Put Myself in Your Shoes: Lifting the Egocentric Perspective from Exocentric Videos

Mi Luo[1], **Zihui Xue**[1,2], **Alex Dimakis**[1], **Kristen Grauman**[1,2]

[1]The University of Texas at Austin, [2]FAIR at Meta

We investigate exocentric-to-egocentric cross-view translation, which aims to generate a first-person (egocentric) view of an actor based on a video recording that captures the actor from a third-person (exocentric) perspective. To this end, we propose a generative framework called Exo2Ego that decouples the translation process into two stages: high-level structure transformation, which explicitly encourages cross-view correspondence between exocentric and egocentric views, and a diffusion-based pixel-level hallucination, which incorporates a hand layout prior to enhance the fidelity of the generated egocentric view. To pave the way for future advancements in this field, we curate a comprehensive exo-to-ego cross-view translation benchmark. It consists of a diverse collection of synchronized ego-exo tabletop activity video pairs sourced from three public datasets: H2O, Aria Pilot, and Assembly101. The experimental results validate that Exo2Ego delivers photorealistic video results with clear hand manipulation details and outperforms several baselines in terms of both synthesis quality and generalization ability to new actions.

Date: March 12, 2024

1 Introduction

Given a third-person video capturing a person opening a milk carton, what would the visual world look like from his perspective? See Figure 1. Due to mirror neurons in the human brain Ardeshir and Borji (2018b), we can easily envision the appearance and spatial relationships of the person's hands and the milk carton from a first-person perspective. However, existing computer vision models struggle to do the same thing, owing to the stark distinctions between the two viewpoints.

We make a step towards addressing this underexplored problem: exocentric-to-egocentric cross-view translation. The goal is to synthesize the corresponding ego view of an actor from an exo video recording[1], with minimal assumptions on the viewpoint relationships (e.g., camera parameters or accurate geometric scene structure). Specifically, we focus on synthesizing ego tabletop activities that involve significant hand-object interactions, such as assembling toys or pouring milk.

This task can benefit many applications, ranging from robotics to virtual and augmented reality (VR/AR). For example, an AR assistant could transform the view shown in a third-person how-to video to demonstrate to a user how things should look from their own perspectives, e.g., revealing finger placements and strumming techniques for a guitar video. Similarly, in robot learning, the robot could better match its actions to those of a human instructor in the room by projecting his/her hand-object interactions to its own perspective. Indeed, recent work demonstrates that the human ego view is valuable for robot learning Bharadhwaj et al. (2023); Bahl et al. (2022); Majumdar et al. (2023); Nair et al. (2022); Mandikal and Grauman (2021).

However, exo-to-ego view translation is highly challenging. It requires understanding the spatial relationships of visible hands and objects and inferring their pixel-level appearance in the novel ego view. Further, the task is not strictly geometric; it is inherently underdetermined. Some parts of an object may not be visible in the exo view—such as the inner pages of a book when only the cover is observed in the exo view—requiring the model to extrapolate the occluded parts. As a result, the recent popular geometry-aware novel view synthesis approaches Mildenhall et al. (2020); Niemeyer et al. (2022); Yu et al. (2021); Jang and Agapito (2021) are

[1]We use "ego" and "exo" as shorthand for egocentric (first-person) and exocentric (third-person).

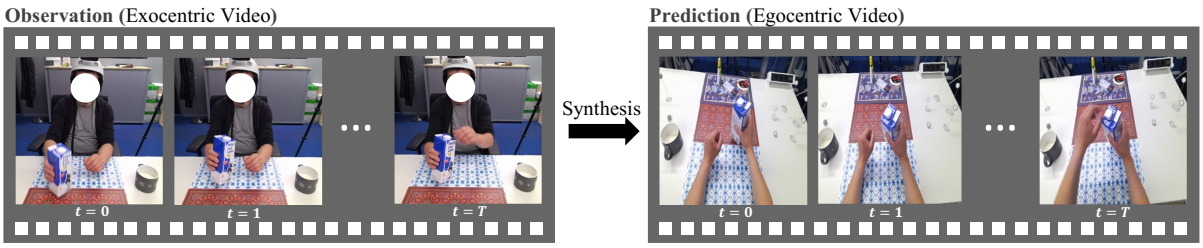

Figure 1 Cross-view translation from exocentric to egocentric video. Given an exocentric video sequence (top), the goal is to generate the corresponding egocentric perspective (bottom).

ill-equipped to solve this problem. The key reason is that they are regressive rather than generative, which limits their ability to deal with sparse input views (single exo view in our case) and major occlusions.

In light of this, we propose a probabilistic exo-to-ego generative framework (termed Exo2Ego) that learns the conditional distribution of the target ego video given the exo video. Exo2Ego differs from generative-based methods for general novel view synthesis Liu et al. (2021a); Ren and Wang (2022); Rombach et al. (2021); Wiles et al. (2020); Watson et al. (2022); Tseng et al. (2023); Chan et al. (2023) as it relaxes the requirement for camera parameters as input. This expands the potential application scenarios since camera parameters are rarely accessible in real-world scenarios, e.g., taking photos or videos with a mobile phone. Additionally, accurately inferring camera parameters remains challenging without precise correspondences or a rigid lab setting for calibration Wang et al. (2021). Moreover, Exo2Ego stands apart from recent cross-view image-to-image translation methods Tang et al. (2019); Ren et al. (2021), as it does not require the ground truth semantic map (object segmentation) of the target domain at the inference stage—typically an infeasible assumption.

Our key insights are to explicitly encourage cross-view correspondence by predicting the ego layout and to introduce a hand layout prior to boost the fidelity of generated hands in the ego view. Specifically, our Exo2Ego framework decouples the exo-to-ego view translation problem into two stages: (1) High-level Structure Transformation, which infers the rough location and interaction manner of the hands and objects in the ego view, using a transformer-based model to transform the exo hand-object interaction layout into its ego counterpart; and (2) Diffusion-based Pixel Hallucination, which trains a conditional diffusion model to refine the details on top of the ego hand-object interaction layout. Overall, Exo2Ego is a generative framework that offers a simple but effective baseline approach for the exo-to-ego view translation problem, accounting for the centrality of hand-object manipulations in this domain.

To evaluate our approach, we construct an exo-to-ego cross-view synthesis benchmark, comprising synchronized ego-exo video pairs curated from three datasets: H2O Kwon et al. (2021), Aria Pilot Lv et al. (2022), and Assembly101 Sener et al. (2022). Empirical results underscore the efficacy of our Exo2Ego framework, which produces realistic video outputs with distinct hand manipulation details. Exo2Ego surpasses single-view translation baselines Wang et al. (2018b,a), a recent cross-view synthesis approach Liu et al. (2020), as well as a NeRF-based method Yu et al. (2021), demonstrating superior generation quality and a marked improvement in generalization capability to new actions.

2 Related work

Relating Egocentric and Exocentric Views There have been several attempts to jointly understand ego and exo views Elfeki et al. (2018); Sigurdsson et al. (2018b); Grauman et al. (2023). Early efforts explore how to localize the person wearing a camera in a third-person view, given their egocentric video Ardeshir and Borji (2016, 2018a); Fan et al. (2017); Xu et al. (2018); Wen et al. (2021). To bridge the ego-exo gap, other work learns view-invariant features from concurrent (paired) Ardeshir and Borji (2018b); Sigurdsson et al. (2018a); Sermanet et al. (2018); Yu et al. (2019, 2020) or unpaired Xue and Grauman (2023) views, or boosts the latent ego signals in exo video during pretraining Li et al. (2021), summarization Ho et al. (2018), and 3D pose estimation Wang et al. (2022). Additionally, fusing ego and exo views can improve action recognition Soran et al. (2015) and robotic manipulation tasks Jangir et al. (2022).

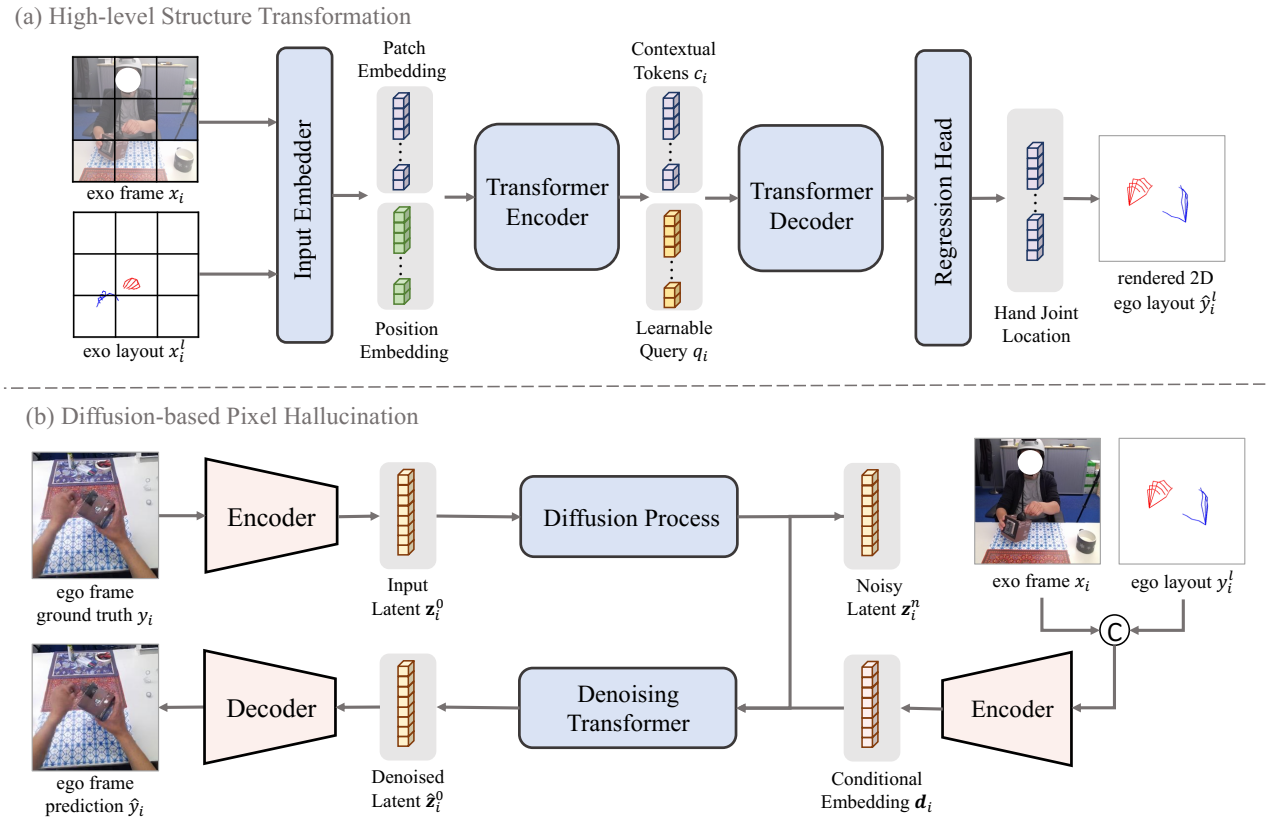

Figure 2 Our Exo2Ego framework comprises two modules: (a) High-level Structure Transformation, which predicts the ego layout, capturing hand position and interactions using a transformer-based encoder-decoder architecture. (b) Diffusion-based Pixel Hallucination, which enhances pixel-level details on top of the ego layout using a conditional diffusion model.

