# Exocentric-to-Egocentric Video Generation

**Jia-Wei Liu**[1*], **Weijia Mao**[1*], **Zhongcong Xu**[1], **Jussi Keppo**[2], **Mike Zheng Shou**[1✉]

[1]Show Lab, [2]National University of Singapore

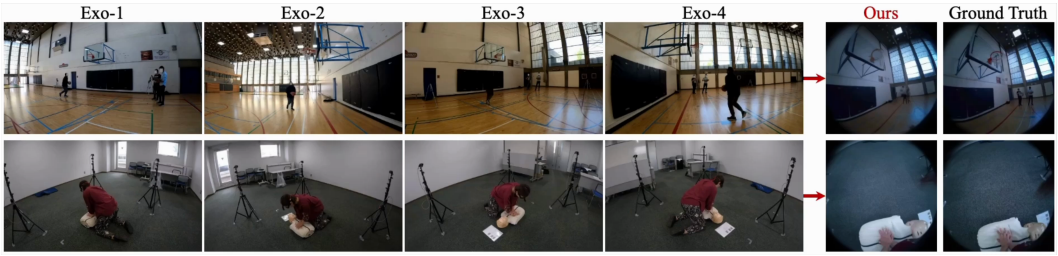


Figure 1: Given sparse 4 exocentric videos configured 360° around daily-life skilled human activities such as playing basketball (upper), CPR training (lower), our Exo2Ego-V can generate corresponding egocentric videos with the same activity and environment as the exocentric videos. We encourage readers to click and play the video clips in this figure using Adobe Acrobat.

## Abstract

We introduce Exo2Ego-V, a novel exocentric-to-egocentric diffusion-based video generation method for daily-life skilled human activities where sparse 4-view exocentric viewpoints are configured 360° around the scene. This task is particularly challenging due to the significant variations between exocentric and egocentric viewpoints and high complexity of dynamic motions and real-world daily-life environments. To address these challenges, we first propose a new diffusion-based multi-view exocentric encoder to extract the dense multi-scale features from multi-view exocentric videos as the appearance conditions for egocentric video generation. Then, we design an exocentric-to-egocentric view translation prior to provide spatially aligned egocentric features as a concatenation guidance for the input of egocentric video diffusion model. Finally, we introduce the temporal attention layers into our egocentric video diffusion pipeline to improve the temporal consistency cross egocentric frames. Extensive experiments demonstrate that Exo2Ego-V significantly outperforms SOTA approaches on 5 categories from the Ego-Exo4D dataset with an average of 35% in terms of LPIPS. Our code and model will be made available on `https://github.com/showlab/Exo2Ego-V`.

## 1 Introduction

When people observe and learn skills such as cooking or playing basketball from an exocentric (third-person) perspective, they can easily envision themselves executing these skills from an egocentric (first-person) perspective [2]. This exocentric-egocentric (Exo-Ego) translation remains the foundation of visual learning [15] for both human beings and AI robots [8, 4, 3, 37, 35], and unleashes new opportunities for AI assistant [54, 7, 13] and augmented reality [1]. However, it remains particularly challenging for computer vision algorithms to achieve such exocentric to egocentric video generation for daily-life skilled human activities, primarily due to 1) significant variations between exocentric

[*] Equal Contribution [✉] Corresponding Author

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

and egocentric viewpoints, and 2) high complexity of dynamic motions and daily-life environments, as illustrated in Fig. 1.

Existing view translation approaches mainly focus on the task of novel view synthesis and have made remarkable progress particularly since the introduction of Neural Radiance Fields (NeRF) [36]. While initially limited to reconstructing static 3D scenes, subsequent studies have extended NeRF to address the challenges of dynamic view synthesis [42, 38, 39, 50, 27, 12, 30]. However, these approaches are limited to per-scene regressive optimization and require dozens or hundreds of views as input. As a result, they fall short in the exocentric to egocentric video generation task due to the sparse yet highly variable viewpoints and significant occlusions.

The remarkable success of powerful image diffusion models [45] provides new opportunities for introducing such generative models in the task of exo-ego generation. Recent attempt [33] leverages action intention consisting of human movement and action description for Ego2Exo video generation. However, it strictly requires the first exocentric frame a priori which largely simplifies the ego-to-exo generation task to optical flow prediction and exocentric frame warping, highly limiting its applications in general exo-ego generation tasks. On the other hand, Exo2Ego [34] attempts to tackle with the exo2ego view translation by first transferring exo hand pose to ego, and then learning the conditional distribution of the target ego image given a single exo image and the predicted ego hand pose. Despite promising, it is limited to image-level translation and requires a carefully designed capturing setup with the exocentric camera configured close to the hand-object region and thus is restricted to desktop activities with simple environments.

