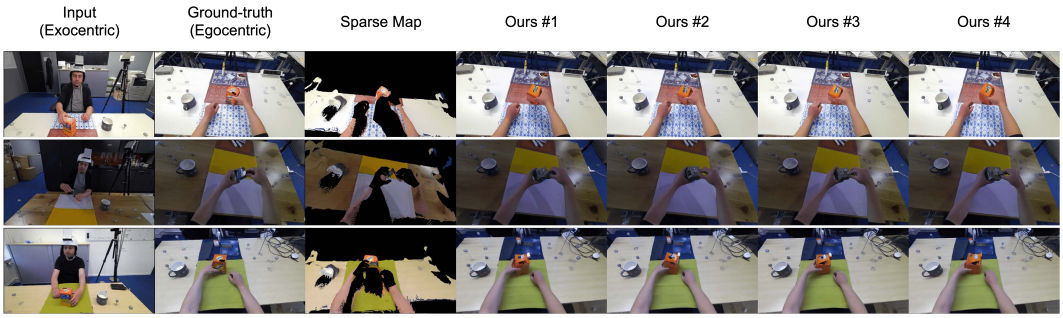

Figure D: **Results for generation consistency of egocentric view reconstruction.** With four iterations, the outputs are consistent, reliable, and similar to ground-truth.

(2) even when arbitrary exocentric images are fed, our model performed sufficient generalization to unseen scenarios.

A.2.4 Generation Consistency of Egocentric View Reconstruction

To evaluate the consistency of our generative model, we generated egocentric images multiple times under identical conditions. As shown in Fig. D, we present four outputs generated from the same exocentric image and corresponding sparse map, and our model consistently produces coherent egocentric images across runs. Despite the inherent variability in generative models, our method achieves stable and reliable exocentric-to-egocentric view translation, demonstrating its robustness and consistency.

A.2.5 Direct Camera Pose Regression of Exocentric View Observation

To compare the performance of the egocentric hand pose estimation and direct camera pose regression, we additionally build a camera pose regression model with ViT (Dosovitskiy et al., 2020) and MLP layers. Specifically, since it is targeted to estimate 4 $\times$ 4 relative pose (*i.e.,* exocentric-to-egocentric camera pose) estimation, we formulate it as 6D rotation representation for continuity in neural networks (Zhou et al., 2019). As shown in Tab. B, we test on H2O (Kwon et al., 2021) unseen action scenarios, and found little difference between these settings. Since we utilize the estimated egocentric hand pose to generate the natural and plausible hand-object interaction image in the egocentric view reconstruction stage, our approach has more advantages in terms of image quality.

Table B: **Results for direct camera pose regression of exocentric view observation.** The case of egocentric hand pose estimation showcases a higher score than that of direct camera pose regression.

Methods	FID↓	PSNR↑	SSIM↑	LPIPS↓
Direct Egocentric Camera Regression	44.907	27.821	0.4311	0.4809
Egocentric Hand Pose Estimation (Ours)	**44.323**	**28.897**	**0.4408**	**0.4590**

A.2.6 Whole-Body Pose Estimation of Exocentric View Observation

Instead of using hand pose estimation models, we review the whole-body pose estimation models (*e.g.,* Hand4Whole (Moon et al., 2022) or OSX (Lin et al., 2023)) and find that the performance was lower than that of the hand pose estimation as shown in Tab. C. In general, hand-object interaction situations in an exocentric view are often the case where a person is occluded by a desk or a table. However, since the hand is relatively visible, the performance is more robust than the whole-body case.

Table C: **Results for whole-body pose estimation of exocentric view observation.** The case of hand pose estimation showcases a higher score than that of whole-body pose estimation.

Methods	MPJPE (left hand) ↓	MPJPE (right hand) ↓
Whole-Body Pose Estimation	19.52	19.49
Hand Pose Estimation (Ours)	**1.005**	**1.161**

Table F: **Results for impact of individual sub-modules.** Whether using the ground-truth or not, *EgoWorld* outperforms baselines which use the ground-truth.