Despite its significance, cross-view translation has received little attention. P-GAN Liu et al. (2020) and STA-GAN Liu et al. (2021b) have explored cross-view image and video synthesis, respectively, yet they only examine basic activities such as walking, jogging, and running, limiting their applicability to more diverse scenarios. To address these limitations, we introduce a new benchmark for exo-to-ego view synthesis covering some diverse activities, from assembling toys to manipulating everyday objects. Importantly, we relax STA-GAN's requirement for the ego semantic map during test time Liu et al. (2021b), enhancing the practical applicability.

Novel View Synthesis Exo-to-ego view translation also relates to novel view synthesis, i.e., given a set of images of a scene, infer how the scene looks from novel viewpoints. State-of-the-art geometry-aware approaches like Scene Representation Networks (SRN) Sitzmann et al. (2019) and Neural Radiance Fields (NeRF) Mildenhall et al. (2020); Niemeyer et al. (2022); Yu et al. (2021); Jang and Agapito (2021) learn 3D representations from 2D images and camera parameters and perform differentiable neural rendering from one or a few input images at inference. Although they excel at interpolating near-input views, they are typically not well-suited for the large view changes required for exo-to-ego view synthesis. Trained purely with regression objectives, they have limited capacity to deal with sparse inputs, ambiguity caused by occluded parts of the scene, and long-range extrapolations caused by drastic camera transformations Watson et al. (2022); Chan et al. (2023)—all challenges that are particularly prominent in exo-to-ego view translation. Alternatively, generative-based novel view synthesis methods Liu et al. (2021a); Ren and Wang (2022); Rombach et al. (2021); Wiles et al. (2020); Watson et al. (2022); Tseng et al. (2023); Chan et al. (2023) use generative priors to synthesize random plausible outputs from the conditional distribution, making them adept at handling ambiguity and extrapolating to occluded parts of the scene. Our Exo2Ego framework belongs to the generative-based vein, but unlike all the above, it does not depend on camera poses as critical conditional inputs. This improves scalability and widens the scope of potential applications, as rigid camera parameters are seldom accessible in real-world scenarios and are challenging to infer solely from RGB images.

Cross-view Image-to-Image Translation Exo-to-ego view synthesis is inherently an image-to-image translation problem Isola et al. (2017); Wang et al. (2018b,a); Bansal et al. (2018)—but with drastic viewpoint change. Prior work centers around aerial-to-ground translation Regmi and Borji (2019, 2018); Tang et al. (2019); Ren et al. (2021) or generating an egocentric view without human-object interactions, i.e. a person walking indoors Tang et al. (2019); Liu et al. (2020, 2021b). Except for P-GAN Liu et al. (2020), all of these approaches require the ground truth semantic map of the target domain at training or inference time to ensure high-quality generation—an unrealistic assumption that means a good portion of the translation task (the boundaries of the objects in the unobserved ego view or the input exo view) is already known, Compared with P-GAN, our Exo2Ego exhibits significantly better qualitative and quantitative results.

Diffusion Models (DMs) have shown competitive performance in image synthesis Ho et al. (2020); Song et al. (2021); Dhariwal and Nichol (2021); Ho et al. (2022), but their use for cross-view translation remains unexplored. We develop an exo-to-ego translation framework incorporating the diffusion model over GANs, considering potential advantages such as addressing mode collapse, promoting sample diversity, and maintaining stable training dynamics for improved image generation Rombach et al. (2022).

3 Methodology

First, we formalize our problem (Sec. 3.1). Then we present the two stages of our model (Sec. 3.2 and 3.3), followed by an overview of training and inference (Sec. 3.4).

3.1 Problem Formulation

Whereas conventional geometry-aware approaches Mildenhall et al. (2020); Yu et al. (2021) generate novel views by specifying camera poses and performing volume rendering, Exo2Ego offers a different perspective. We propose tackling the exo-to-ego translation problem in a purely probabilistic framework. Intuitively, this is essential to address the inherent ambiguity in predicting the ego view from the exo view, e.g., due to entirely unseen portions of objects or the human body.

We define $\mathbf{X}_T = \{x_1, \ldots, x_i, \ldots, x_T\}$ as a series of T video frames captured from an exo viewpoint, characterized by static backgrounds with dynamic actors and other objects present, where i denotes the time index. This exo perspective reveals the actors' motions and (potentially) full-body poses within the scene. Let $\mathbf{Y}_T = \{y_1, \ldots, y_i, \ldots, y_T\}$ denote the corresponding sequence of ego frames captured from a first-person perspective. This perspective emulates the view of a camera mounted on the head or body of the actor, with a focus on their actions and interactions.

The goal of exo-to-ego view translation is to simulate the view of the ego camera-wearer in the scene recorded by an exo camera. Formally, we seek a translation model that can map $\mathbf{X}_T$ into a series of output exo frames, $\hat{\mathbf{Y}}_T = \{\hat{y}_1, \ldots, \hat{y}_i, \ldots, \hat{y}_T\}$. As shown in Eqn. (1), the conditional distribution of $\hat{\mathbf{Y}}^T$ given $\mathbf{X}^T$ should be indistinguishable from the conditional distribution of $\mathbf{Y}^T$ given $\mathbf{X}^T$.

$$p(\hat{\mathbf{Y}}_T|\mathbf{X}_T) = p(\mathbf{Y}_T|\mathbf{X}_T). \tag{1}$$

We prioritize translating daily tabletop activities from the exo to ego view, which is a prevalent setting in egocentric learning and frequently entails extensive interactions between hands and objects. This problem is highly challenging as the translation model must produce photorealistic hand-object interaction sequences in the ego view while also performing geometric and semantic reasoning, i.e., correctly predicting the spatial location and pixel-level appearance of visual concepts. Furthermore, it requires linking the actor's head/body motion captured by the exo camera with the viewpoint change in the ego video. Upon examining the application of the widely used pixel-to-pixel generation method Wang et al. (2018b) for our task, we found it struggles with intricate details of the hands, likely because treats all pixels equally and lacks geometric correspondence between views. At the same time, we know that (at least in manipulation-rich scenarios) the hands of the actor are a common ground between the ego and exo views, despite their many other differences.

Motivated by these points, we propose the Exo2Ego framework, which disentangles the understanding of cross-view correspondences and pixel-level synthesis. It consists of two key modules, as illustrated in Figure 2: (1) High-level Structure Transformation, which addresses the task of inferring the location and interaction

manner of hands and objects in the ego view. To accomplish this, we train a transformer-based encoder-decoder model to translate the exo frame into an ego hand-object interaction layout (detailed below). (2) Diffusion-based Pixel Hallucination, which learns to synthesize realistic and high-quality pixel-level details by training a conditional diffusion model operating on top of the ego hand layout.

3.2 High-level Structure Transformation

Given an exo frame, the purpose of the high-level structure transformation is to train a layout translator which predicts the ego layout showing the location and rough contour of the visual concepts. Specifically, to capture fine-grained hand-object interaction details, we propose to generate the hand layout, instantiated as 2D hand poses. The quality of the generated layout is crucial, as it serves as a crucial reference for further pixel-level hallucination. To achieve this, we draw inspiration from a recent study Rombach et al. (2021) that highlights transformer-based architectures Vaswani et al. (2017); Dosovitskiy et al. (2020) succeed in understanding cross-view correspondence, due to their reduced locality-bias. Considering transformers have the potential to represent the geometric information implicitly, we employ a pure transformer-based encoder-decoder architecture, which enables us to effectively incorporate and process the comprehensive context of the exo scene, and thus help generate a precise ego hand layout.

Exo Contextual Feature Extraction As shown in Figure 2, an exo frame $x_i \in \mathbb{R}^{H\times W\times C_1}$ concatenated with the corresponding exo layout $x_i^l \in \mathbb{R}^{H\times W\times C_2}$ (2D hand pose layout recorded as image size) is first split into a sequence of patches and then processed by an input embedding layer, such as patch embedding for ViTs Dosovitskiy et al. (2020) to get a tokenized input sequence $e_i \in \mathbb{R}^{M\times D}$. To capture the positional information which is critical for spatial relationship reasoning, we add learnable position embedding to the sequence of patches. Then the sequence of tokens is passed into a transformer encoder consisting of N repeated blocks. Each block contains a multi-head attention token mixer to first communicate information among tokens, which is represented as:

$$z_i^{j-1} = \text{Attention}(\text{Norm}(e_i^{j-1})) + e_i^{j-1}, \tag{2}$$

where $j \in \{1, ..., N\}$ and $\text{Norm}(\cdot)$ is the normalization method. Then, the mixed token z_i^{j-1} is passed into a two-layer MLP expressed as follows:

$$e_i^j = \text{MLP}(\text{Norm}(z_i^{j-1})) + z_i^{j-1}. \tag{3}$$

After iterating through all N blocks, the input sequence e_i undergoes a mapping process that transforms it into the final contextual tokens $c_i \in \mathbb{R}^{M\times D}$. We talk about how to decode the contextual tokens into the desired ego layout in the following part.