In contrast, we contend that it is necessary for exo-ego translation algorithms to be resilient to the complexity and diversity of daily-life scenarios such as cooking in kitchens, playing basketball in courts, etc. The introduction of Ego-Exo4D [15] opens new opportunities and challenges for exo-ego translation by providing a large-scale simultaneously-captured egocentric and exocentric videos of daily-life skilled human activities. In order to capture the complete and complex human-environment activities, they configure $4$ exocentric cameras in $360^\circ$ around the dynamic scene, resulting in new challenges of significant variations between exocentric and egocentric viewpoints, as well as complex dynamic motions and daily-life environments, as shown in Fig. 1.

To tackle with these challenges, we propose a novel exocentric-to-egocentric video diffusion pipeline dubbed as Exo2Ego-V. We address the significant challenges of large viewpoint variations and complex environments from two aspects: exo appearance conditions and ego translation prior. Firstly, we propose a diffusion-based multi-view exocentric encoder to extract the multi-scale exocentric features as the appearance conditions for egocentric video generation. We achieve this by concatenating ego hidden states with exo features for self-attention computation, so that the ego hidden states can attend to both the egocentric features as well as the multi-view exocentric features through the self-attention mechanism. In addition, we inject the relative position information into our exocentric encoder by adding exocentric latents with relative Exo2Ego relative camera pose embedding. Our exocentric encoder can extract dense human activity and environment information to guide the appearance of egocentric video generation pipeline. Secondly, we introduce an Exo2Ego view translation prior based on PixelNeRF [60] to provide coarse yet spatially aligned egocentric features as a concatenation guidance for the input of egocentric video diffusion model. Finally, to improve the temporal dynamic motion consistency of egocentric video contents, we insert temporal layers into our egocentric video diffusion pipeline to encode the temporal information across ego frames.

We extensively evaluate our Exo2Ego-V on $5$ categories of skilled human activities from the challenging Ego-Exo4D [15] dataset and H2O dataset [26]. As shown in Fig. 1, our Exo2Ego-V can generate the corresponding egocentric videos given $4$ multi-view exocentric videos, and significantly outperforms SOTA approaches with an average of $35\%$ in terms of LPIPS.

To summarize, the major contributions of our paper are:

- We present a novel framework of Exo2Ego-V, the first work to achieve exocentric-to-egocentric video generation for daily-life real-world skilled human activities.
- We propose a new diffusion-based multi-view exocentric encoder and an Exo2Ego view translation prior that can extract dense exocentric features and spatially aligned egocentric features as conditions for our egocentric video diffusion pipeline.
- Extensive experiments show that Exo2Ego-V significantly outperforms SOTA approaches on the challenging Ego-Exo4D [15] dataset with an average of $35\%$ in terms of LPIPS.

## 2 Related work

### 2.1 Egocentric-exocentric vision

Tremendous progress has been made for exocentric vision with various visual perception and generation tasks due to the large amount of dataset captured at thrid-person views [11, 23, 48]. Recently, egocentric vision has also been scaling up particularly since the introduction of EPIC-Kitchens [9, 10] and Ego4D [14]. Previous attempts for joint egocentric and exocentric vision explore learning the ego-exo view-invariant features on paired small-scale dataset [2, 61] or through unpaired learning [59]. Another line of research focuses on egocentric human localization from exocentric videos [57, 52], as well as egocentric human pose estimation from exocentric videos [51]. More recently, the introduction of Ego-Exo4D dataset [15] opens up new opportunities for joint egocentric and exocentric vision with large-scale synchronized multi-view ego-exo videos with multi-modality annotations.

### 2.2 Egocentric-exocentric cross-view translation

**Ego-exo view translation.** There is limited prior work on ego-exo cross-view translation. Early attempt [28] explores the exo-to-ego image generation with a novel parallel generative adversarial network to learn shared features of exo and ego images. STA-GAN [29] further extends P-GAN [28] to Exo2Ego video synthesis with a spatial temporal attention fusion module. However, they are limited to simple activities such as walking where most contents in egocentric and exocentric views are static environments [29, 28]. More recently, IDE [33] leverages action intention consisting of human movement and action description for ego2exo video generation. However, it requires the first exocentric frame a priori which largely simplifies the ego-to-exo generation task to optical flow prediction and exocentric frame warping. On the other hand, Exo2Ego [34] attempts to tackle with the exo2ego view translation by first transferring exo hand pose to ego, and then learning the conditional distribution of the target ego image given a single exo image and predicted ego hand pose. Despite promising, it is limited to image-level translation and requires a carefully designed capturing setup with the exocentric camera configured close to the hand-object region and thus is restricted to desktop activities with simple environments.

**Novel view synthesis (NVS).** Ego-Exo view translation is also related to NVS, which has made remarkable progress particularly since the introduction of NeRF [36]. While initially limited to reconstructing static 3D scenes, NeRF has been extended to modelling dynamic scenes [42, 38, 39, 50, 27, 12], dynamic humans [41, 53, 22, 32]. However, these approaches are limited to per-scene regressive optimization and require dozens or hundreds of views as input. On the other hand, generalizable scene reconstruction methods [60, 56] are still limited to static scenes. As a result, they fall short in the exocentric to egocentric video generation task due to the sparse yet highly variable viewpoints and significant occlusions.