Methods	Pose	Depth	Text	FID↓	PSNR↑	SSIM↑	LPIPS↓
pix2pixHD (Wang et al., 2018)	GT	–	–	211.10	24.420	0.2854	0.6127
pixelNeRF (Yu et al., 2021)	GT (Camera)	–	–	251.76	27.061	0.3950	0.8159
CFLD (Lu et al., 2024)	GT	–	–	50.953	28.529	0.4324	0.4593
EgoWorld (Ours)	Prediction	Prediction	Prediction (Gemini Team et al. (2023))	42.323	28.897	0.4408	0.4590
	Prediction	GT	Prediction (Qwen-VL Bai et al. (2023))	41.198	29.002	0.4420	0.4379
	GT	Prediction	Prediction (Qwen-VL Bai et al. (2023))	37.040	30.017	0.4487	0.4092
	GT	GT	Prediction (Gemini Team et al. (2023))	34.891	30.998	0.4501	0.3820
	GT	GT	Prediction (Qwen-VL Bai et al. (2023))	**33.284**	**31.620**	**0.4566**	**0.3780**

A.2.7 Representations of Estimated Hand Pose

To examine the effect of MANO (Romero et al., 2017) representation for hand pose, we build an egocentric MANO parameter estimator based on ViT (Dosovitskiy et al., 2020) and MLP layers, and validate final results on the egocentric view reconstruction stage. As shown in Tab. D, we test on H2O (Kwon et al., 2021) unseen actions scenarios, and the trivial difference of performance on MANO is revealed. Although MANO representation contains richer visual information than keypoints, it does not exert a strong influence in the egocentric view reconstruction stage, as hand pose is fused with other modalities, *i.e.,* sparse maps and text descriptions.

Table D: **Results for representations of estimated hand pose.** The representation of the hand pose does not have a significant impact on performance.

Representations	FID↓	PSNR↑	SSIM↑	LPIPS↓
MANO (Romero et al., 2017)	33.208	31.632	0.4609	0.3771
Keypoints (Ours)	33.284	31.620	0.4566	0.3780

A.2.8 Robustness on Noisy Input

With our proposed pipeline, the heavy reliance on off-the-shelf estimators is likely to create error propagation vulnerabilities under occlusion or noisy inputs. Thus, we conduct additional experiments on how much the noisy input affects the final result. We newly define a noisy test set from H2O (Kwon et al., 2021) unseen actions scenario, which contains the cases causing incorrect depth or hand pose estimation (*e.g.,* occluded hands by object or hand, or blurry hand). We manually select hard cases. As shown in Tab. E, there was a slight deterioration in performance for the noisy cases, but it still achieved outstanding performance compared to other baselines. Although the off-the-shelf estimators may introduce some noise or slightly lower accuracy, our model demonstrates significantly greater robustness compared to other baselines. This indicates that even with current state-of-the-art estimators, our framework can produce reliable results. We expect even better performance in the future as estimation models continue to improve.

Table E: **Results for robustness on noisy input.** *EgoWorld* showcases robustness on noisy exocentric input and alleviates the heavy reliance on off-the-shelf estimators.

Test Sets	Methods	FID↓	PSNR↑	SSIM↑	LPIPS↓
All Cases	pix2pixHD (Wang et al., 2018)	211.10	24.420	0.2854	0.6127
	pixelNeRF (Yu et al., 2021)	251.76	27.061	0.3950	0.8159
	CFLD (Lu et al., 2024)	50.953	28.529	0.4324	0.4593
	EgoWorld (Ours)	**33.284**	**31.620**	**0.4566**	**0.3780**
Noisy Cases	pix2pixHD (Wang et al., 2018)	233.09	23.897	0.2612	0.6553
	pixelNeRF (Yu et al., 2021)	255.10	26.352	0.3892	0.8236
	CFLD (Lu et al., 2024)	52.879	27.090	0.4037	0.4701
	EgoWorld (Ours)	**34.910**	**30.284**	**0.4455**	**0.3835**

A.2.9 Impact of Sub-Modules of Exocentric View Observation

To evaluate the impact of individual sub-modules in the observation pipeline (*i.e., hand pose estimator, depth estimator, and vision-language model*), we conduct an experiment on H2O (Kwon et al., 2021) unseen actions scenario by distinguishing whether each sub-module (pose estimator, depth estimator, and VLM) is used or the ground-truth is used. Note that since there are no ground-truths for text description in the H2O dataset, we quantify the impact of VLM by comparing Qwen-VL (Bai et al., 2023), which we already adopted, with Gemini (Team et al., 2023), which is the popular foundation model. As shown in Tab. F, all prediction cases (last row) record the lowest score for all metrics. However, this case outperforms all state-of-the-art baselines, which use ground-truth hand

Figure E: **Results for additional comparisons with state-of-the-arts on unseen scenarios.** Compared to state-of-the-arts (*i.e.,* pix2pixHD (Wang et al., 2018), pixelNeRF (Yu et al., 2021), and CFLD (Lu et al., 2024))), *EgoWorld* outperforms for all unseen scenarios.

pose or camera pose. It implies although the performance of each sub-module is crucial, we expect the improvement of sub-modules will further increase our framework's performance in the future.