Layout Decoder Consider a hand pose image with shape $\{H \times W \times 3\}$ as the target ego layout. The contextual tokens c_i should be decoded into 2D hand joint coordinates and then rendered to the RGB image. First, the contextual tokens c_i are combined with a sequence of learnable query embeddings $q_i \in \mathbb{R}^{E\times D}$ and passed into a decoder architecture, consisting of K transformer blocks, as described earlier. Following the transformer blocks, a regression head is employed to estimate the 2D coordinates of the joints whose range is restricted to $[0, 1]$. Then we apply the classic bipartite matching loss Carion et al. (2020); Redmon et al. (2016) to map the regressed hand joints to the ground truth joints.

3.3 Diffusion-based Pixel Hallucination

Having inferred the ego hand pose layout y_i^l, next the aim of pixel-level hallucination is to synthesize photorealistic ego frames by taking into account an exo frame x_i and the target ego layout. We use the diffusion formulation proposed in the denoising diffusion probabilistic model (DDPM) Ho et al. (2020) and train the diffusion model in the latent space. As shown in Figure 2, we first adopt a pre-trained variational autoencoder (VAE) model Kingma and Welling (2013) as used in Rombach et al. (2022) to encode the original ego frames y_i and conditional information exo frame x_i and ego layout y_i^l into the latent space. Then, we train a diffusion transformer Peebles and Xie (2022) to learn the latent data distribution by denoising a latent

vector sampled from a Gaussian distribution gradually. Specifically, given an initial noise map $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and a conditioning vector $\mathbf{d}$, the diffusion model generates the corresponding ego latent $\hat{\mathbf{z}}_\theta$. A squared error loss is used to denoise a variably-noised latent code $\mathbf{z}^n = \alpha^n \mathbf{z} + \sigma^n \boldsymbol{\epsilon}$ as follows:

$$\mathcal{L} = \mathbb{E}_{\mathbf{d},\mathbf{z},\boldsymbol{\epsilon},n}\left[\|\mathbf{z} - \hat{\mathbf{z}}_\theta(\alpha^n \mathbf{z} + \sigma^n \boldsymbol{\epsilon}, \mathbf{d})\|_2^2\right] \tag{4}$$

where $\mathbf{z}$ is the ground-truth latent vector, $\mathbf{d}$ is the conditioning vector, and α^n, σ^n are functions of the diffusion process time step n, which control the noise schedule and sample quality. We simply concatenate the noisy latent vector $\mathbf{z}^n$ and the conditional embedding $\mathbf{d}$ as the input of the denoising transformer. In contrast to earlier cross-view translation methods based on GANs Wang et al. (2018b,a); Liu et al. (2020), our conditional diffusion model sequentially updates the target ego outputs, excels in capturing intricate ego-exo dependencies and faithfully reproducing the mapping of the conditional distributions in Eqn. (1). Notably, our model exhibits enhanced stability throughout the training and consistently produces higher-quality samples, as verified by our experiments.

Our diffusion model operates on a per-frame basis, independent of any previously generated ego frames or observed exo frames from the past. However, our Exo2Ego framework can be seamlessly integrated into video-to-video synthesis techniques, such as vid2vid Wang et al. (2018a), to enhance the temporal coherence in the resulting videos. For example, Exo2Ego can generate the initial ego frame which could serve as the initialization of vid2vid's sequential generation.

3.4 Training and Inference

At the training stage, the two modules described in Section 3.2 and 3.3 are trained separately. During inference, we first generate the corresponding ego hand layout prediction $\hat{y}_i^l$ for each input exo frame x_i with the layout translator described in Section 3.2. Subsequently, $\hat{y}_i^l$ is concatenated with x_i in a channel-wise manner and then used as the conditional input of the pixel-level diffusion generator described in Section 3.3.

4 Experimental Evaluation

We introduce an exo-to-ego synthesis benchmark (Sec. 4.1) and then validate our approach against state-of-the-art synthesis methods (Sec. 4.2).

4.1 A Benchmark for Exo-to-Ego View Translation

To facilitate new work on the exo-to-ego translation task, we contribute a new cross-view synthesis benchmark sourced from three public time-synchronized multi-view datasets: H2O Kwon et al. (2021), Aria Pilot Lv et al. (2022), and Assembly101 Sener et al. (2022). **H2O** 🙌 Kwon et al. (2021) has indoor videos featuring actors manipulating 8 different objects using both hands. **Assembly101** 🚚 Sener et al. (2022) shows people (dis)assembling 101 take-apart toy vehicles. We take 6 activity sequences involving different individuals manipulating a toy roller. **Aria Pilot** 👓 Lv et al. (2022) provides 16 multi-view recordings of an actor wearing Project Aria glasses Somasundaram et al. (2023) manipulating YCB Calli et al. (2015) objects on a desktop.

For all datasets, we select the exo camera viewpoints that offer the clearest capture of the action ('cam 2' for H2O, '214-7' for Aria Pilot, and 'v4' for Assembly101). All frames are cropped and resized into 256×256. We divide each video into 30-frame clips. Overall, this dataset showcases an array of tabletop activities with various objects, from toy assembly to the manipulation of everyday objects. More details are provided in Appendix.

We employ the following split settings to evaluate four kinds of generalization: (1) **new actions**, where we put the first 80% of a video's clips into the training set and the remaining 20% in the test set, hence splitting up action steps over time (e.g., for a video depicting picking up a coffee box and then taking coffee capsules out, the former will be assigned to the training set and the latter to the test set); (2) **new objects**, where we train with videos involving any of 7 objects and test with videos containing a novel 8th object; (3) **new subjects**, where we train with all clips from one subject and test on clips from another (subject 1 $\rightarrow$ subject 2); (4) **new scenes (backgrounds)**, where we train and test with disjoint scenes (scene h1, h2, k1, k2 $\rightarrow$ scene o1, o2). We

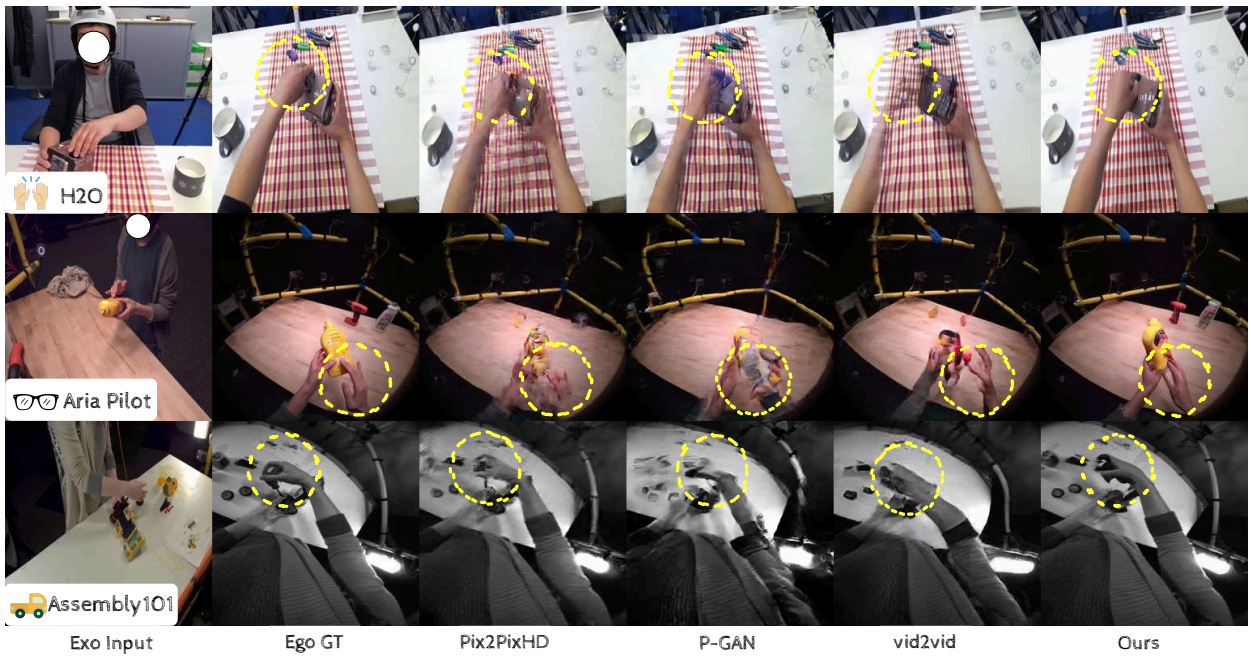

Figure 3 Qualitative examples when generalizing to new actions on all datasets. More examples are in Appendix.

analyze new action generalization on all 3 datasets, and the rest on H2O only, as it is the only dataset with the variations and meta-data to allow such split settings. For action generalization, H2O, Aria Pilot, and Assembly101 have 704, 343, and 682 training clips, and 199, 95, and 205 test clips, respectively. For other generalization settings (new objects, subjects, and scenes), there are 691, 704, and 591 training clips, and 108, 194, and 312 test clips.

4.2 Evaluating Synthesis Quality

With this well-constructed benchmark at hand, we proceed to validate our Exo2Ego alongside baseline models.

Baselines We consider several baselines: (1) Pix2PixHD Wang et al. (2018b), a single-view image translation method that processes the videos frame by frame; (2) P-GAN Liu et al. (2020), a recent exo-to-ego view translation method that proposes a parallel generative network to facilitate cross-view image translation; (3) Vid2Vid Wang et al. (2018a), a single-view video translation method that models the temporal dynamics in the video; (4) pixelNeRF Yu et al. (2021), a NeRF-like model known for its remarkable generalization capabilities and its accommodating stance toward sparser input views. Note that pixelNeRF assumes camera parameters are known, while other methods do not. We conduct experiments with pixelNeRF only on H2O, as the other two datasets lack the camera information needed by pixelNeRF.

Implementation Details For Exo2Ego's high-level structure transformation transformer, the number of blocks N is set to 6. We use DiT-XL/2 Peebles and Xie (2022) as the denoising architecture for Exo2Ego. Our diffusion model is trained for 40,000 steps for all datasets. pixelNeRF is trained for 70,000 steps for all datasets. For H2O and Assembly101, pix2pixHD, and P-GAN are trained for 100 epochs, and vid2vid is trained for 40 epochs. For Aria Pilot, we use 400 and 100 epochs respectively. More details are in Appendix.