### 2.3 Video generation

Recent works have extended the power of image diffusion models to video editing [55, 43, 31] and generation [6, 62, 63, 18, 20]. Tune-A-Video [55] inflates the image diffusion with cross-frame attention and fine-tunes the source video, aiming to implicitly learn the source motion and transfer it to the target video. Video Diffusion Models(VDM) [20] designs a factorized space-time UNet to generate videos. Stable Video Diffusion [5] introduces a systematic data curation workflow, enabling the training of a state-of-the-art text-to-video and image-to-video models. AnimateDiff [16] proposes a plug-and-play motion module on temporal layers for personalized text-to-image animation. Other approaches [21, 58] introduce such video generation architectures to human image animation and achieve faithful performances. Our Exo2Ego-V is a new video diffusion pipeline for the challenging Exo2Ego generation on daily-life skilled human activities.

## 3 Method

### 3.1 Preliminaries

**Latent diffusion models (LDMs).** LDMs encode input images to a latent representation using a pretrained variational auto-encoder (VAE) and operate the diffusion and denoising process following

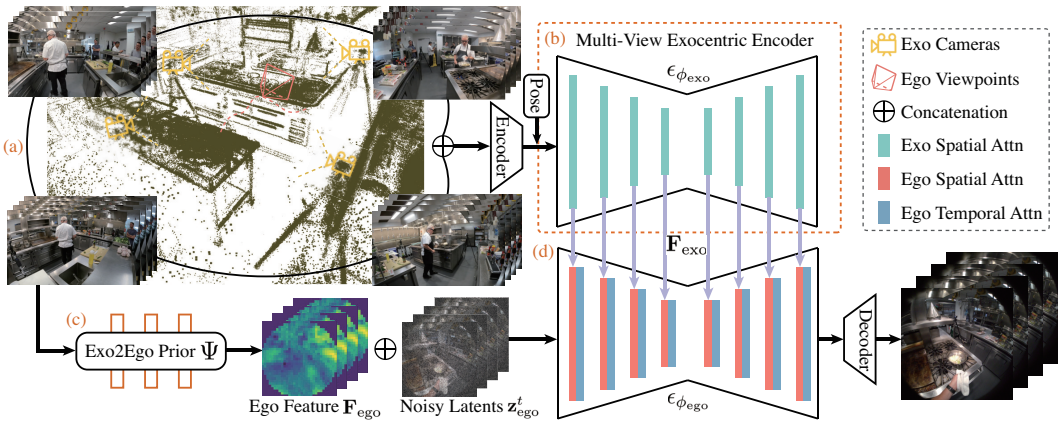


Figure 2: **Overview of Exo2Ego-V.** Given 4 exocentric videos configured 360° around daily-life skilled human activities such as cooking (a), our multi-view exocentric encoder (b) extracts the multi-scale exocentric features as the appearance conditions for egocentric video generation, and our Exo2Ego view translation prior (c) predicts the egocentric features as the concatenation guidance for the egocentric noisy latents input. With these information, our egocentric video diffusion pipeline (d) generates the egocentric videos with the same activity and environment as the exocentric videos.

denoising diffusion probabilistic models (DDPMs) [19] (see Sec. A.6 for more details) in the reduced-dimension latent space, and finally decode the denoised latents to the image space. Our model is an extention of the pretrained text-to-image latent diffusion model [45] that utilizes the UNet architecture [49] for denoising noise prediction with multiple down, middle, and up blocks. Each block consists of a ResNet2D layer, a self-attention layer, and a cross-attention layer.

### 3.2 Overall framework

**Task definition.** Given 4 exocentric videos $\mathbf{V}_{\text{exo}} = \left[\mathbf{V}^1_{\text{exo}}, \mathbf{V}^2_{\text{exo}}, \mathbf{V}^3_{\text{exo}}, \mathbf{V}^4_{\text{exo}}\right]$ configured 360° around daily-life skilled human activities, our objective is to generate their corresponding egocentric video $\mathbf{V}_{\text{ego}}$, as shown in Fig. 2(a). Since these daily-life activities happen naturally in real-world scenarios such as kitchens, basketball courts, bike stores, etc, our task features the diversity of the real world with complex activities and environments. Therefore, the sparse 4 exocentric cameras have to be configured evenly in 360° around the dynamic scene in order to capture both the complex human activities and real-world environments [15], resulting in **significant variations between exocentric and egocentric viewpoints**. Furthermore, such real-world skilled human activities such as cooking, repairing bike, and playing basketball are also highly challenging in terms of the **complexity of dynamic motions and real-world environments**.