A.2.10 Additional Comparisons with State-of-the-Arts

We provide additional state-of-the-art comparisons in this appendix as shown in Fig. E. We evaluate our method across four unseen scenarios (*i.e.,* unseen objects, actions, scenes, and subjects) and observe that it consistently outperforms baseline models. pix2pixHD (Wang et al., 2018), which depends on label map-based image-to-image translation, generates egocentric images with significant noise; it implies pix2pixHD is ill-suited for tackling the exocentric-to-egocentric view translation task. Likewise, pixelNeRF (Yu et al., 2021), which is originally intended for novel view synthesis using multiple inputs, produces blurry results that lack fine-grained details; it means pixelNeRF is less effective for one-to-one view translation. On the other hand, CFLD (Lu et al., 2024), which focuses on generating view-aware person images using hand pose maps, shows better performance than the previous methods. However, its strengths are largely confined to hand region translation only, and it struggles to accurately reconstruct surrounding information like objects and scenes. In contrast, our approach, *EgoWorld*, produces robust and coherent results even in complex and previously unseen scenarios involving rich contextual elements. Therefore, we verify *EgoWorld*'s generalization ability across diverse, unseen situations.

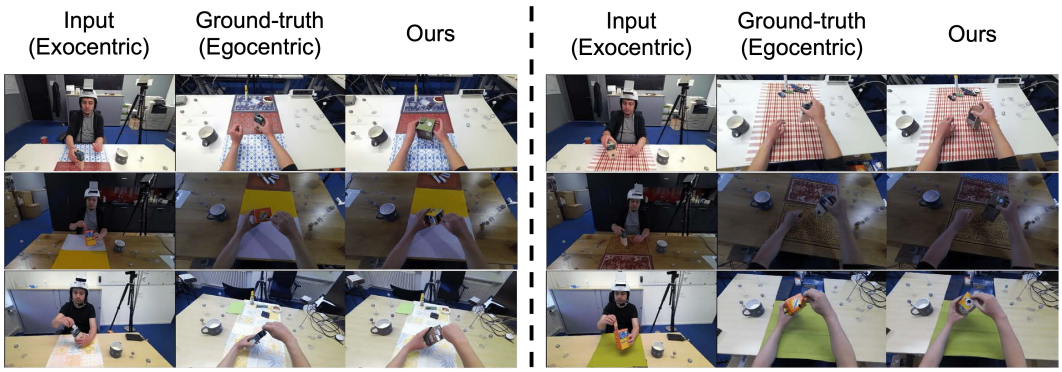

Figure F: **Results for failure examples in H2O (Kwon et al., 2021).** Subtle finger movements and dependency of VLMs make the reconstructed ouptuts of hands and objects quite unsatisfying.

A.2.11 Limitations and Future Work

In Fig. F, we illustrate failure examples on H2O (Kwon et al., 2021). In some cases, the reconstructed hand poses or objects are different from the ground-truth. For hand poses, subtle finger movements that are difficult to observe from the exocentric view are similarly hard to reproduce accurately in the egocentric view. We believe these limitations could be mitigated by developing a more advanced 3D egocentric hand pose estimator or by leveraging improved depth estimation to generate more reliable sparse maps, leading to better hand-aligned image completions. For objects, parts that are occluded or not visible in the exocentric image may appear distorted or inaccurately reconstructed in the egocentric view. Additionally, the object could be reconstructed incorrectly from the egocentric image when VLMs generates inaccurate text descriptions from the exocentric image. We anticipate that such issues can be addressed in the future with the development of more powerful VLMs.

REFERENCES

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.

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.

William Falcon and The PyTorch Lightning team. Pytorch lightning, 2019. URL `https://github.com/Lightning-AI/lightning`.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pp. 16000–16009, 2022.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015.

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

Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. Mat: Mask-aware transformer for large hole image inpainting. In *CVPR*, pp. 10758–10768, 2022.

Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3d whole-body mesh recovery with component aware transformer. In *CVPR*, pp. 21159–21168, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

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*, pp. 6420–6429, 2024.

Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Accurate 3d hand pose estimation for whole-body 3d human mesh estimation. In *CVPRW*, pp. 2308–2317, 2022.

Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 807–814, 2010.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 32, 2019.

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

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

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

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*, pp. 8798–8807, 2018.

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

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, pp. 5745–5753, 2019.