Evaluation Metrics We measure the quality of the synthesized videos as follows. First, we evaluate the feasibility of generated hands, denoted as Feasi. We adopt an off-the-shelf egocentric hand object detector Shan et al. (2020) to measure the physical feasibility of the synthesized hands. This metric is calculated as the confidence score of hand detection results produced by the detector, averaged over all synthesized frames. The purpose of this metric is to check whether the synthesized hands are realistic enough to be faithfully detected. It's noteworthy that Ye et al. (2023) also employs the confidence score of the hand as an evaluation metric, but they focus on in-contact confidence rather than the hand detection confidence utilized in our approach. We also use structural similarity (SSIM) and peak signal-to-noise ratio (PSNR) Hore and Ziou (2010) to

	H2O						
Method	SSIM↑	PSNR↑	FID↓	P_{Squeeze}↓	P_{Alex}↓	P_{Vgg}↓	Feasi.↑
pix2pixHD Wang et al. (2018b)	0.428	30.370	132.03	0.163	0.220	0.337	0.8952
P-GAN Liu et al. (2020)	0.272	29.079	192.26	0.242	0.308	0.449	0.8353
vid2vid Wang et al. (2018a)	0.402	29.882	85.76	0.166	0.226	0.342	0.9116
pixelNeRF* Yu et al. (2021)	0.219	27.871	470.13	0.711	0.807	0.731	0.0086
Exo2Ego (Ours)	**0.433**	**30.564**	**38.03**	**0.128**	**0.177**	**0.295**	**0.9758**
	Aria Pilot						
Method	SSIM↑	PSNR↑	FID↓	P_{Squeeze}↓	P_{Alex}↓	P_{Vgg}↓	Feasi.↑
pix2pixHD Wang et al. (2018b)	0.350	29.553	159.91	0.304	0.362	0.480	0.0790
P-GAN Liu et al. (2020)	0.343	29.769	122.40	0.287	0.345	0.475	0.1014
vid2vid Wang et al. (2018a)	0.359	29.858	52.38	0.266	0.328	0.443	0.2646
Exo2Ego (Ours)	**0.371**	**29.952**	**26.01**	**0.245**	**0.305**	**0.421**	**0.5869**
	Assembly101						
Method	SSIM↑	PSNR↑	FID↓	P_{Squeeze}↓	P_{Alex}↓	P_{Vgg}↓	Feasi.↑
pix2pixHD Wang et al. (2018b)	0.405	29.909	112.71	0.216	0.295	0.401	0.1184
P-GAN Liu et al. (2020)	0.372	29.303	114.68	0.216	0.300	0.436	0.0800
vid2vid Wang et al. (2018a)	0.368	29.481	78.47	0.224	0.312	0.424	0.1358
Exo2Ego (Ours)	**0.406**	**30.027**	**33.16**	**0.178**	**0.254**	**0.365**	**0.3791**

Table 1 Generalizing to new actions on three datasets. *pixelNeRF requires privileged camera information relative to the other baseline models.

check the pixel-level similarity between the predicted ego frame and the ground truth (GT) ego frame. The evaluation metrics should ideally remain stable under minor transformations (such as small translations or affine transformations). So we further adopt perceptual metrics, including FID Seitzer (2020) and Learned Perceptual Image Patch Similarity (LPIPS) Zhang et al. (2018) which evaluates the feature-level similarity of the predicted ego frame and the ground truth (GT) ego frame. LPIPS are denoted as P_{Squeeze}, P_{Alex}, and P_{Vgg}, which uses SqueezeNet Iandola et al. (2016), AlexNet Krizhevsky et al. (2017), and VGG Simonyan and Zisserman (2014) as the feature extractor, respectively.

Generalization to New Actions Figure 3 showcases qualitative results on all datasets, demonstrating the superiority of our Exo2Ego framework. Compared to all baselines, our approach produces realistic hands with correct poses, especially noticeable in the highlighted yellow circle regions when dealing with new actions during test time. Table 1 presents the feasibility of synthesized hands and the various measures of pixel-level similarity and perceptual similarity between the predicted ego frame and the ground truth ego frame for all datasets. Our Exo2Ego greatly increases the feasibility of synthesized hands, with a confidence gain up to 32.23% on Aria Pilot, compared with the best score of 26.46% achieved by vid2vid. This improvement indicates enhanced realism in the generated hands, resulting in greater confidence in hand detection results. Additionally, our Exo2Ego can produce ego frames with higher pixel-level similarity to the ground truth frames and better perceptual metrics (much lower FID scores). This highlights the benefits of the 'decoupling' idea of our Exo2Ego framework in generating photorealistic ego frames.

Generalization to New Objects, Subjects, and Backgrounds Illustrated in Table 2, our Exo2Ego outperforms the baselines in terms of quantitative metrics for generalization to new objects/subjects/backgrounds. Figure 4 shows that our Exo2Ego can produce realistic hand-object interactions when encountering exo views with new distributions, while other baselines fail miserably. This is mainly because our diffusion-based generative mechanism gradually denoises the target ego view, capturing the inherent intricate mapping from exo to ego views better. Another observation is that pixelNeRF produces ego views characterized by noticeable blurring and a lack of fine details in hands and objects. This observation supports our earlier discussions on the limitations of geometry-aware synthesis approaches.

Modeling Egocentric Viewpoint Change Due to the probabilistic nature of our Exo2Ego framework, it excels in capturing the relationship between the viewpoint changes in the ego view and the head motions observed in the exo view, holding the potential to deliver more immersive virtual experiences. Intrinsically, our Exo2Ego

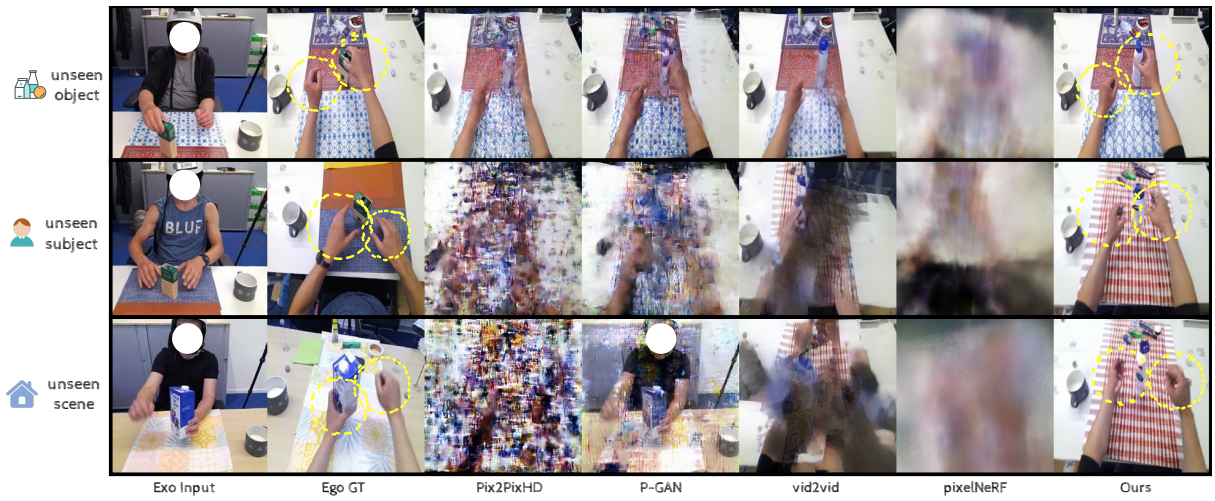

Figure 4 Qualitative examples when generalizing to new objects, subjects, and scenes (backgrounds) on H2O dataset.

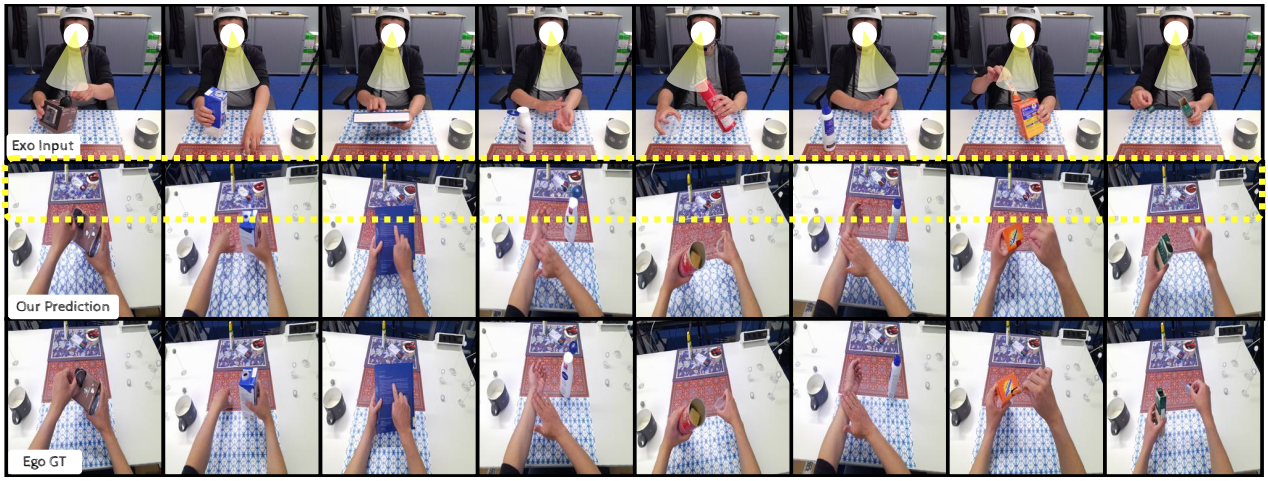

Figure 5 Exo2Ego framework generates ego videos with reasonable viewpoint changes.

framework leverages the fact that ego-camera pose is conditioned on the reference exo view, as the ego-camera wearer is usually visible in the exo view. This is highlighted by the yellow gaze direction and dashed box in Figure 5. As the actor moves their head when interacting with objects, the slope of the desk edge in the ego view also changes, indicating a corresponding change in ego viewpoint. This implies the ego-camera motion can be implicitly inferred via conditional generative modeling, even if it's not enforced explicitly. The ability to model ego viewpoint changes is crucial, especially in physical activities like playing ball games which involve heavy head and body motions.