**Overall framework.** In order to tackle with the above challenges, we propose a novel exocentric-to-egocentric video diffusion pipeline dubbed as Exo2Ego-V, as shown in Fig. 2. To address the significant challenges of large viewpoint variations and complex environments, we propose a diffusion-based multi-view exocentric encoder (Fig. 2(b)) to extract the multi-scale multi-view exocentric features as the appearance conditions for egocentric video generation. In addition, we design an exocentric-to-egocentric view translation prior (Fig. 2(c)) based on PixelNeRF [60] to provide coarse yet spatially aligned egocentric features as a concatenation guidance for egocentric video generation. Finally, we introduce the temporal attention layers into our egocentric video diffusion pipeline to improve the temporal consistency cross egocentric frames (Fig. 2(d)).

### 3.3 Multi-view exocentric encoder

**Motivation.** Our task is featured with the significant variations of Exo2Ego viewpoints and complex environments. The core to tackle with this challenge is to fully explore the multi-view exocentric information for the purpose of guiding the egocentric video generation. A naive solution is to utilize the CLIP [44] image encoder to extract latent features from low-resolution images. However, such semantic-level CLIP image features fall short in extracting the dense and fine-grained detail information from exocentric videos. Another naive solution is to train a 4D dynamic scene reconstruc-

tion model for each multi-view exocentric sequences, but it requires high computation and storage resources and current methods cannot handle sparse $4$ views complex dynamic scene reconstruction.

**Framework.** Inspired by recent reference image animation methods [58, 21] that extract dense image features with a reference UNet to preserve reference human identity, we propose our multi-view exocentric encoder with a different purpose of extracting the dense multi-view exocentric intricate details as appearance conditions for egocentric video generation. Specifically, our exocentric encoder creates a trainable copy of the base Ego UNet and inject the relative camera poses from exocentric viewpoints to the egocentric viewpoints as additional embeddings, as shown in Fig. 2(b). We additionally explore adding temporal layers for exocentric encoder in our ablation study.

Given $\mathbf{V}_{\text{exo}} \in \mathbb{R}^{N\times C\times F\times H\times W}$ and their relative camera poses $\mathbf{P} \in \mathbb{R}^{N\times 4\times 4}$, our exocentric encoder computes the multi-view multi-scale appearance condition features $\mathbf{F}_{\text{exo}}$ for the egocentric video generation at denoising step $t = 0$:

$$\mathbf{F}_{\text{exo}} = \epsilon_{\phi_{\text{exo}}}\left(\mathbf{z}^t_{\text{exo}};\, \mathbf{V}_{\text{exo}},\, \mathbf{P},\, t\right)\,, \tag{1}$$

where $N = 4$ is the number of exocentric views and $F = 8$ is the number of frames for each video. $\mathbf{F}_{\text{exo}}$ are the normalized attention features for the downsampling, middle, and upsampling blocks of Exo UNet. We set $t = 0$ to preserve the appearance details of the noise-free exocentric videos.

Then, the exocentric features $\mathbf{F}_{\text{exo}}$ are utilized as the appearance conditions for egocentric video generation by concatenating $\mathbf{F}_{\text{exo}}$ with the corresponding egocentric UNet hidden states $\mathbf{z}^t_{\text{ego}}$ for the self-attention layers in every block $b$ at each denoising step $t$:

$$\mathbf{Q}_b = \mathbf{W}^{\mathbf{Q}}_b \cdot \mathbf{z}^t_{\text{ego},b},\, \mathbf{K}_b = \mathbf{W}^{\mathbf{K}}_b \cdot \left[\mathbf{z}^t_{\text{ego},b},\, \mathbf{F}_{\text{exo},b}\right],\, \mathbf{V}_b = \mathbf{W}^{\mathbf{V}}_b \cdot \left[\mathbf{z}^t_{\text{ego},b},\, \mathbf{F}_{\text{exo},b}\right]\,, \tag{2}$$

where $[\cdot]$ denotes concatenation operation, and the Ego UNet self-attention is: $\text{Softmax}\left(\frac{\mathbf{Q}_b\cdot\mathbf{K}_b^{\text{T}}}{\sqrt{d}}\right)\cdot\mathbf{V}_b$. The queried egocentric noisy latents can attend to both the egocentric features as well as the multi-view exocentric features through the self-attention mechanism, and thus translate the appearance of complex human skill activities and environments from exocentric views to the egocentric view.

**Camera pose.** In order to inject the relative position information into our Exo2Ego generation pipeline, inspired by MVDream [47], we embed the relative Exo2Ego camera poses with a 2-layer MLP and add the embedding with the denoising timestep embedding for our multi-view exocentric encoder. Since we set $t = 0$ for our Exo UNet, the relative camera pose embeddings are the main embedding to inject the relative position information into the exocentric feature extraction.

### 3.4 Exocentric-to-egocentric view translation prior

**Motivation.** Although our multi-view exocentric encoder can extract dense and intricate appearance details, it entirely relies on the self-attention mechanism to explore the correspondences from the egocentric contents to the exocentric features, which are still challenging for our scenarios with large viewpoints variations. To tackle with this, we design an Exo2Ego view translation prior based on PixelNeRF [60] to generate a coarse yet spatially aligned egocentric latent feature as the concatenation guidance for our egocentric video generation pipeline.

**Framework.** Given multi-view exocentric videos $\mathbf{V}_{\text{exo}}$, exo camera poses $\mathbf{P}_{\text{exo}}$, egocentric videos $\mathbf{V}_{\text{ego}}$, and ego camera poses $\mathbf{P}_{\text{ego}}$, we learn an Exo2Ego view translation prior $\Psi$ by training a generalizable PixelNeRF [60] for all timesteps. For each synchronized timestep of the $4$ exocentric videos and $1$ egocentric video, at each iteration we extract $4$ exo frames and $1$ ego frame and randomly sample rays from these $5$ images for optimization. Inspired by ReconFusion [56], we utilize a light PixelNeRF with 6-layer MLPs for higher efficiency. Please see Sec. A.1 for more details.

As shown in Fig. 2(c), with this Exo2Ego translation prior, we can render both the egocentric features $\mathbf{F}_{\text{ego}}$ and egocentric pixels $\mathbf{I}_{\text{ego}}$ given the multi-view exocentric videos, exo camera poses, and queried ego camera pose:

$$\Psi\left(\mathbf{V}_{\text{exo}},\, \mathbf{P}_{\text{exo}},\, \mathbf{P}_{\text{ego}}\right) \longmapsto \left(\mathbf{F}_{\text{ego}},\, \mathbf{I}_{\text{ego}}\right), \tag{3}$$

Inspired by ReconFusion [56], we design our translation prior to render egocentric features $\mathbf{F}_{\text{ego}}$ at the egocentric viewpoint with the same spatial resolution as the egocentric latents, so that $\mathbf{F}_{\text{ego}}$ is spatially aligned with the noisy egocentric latents. Therefore, we concatenate $\mathbf{z}^t_{\text{ego}}$ with $\mathbf{F}_{\text{ego}}$ along channel dimension as the input to the egocentric video diffusion model to predict the noise $\epsilon_{\text{ego}}$ at

each denoising timestep $t$. In addition, we extract the CLIP image feature of the rendered ego images $\mathbf{I}_{\text{ego}}$ as the cross-attention information for our egocentric video diffusion pipeline:

$$\epsilon_{\text{ego}} = \epsilon_{\phi_{\text{ego}}}\left(\left[\mathbf{z}^t_{\text{ego}},\, \mathbf{F}_{\text{ego}}\right];\, \mathbf{F}_{\text{exo}},\, \mathbf{I}_{\text{ego}},\, t\right) . \tag{4}$$

### 3.5 Temporal dynamic motion layer

To improve the temporal dynamic motion consistency of egocentric and exocentric video contents, we follow common practice [18, 20] to insert temporal attention layers within the 2D UNet blocks, as shown in Fig. 2(d). Specifically, we insert the temporal layers on the egocentric video generation pipeline and we additionally ablate on inserting the temporal layers on the exocentric encoder. As such, the input egocentric latents $\mathbf{z}^t_{\text{ego}} \in \mathbb{R}^{N\times C\times F\times H\times W}$ are first reshaped to $\mathbb{R}^{(NF)\times H\times W\times C}$ for computing the spatial attentions with egocentric and exocentric features in spatial layers, and then reshaped to $\mathbb{R}^{(NHW)\times F\times C}$ to compute the temporal cross-frame information in temporal layers.

### 3.6 Optimization

We optimize our Exo2Ego-V in a 2-stage training strategy. In the first stage, we remove the temporal layers and optimize the Exo2Ego spatial appearance translation modules, including the multi-view exocentric encoder, Exo2Ego view translation prior, as well as the Ego UNet. In the second stage, we only optimize the Ego temporal layers for temporal consistency and freeze other modules.

**Exo2Ego spatial appearance translation.** We first pre-train our Exo2Ego view translation prior with the pixel-level reconstruction loss $\mathcal{L}_{\text{REC}}$. Then, we alternately finetune the Exo2Ego view translation prior with the reconstruction loss $\mathcal{L}_{\text{REC}}$, and multi-view exocentric encoder and the Ego UNet with the noise prediction loss $\mathcal{L}_{\text{S}}$.

$$\mathcal{L}_{\text{REC}} = \|\mathbf{R}_{\text{render}} - \mathbf{R}_{\text{gt}}\|_2^2 , \quad \mathcal{L}_{\text{S}} = \mathbb{E}_{\mathbf{z}^t_{\text{ego}},\mathbf{V}_{\text{exo}},\mathbf{P}_{\text{exo}},t,\epsilon}\left[\omega\left(t\right)\|\epsilon - \epsilon_{\text{ego}}\|_2^2\right] , \tag{5}$$

where $\mathbf{R}_{\text{render}}$ is the rendered ray pixels sampled randomly from exocentric and egocentric frames, $\mathbf{R}_{\text{gt}}$ is the corresponding ground-truth pixels. $\epsilon$ is the ego noise sampled from $\mathcal{N}(0,1)$. $w(t)$ is a weighting function that depends on the noise level $t$.