5 Limitations and Future Work

As an early-stage effort for exo-to-ego cross-view translation, our work makes non-trivial progress in producing reasonable hand manipulation details for novel actions. However, despite establishing stronger generalizations than the existing models, our model does not generalize perfectly to in-the-wild objects, subjects, and backgrounds, due to the modest scale of available training data. Moreover, our qualitative results expose limitations in existing baselines and our Exo2Ego framework in generating 3D-consistent views for objects during test time, attributed to the absence of geometric priors for common objects, see Figure 6. Future work involves integrating robust object geometric priors. While leveraging predicted camera parameters for encoding 3D geometry is a promising approach, as shown Figure 4, the popular 3D-aware method pixelNeRF Yu et al. (2021) faces challenges in the dynamic exo-ego setting characterized by hand and object interactions, due

	H2O-unseen object						
Method	SSIM↑	PSNR↑	FID↓	P_{Squeeze}↓	P_{Alex}↓	P_{Vgg}↓	Feasi.↑
pix2pixHD Wang et al. (2018b)	0.333	29.504	253.73	0.269	0.360	0.467	0.7527
P-GAN Liu et al. (2020)	0.270	28.915	264.76	0.277	0.350	0.492	0.7720
vid2vid Wang et al. (2018a)	**0.349**	29.585	145.39	0.242	0.324	0.428	0.8612
pixelNeRF Yu et al. (2021)	0.262	27.912	473.73	0.750	0.841	0.736	0.0000
Exo2Ego (Ours)	0.332	**29.701**	**102.58**	**0.228**	**0.308**	**0.419**	**0.9876**
	H2O-unseen subject						
Method	SSIM↑	PSNR↑	FID↓	P_{Squeeze}↓	P_{Alex}↓	P_{Vgg}↓	Feasi.↑
pix2pixHD Wang et al. (2018b)	0.069	28.028	477.47	0.661	0.754	0.765	0.0140
P-GAN Liu et al. (2020)	0.088	28.063	451.37	0.623	0.708	0.725	0.2002
vid2vid Wang et al. (2018a)	0.136	28.094	308.52	0.564	0.690	0.686	0.3838
pixelNeRF Yu et al. (2021)	0.170	27.874	459.49	0.769	0.836	0.768	0.0000
Exo2Ego (Ours)	**0.178**	**28.266**	**150.87**	**0.503**	**0.629**	**0.665**	**0.9666**
	H2O-unseen scene						
Method	SSIM↑	PSNR↑	FID↓	P_{Squeeze}↓	P_{Alex}↓	P_{Vgg}↓	Feasi.↑
pix2pixHD Wang et al. (2018b)	0.024	27.918	558.87	0.729	0.789	0.788	0.0013
P-GAN Liu et al. (2020)	0.085	27.928	325.34	0.608	**0.699**	0.724	0.6041
vid2vid Wang et al. (2018a)	0.033	27.942	398.78	0.636	0.731	0.734	0.2095
pixelNeRF Yu et al. (2021)	0.103	27.810	506.10	0.773	0.908	0.800	0.0000
Exo2Ego (Ours)	**0.157**	**27.943**	**249.04**	**0.531**	0.709	**0.658**	**0.9635**

Table 2 Generalizing to different objects, subjects, and scenes.

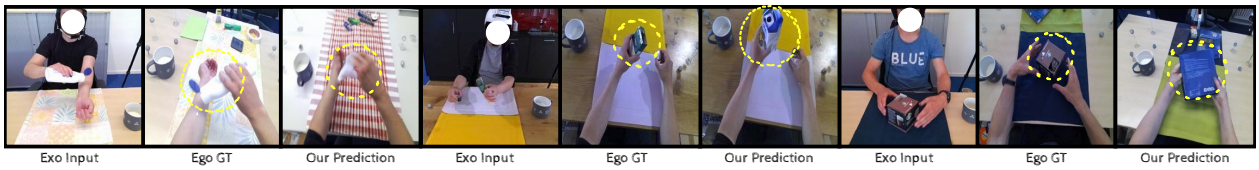

Figure 6 Failure cases: Exo2Ego generates hands with reasonably accurate positions/poses but incomplete/wrong objects in the ego view.

to limitations in handling sparse inputs, occluded scenes, and long-range camera extrapolations. Notably, our Exo2Ego outperforms pixelNeRF, highlighting the potential for future work to explore explicit geometric reasoning within our framework. We focus on hand-object interactions due to their significant value for applications in augmented reality and robotics, yet more general ego-exo settings will be interesting.

6 Conclusion

Overall, our work contributes to the growing body of research in cross-view translation and lays the groundwork for the important case of exo-to-ego. Our generative framework Exo2Ego integrates high-level structure transformation and pixel-level hallucination to yield very encouraging experimental results. We also provide a curated benchmark task for supporting continued work on this problem. The core technology holds immense potential for applications in robot learning and AR skill coaching, where an ego actor needs to replicate the actions of a demonstrator observed from the exo perspective.

7 Acknowledgment

This research has been supported by NSF Grants AF 1901292, CNS 2148141, Tripods CCF 1934932, IFML CCF 2019844 and research gifts by Western Digital, Amazon, WNCG IAP, UT Austin Machine Learning Lab (MLL), Cisco and the Stanly P. Finch Centennial Professorship in Engineering. UT Austin is supported in part by the IFML NSF AI Institute. K.G. is paid as a research scientist at Meta.

References

Shervin Ardeshir and Ali Borji. Ego2top: Matching viewers in egocentric and top-view videos. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 253–268. Springer, 2016.

Shervin Ardeshir and Ali Borji. Egocentric meets top-view. *IEEE transactions on pattern analysis and machine intelligence*, 41(6):1353–1366, 2018a.

Shervin Ardeshir and Ali Borji. An exocentric look at egocentric actions and vice versa. *Computer Vision and Image Understanding*, 171:61–68, 2018b.

Shikhar Bahl, Abhinav Gupta, and Deepak Pathak. Human-to-robot imitation in the wild. *arXiv preprint arXiv:2207.09450*, 2022.

Aayush Bansal, Shugao Ma, Deva Ramanan, and Yaser Sheikh. Recycle-gan: Unsupervised video retargeting. In *Proceedings of the European conference on computer vision (ECCV)*, pages 119–135, 2018.

Homanga Bharadhwaj, Abhinav Gupta, Shubham Tulsiani, and Vikash Kumar. Zero-shot robot manipulation from passive human videos. *arXiv preprint arXiv:2302.02011*, 2023.

Berk Calli, Arjun Singh, Aaron Walsman, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M Dollar. The ycb object and model set: Towards common benchmarks for manipulation research. In *2015 international conference on advanced robotics (ICAR)*, pages 510–517. IEEE, 2015.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020.

Eric R Chan, Koki Nagano, Matthew A Chan, Alexander W Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. Generative novel view synthesis with 3d-aware diffusion models. *arXiv preprint arXiv:2304.02602*, 2023.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

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.

Chenyou Fan, Jangwon Lee, Mingze Xu, Krishna Kumar Singh, Yong Jae Lee, David J Crandall, and Michael S Ryoo. Identifying first-person camera wearers in third-person videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5125–5133, 2017.

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. *arXiv preprint arXiv:2311.18259*, 2023.

Hsuan-I Ho, Wei-Chen Chiu, and Yu-Chiang Frank Wang. Summarizing first-person videos from third persons' points of view. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 70–85, 2018.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *The Journal of Machine Learning Research*, 23(1):2249–2281, 2022.

Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010.

Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

Wonbong Jang and Lourdes Agapito. Codenerf: Disentangled neural radiance fields for object categories. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12949–12958, 2021.

Rishabh Jangir, Nicklas Hansen, Sambaran Ghosal, Mohit Jain, and Xiaolong Wang. Look closer: Bridging egocentric and third-person views with transformers for robotic manipulation. *IEEE Robotics and Automation Letters*, 7(2): 3046–3053, 2022.

Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human pose fitting towards in-the-wild 3d human pose estimation. *3DV*, 2021.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10138–10148, October 2021.

Yanghao Li, Tushar Nagarajan, Bo Xiong, and Kristen Grauman. Ego-exo: Transferring visual representations from third-person to first-person videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6943–6953, 2021.

Andrew Liu, Richard Tucker, Varun Jampani, Ameesh Makadia, Noah Snavely, and Angjoo Kanazawa. Infinite nature: Perpetual view generation of natural scenes from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14458–14467, 2021a.

Gaowen Liu, Hao Tang, Hugo Latapie, and Yan Yan. Exocentric to egocentric image generation via parallel generative adversarial network. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1843–1847. IEEE, 2020.

Gaowen Liu, Hao Tang, Hugo M Latapie, Jason J Corso, and Yan Yan. Cross-view exocentric to egocentric video synthesis. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 974–982, 2021b.

Zhaoyang Lv, Edward Miller, Jeff Meissner, Luis Pesqueira, Chris Sweeney, Jing Dong, Lingni Ma, Pratik Patel, Pierre Moulon, Kiran Somasundaram, Omkar Parkhi, Yuyang Zou, Nikhil Raina, Steve Saarinen, Yusuf M Mansour, Po-Kang Huang, Zijian Wang, Anton Troynikov, Raul Mur Artal, Daniel DeTone, Daniel Barnes, Elizabeth Argall, Andrey Lobanovskiy, David Jaeyun Kim, Philippe Bouttefroy, Julian Straub, Jakob Julian Engel, Prince Gupta, Mingfei Yan, Renzo De Nardi, and Richard Newcombe. Aria pilot dataset. https://about.facebook.com/realitylabs/projectaria/datasets, 2022.

Arjun Majumdar, Karmesh Yadav, Sergio Arnaud, Yecheng Jason Ma, Claire Chen, Sneha Silwal, Aryan Jain, Vincent-Pierre Berges, Pieter Abbeel, Jitendra Malik, et al. Where are we in the search for an artificial visual cortex for embodied intelligence? *arXiv preprint arXiv:2303.18240*, 2023.

Priyanka Mandikal and Kristen Grauman. Dexvip: Learning dexterous grasping with human hand pose priors from video. In *Conference on Robot Learning*, 2021.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. *arXiv preprint arXiv:2203.12601*, 2022.

Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5480–5490, 2022.

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. *Advances in neural information processing systems*, 32, 2019.

William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

Krishna Regmi and Ali Borji. Cross-view image synthesis using conditional gans. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Krishna Regmi and Ali Borji. Cross-view image synthesis using geometry-guided conditional gans. *Computer Vision and Image Understanding*, 2019. ISSN 1077-3142. doi: https://doi.org/10.1016/j.cviu.2019.07.008. http://www.sciencedirect.com/science/article/pii/S1077314219301043.

Bin Ren, Hao Tang, and Nicu Sebe. Cascaded cross mlp-mixer gans for cross-view image translation. *arXiv preprint arXiv:2110.10183*, 2021.

Xuanchi Ren and Xiaolong Wang. Look outside the room: Synthesizing a consistent long-term 3d scene video from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3563–3573, 2022.

Robin Rombach, Patrick Esser, and Björn Ommer. Geometry-free view synthesis: Transformers and no 3d priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14356–14366, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *IEEE International Conference on Computer Vision Workshops*, 2021.

Maximilian Seitzer. pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid, August 2020. Version 0.3.0.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21096–21106, 2022.

Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, and Sergey Levine. Time-contrastive networks: Self-supervised learning from video. *Proceedings of International Conference in Robotics and Automation (ICRA)*, 2018. http://arxiv.org/abs/1704.06888.

Dandan Shan, Jiaqi Geng, Michelle Shu, and David F. Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Actor and observer: Joint modeling of first and third-person videos. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7396–7404, 2018a.

Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Charades-ego: A large-scale dataset of paired third and first person videos. *arXiv preprint arXiv:1804.09626*, 2018b.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *Advances in Neural Information Processing Systems*, 32, 2019.

Kiran Somasundaram, Jing Dong, Huixuan Tang, Julian Straub, Mingfei Yan, Michael Goesele, Jakob Julian Engel, Renzo De Nardi, and Richard Newcombe. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. https://openreview.net/forum?id=PxTIG12RRHS.

Bilge Soran, Ali Farhadi, and Linda Shapiro. Action recognition in the presence of one egocentric and multiple static cameras. In *Computer Vision–ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part V 12*, pages 178–193. Springer, 2015.

Hao Tang, Dan Xu, Nicu Sebe, Yanzhi Wang, Jason J Corso, and Yan Yan. Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2417–2426, 2019.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.

Hung-Yu Tseng, Qinbo Li, Changil Kim, Suhib Alsisan, Jia-Bin Huang, and Johannes Kopf. Consistent view synthesis with pose-guided diffusion models. *arXiv preprint arXiv:2303.17598*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Jian Wang, Lingjie Liu, Weipeng Xu, Kripasindhu Sarkar, Diogo Luvizon, and Christian Theobalt. Estimating egocentric 3d human pose in the wild with external weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13157–13166, 2022.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. *arXiv preprint arXiv:1808.06601*, 2018a.

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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018b.

Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. NeRF−−: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021.

Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*, 2022.

Yangming Wen, Krishna Kumar Singh, Markham Anderson, Wei-Pang Jan, and Yong Jae Lee. Seeing the unseen: Predicting the first-person camera wearer's location and pose in third-person scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 3446–3455, October 2021.

Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7467–7477, 2020.

Mingze Xu, Chenyou Fan, Yuchen Wang, Michael S Ryoo, and David J Crandall. Joint person segmentation and identification in synchronized first-and third-person videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 637–652, 2018.

Zihui Xue and Kristen Grauman. Learning fine-grained view-invariant representations from unpaired ego-exo videos via temporal alignment. In *NeurIPS*, 2023.

Yufei Ye, Xueting Li, Abhinav Gupta, Shalini De Mello, Stan Birchfield, Jiaming Song, Shubham Tulsiani, and Sifei Liu. Affordance diffusion: Synthesizing hand-object interactions. In *CVPR*, 2023.

Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4578–4587, June 2021.

Huangyue Yu, Minjie Cai, Yunfei Liu, and Feng Lu. What i see is what you see: Joint attention learning for first and third person video co-analysis. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1358–1366, 2019.

Huangyue Yu, Minjie Cai, Yunfei Liu, and Feng Lu. First-and third-person video co-analysis by learning spatial-temporal joint attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

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.

Appendix

A Experimental Setup

Dataset Details Figure 7, 8, and 9 show the preprocessed exo-ego frame pairs selected from H2O, Aria Pilot and Assembly101 respectively. For all the experiments, we perform frame cropping manually per scene to eliminate unnecessary background details and ensure that the actor is the focal point of each video. In the case of H2O, we focus on the videos of 'subject 1' and 'subject 2', in which the actors engage with 8 distinct everyday objects across 6 unique scenarios. We blurred human faces in the H2O dataset. Regarding Aria Pilot, we focus on the desktop activities set. We exclude the frames in which the actor is not wearing the glasses. These desktop activities involve tasks such as tidying up the desk, manipulating multiple objects, and manipulating single objects. The objects were primarily selected from the YCB Benchmark object set Calli et al. (2015), consisting of 10 commonly used items. Human faces are blurred in the Aria Pilot dataset. For Assembly101, we select 6 sequences that depict 5 different actors assembling a toy roller. The rationale behind our downselection lies in the visual homogeneity identified in Assembly101's video content. In essence, we aim to guarantee that the selected sequences are representative of the overall dataset. Note that Assembly101 has 4 ego views. We select 'e3' as the target ego view since it shows the clearest hand-object interaction details.

Layout Annotations As shown in Figure 10, in our Exo2Ego framework, we utilize the 2D hand pose as the layout for H2O and Assembly101. To obtain the 2D hand poses, we project them from the 3D hand poses available in the original dataset. In the case of Aria Pilot, it does not have 3D hand pose annotations, so we employ the hand mask as the layout. We first utilize an off-the-shelf hand detector Shan et al. (2020) to obtain the 2D hand bounding boxes. Subsequently, we apply Segment Anything Kirillov et al. (2023) to generate the hand masks with the bounding boxes.

Implementation Details For all baselines except PixelNeRF, we utilize the Adam optimizer Kingma and Ba (2014) with a learning rate of 0.0002 and betas set to (0.5, 0.999). For PixelNeRF and Exo2Ego, we use the Adam optimizer with a learning rate of 0.0001. Regarding Exo2Ego's high-level structure transformation module, we adopt DeiT Touvron et al. (2021) as the transformer backbone. For all methods, we monitor the training loss curve, ensuring that the losses exhibit a satisfactory convergence by the specified final epochs. Additionally, an evaluation of the synthesis results on the training set is performed to confirm the stability of synthesis quality and its tendency to plateau — further improvements are unlikely. All experiments were conducted using PyTorch 1.10 Paszke et al. (2019).

B Mask as Layout

In the hand mask layout case, we employ a hand mask with dimensions of $\{H \times W \times 1\}$ as the target ego layout. The overall architectures of the transformer encoder and decoder remain consistent with the hand pose case. The only distinction lies in the decoder, which produces a feature map with dimensions of $\{H \times W \times 2\}$ rather than hand joint locations. Treating mask generation as a pixel-wise classification problem (hand region vs. non-hand region), we define the number of classes to be 2. For the decoder output, we apply softmax to the class dimension to derive the final mask prediction. The model is trained end-to-end using per-pixel cross-entropy loss.

C Extra Experimental Results

Additional Qualitative Results We offer additional qualitative results of generalizing to new actions on all datasets in Figure 12, 13, and 14. Our Exo2Ego framework has demonstrated superior performance compared to all baseline methods when it comes to synthesizing hand-object interactions that are realistic and visually coherent.

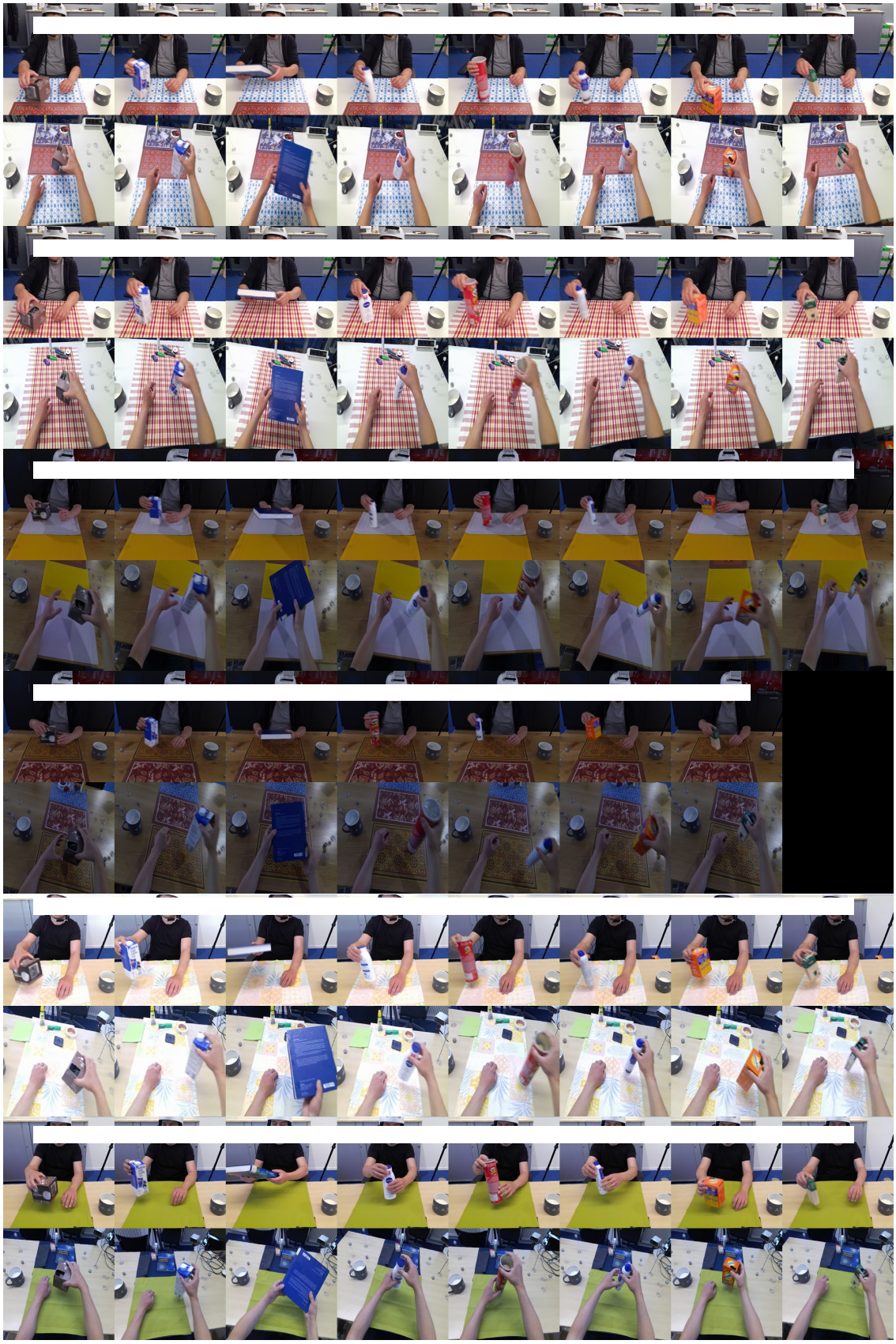

Figure 7 Selected samples from the H2O dataset.