**Temporal motion finetuning.** In the second stage, we freeze the translation prior and Exo and Ego UNets, and finetune the pretrained temporal layers from AnimateDiff [16] on our egocentric and exocentric videos with $F$ frames in temporal dimension.

$$\mathcal{L}_{\text{T}} = \mathbb{E}_{\mathbf{z}^{t,F}_{\text{ego}},\mathbf{V}^F_{\text{exo}},\mathbf{P}^F_{\text{exo}},t,\epsilon^F,}\left[\omega^F\left(t\right)\left\|\epsilon^F - \epsilon^F_{\text{ego}}\right\|_2^2\right] . \tag{6}$$

## 4 Experiments

### 4.1 Dataset

We evaluate our method on 5 categories of Ego-Exo4D dataset [15] featuring both exocentric and egocentric human activities: Cooking, Covid Test, Basketball, CPR, and Bike. Each category contains synchronized captured exo and ego videos of different participants performing these activities at different locations around the world. Specifically, **Cooking** captures people preparing various dishes in kitchens. **Basketball** captures participants playing basketball in courts. **Covid Test** captures individuals conducting covid tests for themselves in various scenes. **CPR** captures scenes where participants perform cardiopulmonary resuscitation on a CPR model. **Bike** captures scenes of participants repairing bikes in bike stores. We set the the number of temporal frames to 8 and spatial resolution to $480 \times 270$ and $256 \times 256$ for exocentric and egocentric videos, respectively. For each video from the above categories, we extract frames at 7.5 fps and split them into multiple action clips according to the action annotations and turning timesteps where the participant's head pose turns more than $45°$ within 1 second. We retain the videos that contain both ego and exo intrinsic and extrinsic parameters. Finally, we processed 489 videos from Cooking category, 909 videos from Basketball category, 127 videos from Covid Test category, 66 videos from CPR category, and 359 videos from Bike category. The videos lengths vary between $3 \sim 15$ minutes for different categories. We also evaluate our method on the H2O dataset [26], which provides synchronized multi-view Exo-Ego images for desktop activities.

We train our Exo2Ego-V (see Sec. A.2 for training details) and baselines for each category and utilize the following 3 test evaluation: (1) **Unseen action**: We split each video into multiple clips based on the action annotations, so that each clip features a different action. We use $80\%$ of action clips as our train set and the remaining $20\%$ unseen action clips as test set. (2) **Unseen take**: Each take refers to a complete human activity video. Each participant conduct $2 \sim 4$ takes for an activity. We randomly select one take as our test set and use the remaining takes for training. (3) **Unseen scenes**: Each category is captured in multiple different scenes around the world. We randomly select one entire scene with multiple takes out of a category as our test set and use the remaining scenes for training.

### 4.2 Comparisons with SOTA approaches

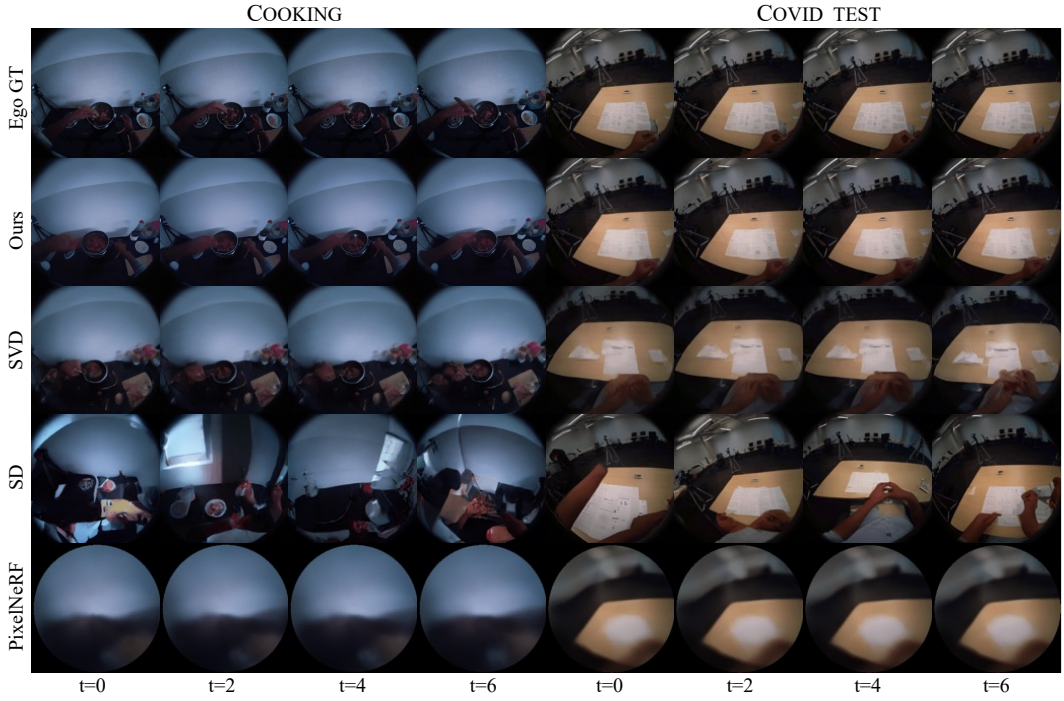


Figure 3: Qualitative comparisons of our method against SOTA approaches on **unseen actions**.

**Baselines.** We compare our Exo2Ego-V with three baselines. (1) *Stable Video Diffusion (SVD)* [5], a recent state-of-the-art image-to-video diffusion model. (2) *Stable Diffusion (SD)* [46], a powerful text-to-image diffusion model. (3) *PixelNeRF* [60], a general 3D scene reconstruction model. We adapt SVD and SD for our Exo2Ego generation task by inputting $4$ exo views as their conditions and train the models to generate ego views. Specifically, we first use a VAE model to obtain the latent contents of each exo view and concatenate ego noisy latent with these $4$ exo latents along the channel dimension as input. Then, we use the CLIP model to obtain the exo image CLIP features as the cross-attention information for SVD and SD. We train these three baselines for each categories.

**Quantitative results.** We report PSNR, SSIM, and LPIPS with AlexNet [25] that measure the differences between generated ego frames and groundtruth on Tab. 1. Our Exo2Ego-V achieves the best performance for both unseen actions and unseen takes in terms of all metrics. It is noted PixelNeRF [60] achieves good PSNR scores since PSNR favors blurry images [38] as shown in Fig. 3 and 4. Most importantly, our Exo2Ego-V significantly outperforms SOTA approaches on all categories with an average of $35\%$ in terms of LPIPS, which clearly demonstrates the superiority of our Exo2Ego-V. We additionally evaluate our Exo2Ego-V and best-performing baseline SVD [5] on the H2O dataset [26] in Tab. 2. Our model still achieves the best performance, which demonstrate the generalizability of our method on different datasets.

Table 2: Quantitative comparison of our method against SVD on H2O dataset.

| | PSNR↑ | SSIM↑ | LPIPS ↓ |
|---|---|---|---|
| SVD [5] | 16.530 | 0.468 | 0.271 |
| Ours | 18.600 | 0.581 | 0.189 |

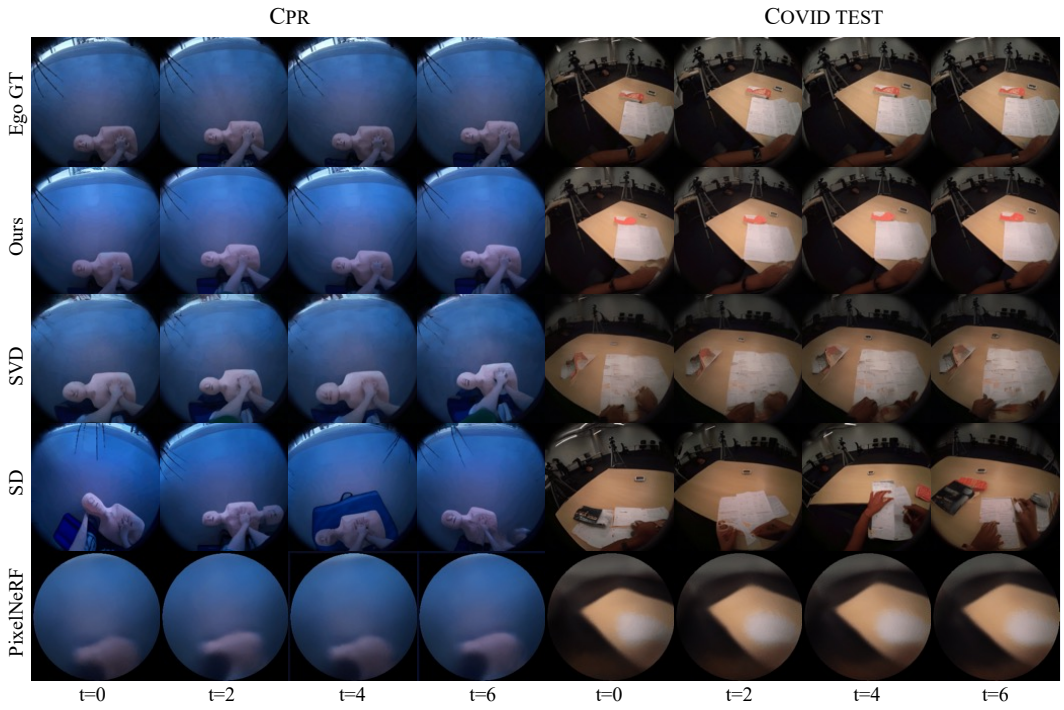


Figure 4: Qualitative comparisons of our method against SOTA approaches on **unseen takes**.