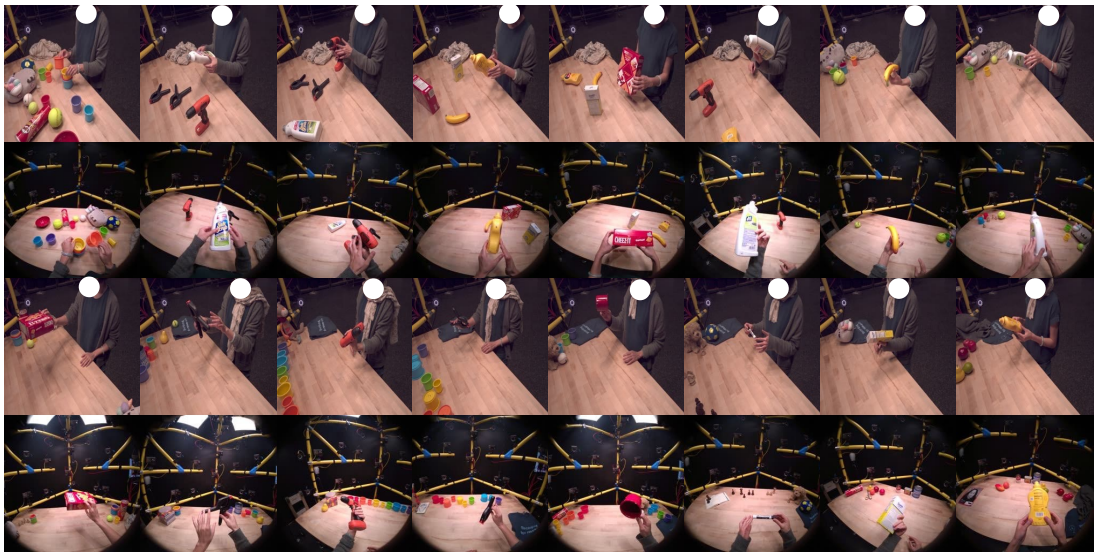

Figure 8 Selected samples from the Aria Pilot dataset.

Figure 9 Visualization of the toy roller and selected samples from the Assembly101 dataset.

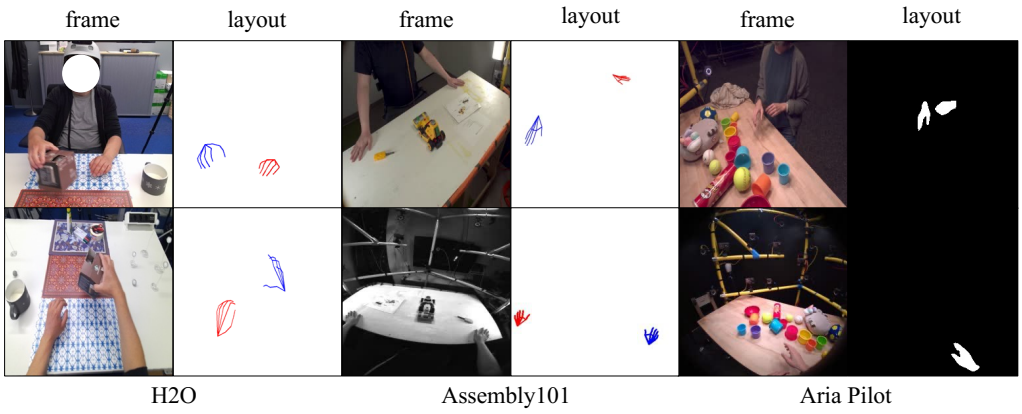

Figure 10 Layout visualizations for all datasets.

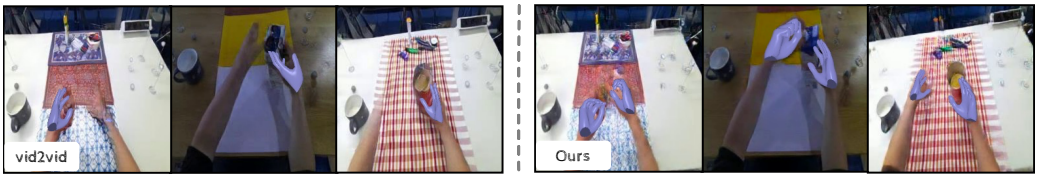

Figure 11 Visualizing mesh rendered by a pretrained 3D hand estimation model.

Table 2 Using different exo cameras as source views on H2O.

Source Camera	SSIM↑	PSNR↑	FID↓	P_{Squeeze}↓	P_{Alex}↓	P_{Vgg}↓
cam 0	0.427	**30.394**	135.74	0.165	0.222	0.338
cam 1	0.416	30.226	142.59	0.172	0.231	0.347
cam 2	**0.428**	30.370	**132.03**	**0.163**	**0.220**	**0.337**
cam 3	0.410	30.281	143.21	0.184	0.243	0.356

Ablation Study on the Choice of Source Exo View We simply choose the exo view with the largest FOV of the ego view – showcasing the intricate details of hand and object interaction most effectively. We have conducted ablation studies on the selection of different exo views (cam 0-4) within the H2O dataset. As shown in Table 2, we observe that different sources for the exo view do impact the model's performance, and the exo view – cam 2, chosen by us – with the highest FOV alongside the ego view, delivers the best translation performance.

Extracting 3D Hand Pose Following Ye et al. (2023), we employ FrankMocap Rong et al. (2021); Joo et al. (2021), an off-the-shelf hand pose estimator, to directly extract 3D poses from the generated ego frames. The extracted 3D poses are visualized in Figure 11. It is observed that our Exo2Ego framework surpasses vid2vid in synthesizing highly realistic and visually coherent hands, enabling the rendering of exceptional 3D hand meshes that exhibit accurate positioning and overall realism. Note that the quality of 3D hand mesh is essential for creating vivid and lifelike experiences in virtual and augmented reality (VR/AR).

D Limitations and Future Work

Generalization Ability There are different types of generalization. As an early-stage work for exo-to-ego cross-view translation, our paper has made non-trivial progress in producing reasonable hand manipulation details for novel actions. However, our model does not generalize well to in-the-wild objects, subjects, and backgrounds. This is expected, given its training on a modest scale of data. Considering the intricate nature of the generalization to new environments, we regard it as a challenging future research direction. Nonetheless, our focus on generalizing to new actions within the same environment is meaningful for several applications. Some examples: Real-time monitoring during physical rehabilitation (exo-to-ego view translation can help therapists guide patients during rehabilitation sessions), assistive cooking aid for visually impaired individuals (ego views are provided to the visually impaired individuals to cook independently), or real-time fitness training monitoring (coaches use synthesized ego views to help the trainee correct the fitness movements) — in each case, the environment is stable between training and testing, and yet the model must generalize to new actions for cross-view translation.

Object 3D prior Our empirical qualitative results reveal the limitation of the existing baselines and our Exo2Ego framework when it comes to generating 3D consistent views for novel objects during test time. This can be attributed to the lack of geometric priors for common objects in our daily lives. In order to address this issue, our future works include incorporating robust object geometric priors into our framework, enhancing the realism and precision of the generated objects in ego views, and extending our approach to in-the-wild scenarios.

Action Semantics There is currently no dataset available that encompasses pairs of exo-ego videos showcasing a wide spectrum of action semantics. Given the absence of such a dataset, our emphasis in this work lies

on the domain of hand-object tabletop interactions. However, our task setting still holds significant value, particularly considering its relevance to a range of applications in augmented reality and robotics.

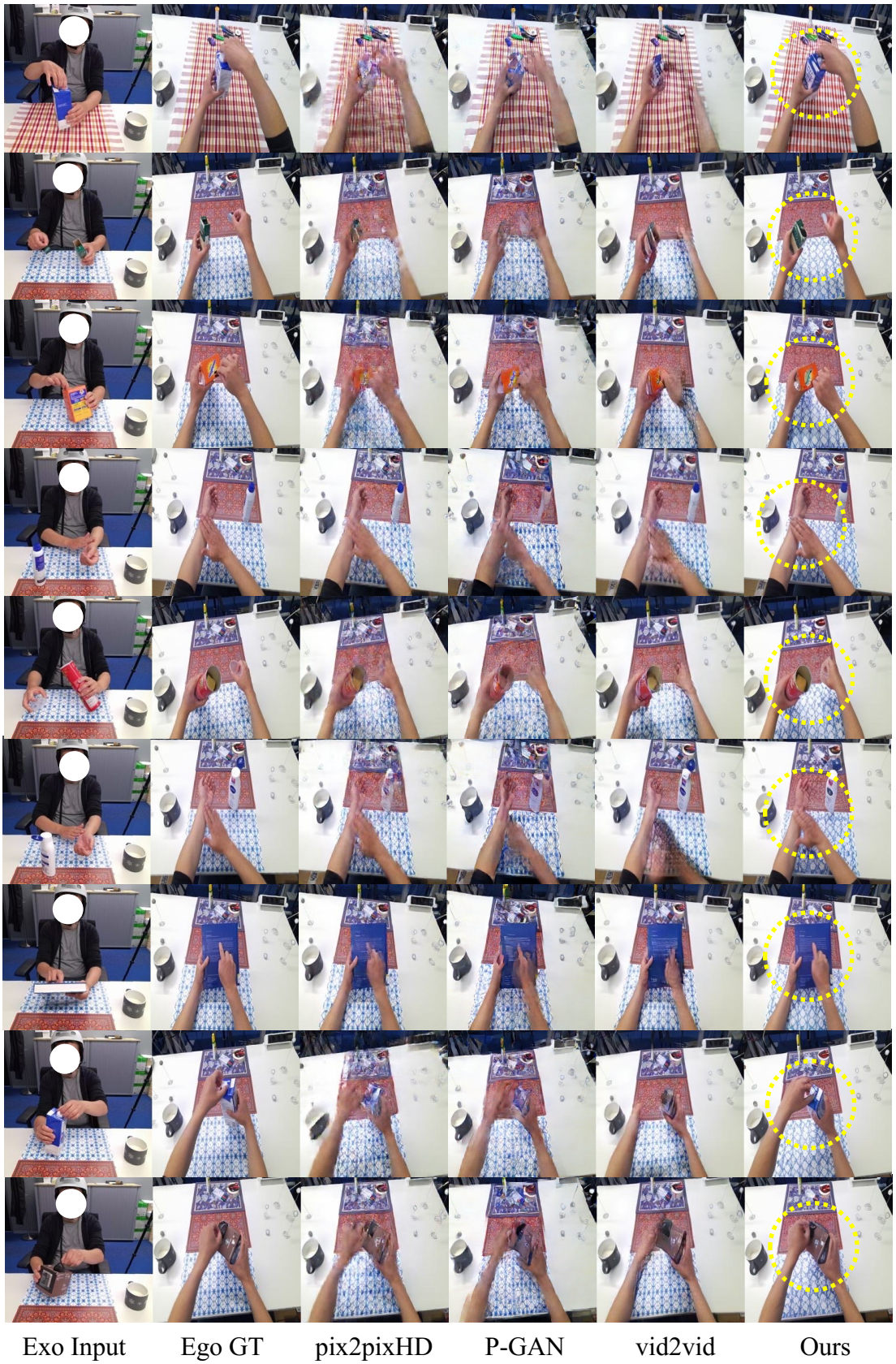

Figure 12 Extra qualitative results of generalizing to new actions on H2O.

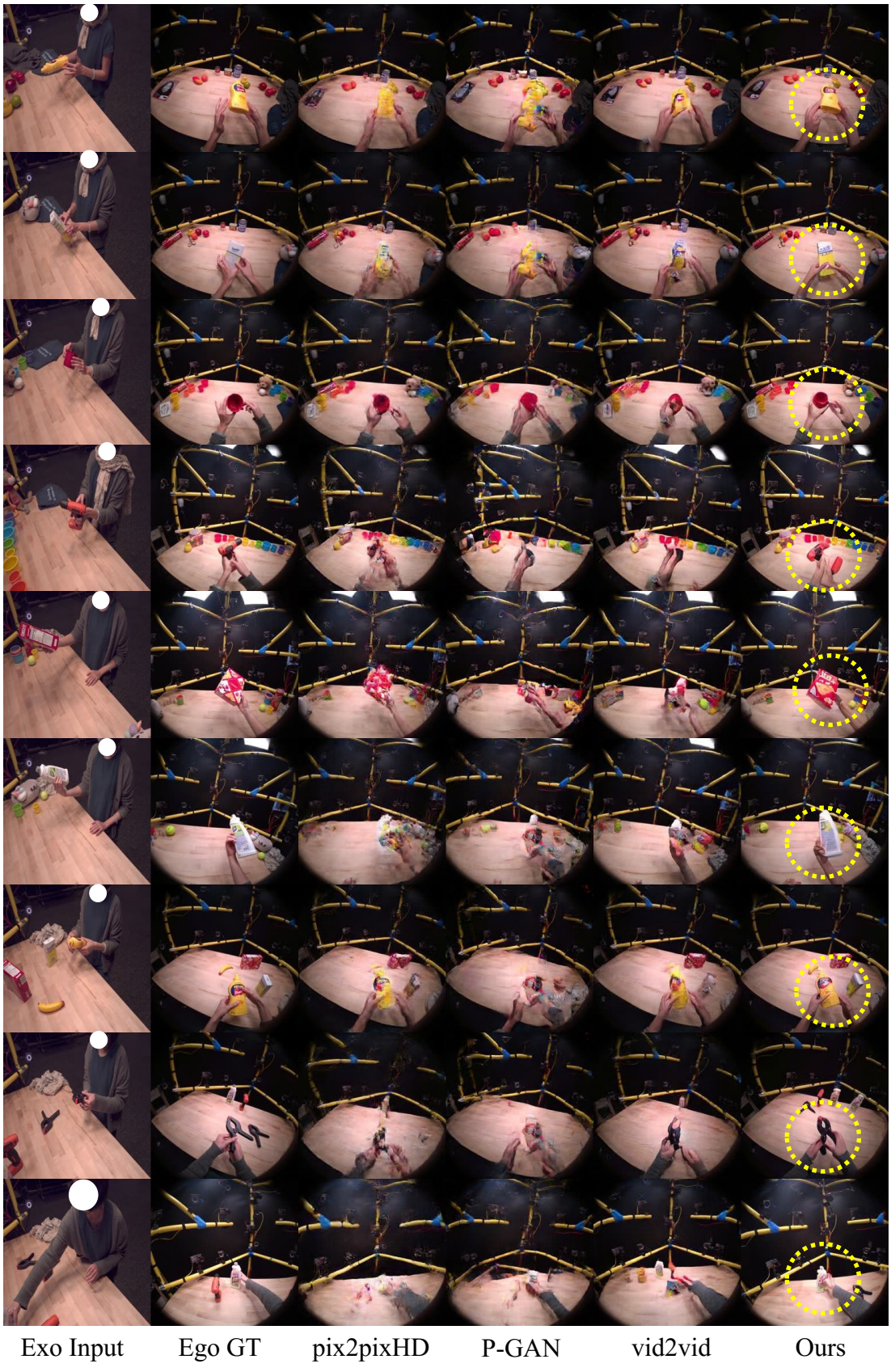

Figure 13 Extra qualitative results of generalizing to new actions on Aria Pilot.

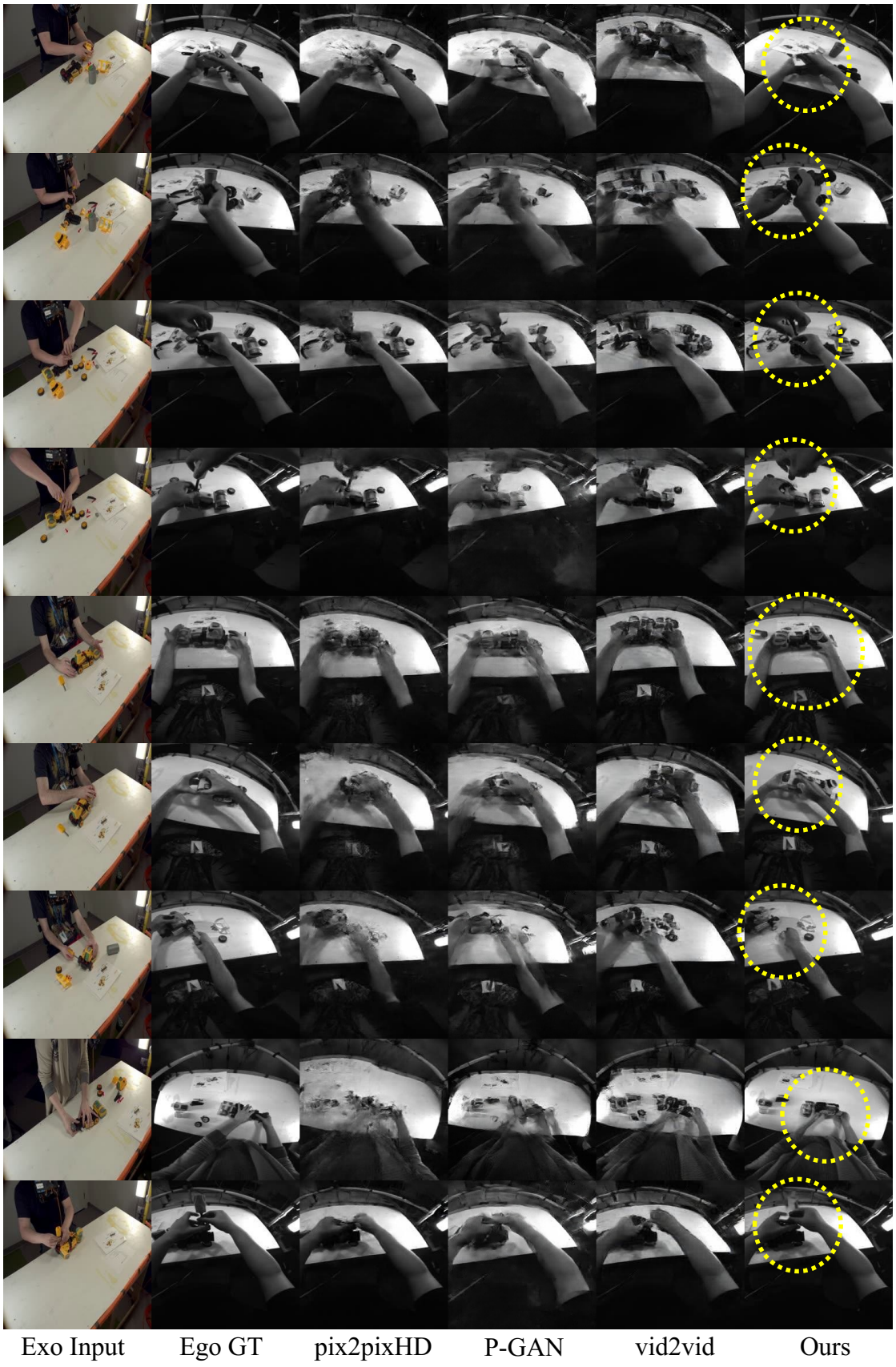

Figure 14 Extra qualitative results of generalizing to new actions on Assembly101.