Table 1: Averaged quantitative evaluation on different categories. We color code each cell as **best**.

| | UNSEEN ACTION | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | COOKING | | | BASKETBALL | | | COVID TEST | | | CPR | | | BIKE | | |
| | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ |
| PixelNeRF [60] | 17.278 | 0.412 | 0.640 | 18.054 | 0.512 | 0.599 | 19.707 | 0.548 | 0.543 | 18.444 | 0.561 | 0.558 | 16.070 | 0.351 | 0.679 |
| SD [46] | 12.167 | 0.313 | 0.583 | 12.480 | 0.400 | 0.605 | 14.200 | 0.413 | 0.538 | 16.543 | 0.573 | 0.454 | 12.510 | 0.289 | 0.577 |
| SVD [5] | 14.318 | 0.407 | 0.519 | 15.529 | 0.491 | 0.533 | 16.584 | 0.507 | 0.477 | 17.807 | 0.630 | 0.397 | 14.541 | 0.364 | 0.516 |
| Ours | 17.367 | 0.493 | 0.408 | 20.062 | 0.624 | 0.249 | 21.462 | 0.668 | 0.235 | 18.533 | 0.647 | 0.305 | 16.310 | 0.413 | 0.486 |
| | UNSEEN TAKE | | | | | | | | | | | | | | |
| | COOKING | | | BASKETBALL | | | COVID TEST | | | CPR | | | BIKE | | |
| | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ |
| PixelNeRF [60] | 17.177 | 0.426 | 0.644 | 18.744 | 0.549 | 0.600 | 20.402 | 0.561 | 0.531 | 16.674 | 0.634 | 0.476 | 16.129 | 0.393 | 0.663 |
| SD [46] | 12.542 | 0.324 | 0.582 | 13.048 | 0.433 | 0.612 | 14.654 | 0.416 | 0.541 | 18.143 | 0.667 | 0.356 | 12.510 | 0.314 | 0.581 |
| SVD [5] | 14.554 | 0.416 | 0.532 | 16.563 | 0.546 | 0.555 | 16.875 | 0.500 | 0.490 | 18.302 | 0.680 | 0.353 | 14.492 | 0.383 | 0.527 |
| Ours | 17.712 | 0.504 | 0.456 | 21.417 | 0.646 | 0.300 | 21.590 | 0.667 | 0.245 | 18.473 | 0.711 | 0.219 | 16.343 | 0.441 | 0.489 |

**Qualitative results.** Fig. 3 and 4 visualizes the qualitative comparison of Exo2Ego-V over SOTA approaches on unseen actions and unseen takes, respectively, where Exo2Ego-V achieves substantially better egocentric videos quality than other approaches for all categories (see Sec. A.3 for more results). SVD [5] and SD [46] encounter significant difficulties by conditioning on the highly semantic exo images features to generate egocentric videos. In addition, SD [46] falls short in temporal consistency due to its 2D image-level generation. PixelNeRF [60] renders very blurry results due to the significant difficulty of sparse yet highly variable viewpoints and large occlusions. In addition, Fig. 6 visualizes the comparison of our method against SVD [5] on the H2O dataset [26]. Our method achieves the best performance and generates photorealistic hand-object interactions. Please see **supplementary video** for more results on video comparisons, which demonstrates the superiority of our Exo2Ego-V on both much higher spatial appearance quality and temporal consistency compared to other methods.

Table 3: Averaged quantitative evaluation on different categories against baselines for **unseen scenes**.

| | UNSEEN SCENE | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | COOKING | | | BASKETBALL | | | COVID TEST | | | CPR | | |
| | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ |
| PixelNeRF [60] | 13.393 | 0.355 | 0.700 | 16.091 | 0.418 | 0.676 | 14.885 | 0.438 | 0.697 | 14.349 | 0.348 | 0.745 |
| SD [46] | 10.617 | 0.268 | 0.587 | 11.686 | 0.334 | 0.640 | 13.072 | 0.409 | 0.624 | 14.937 | 0.428 | 0.586 |
| SVD [5] | 11.960 | 0.321 | 0.553 | 14.468 | 0.385 | 0.586 | 14.392 | 0.484 | 0.627 | 14.934 | 0.447 | 0.580 |
| Ours | 13.926 | 0.389 | 0.602 | 16.201 | 0.462 | 0.560 | 14.127 | 0.387 | 0.604 | 15.387 | 0.553 | 0.654 |

**Comparisons on unseen scenes.** We conduct additional experiments on unseen scenes of our Exo2Ego-V and baselines, and report PSNR, SSIM, and LPIPS with AlexNet [25] that measure the differences between generated ego frames and groundtruth on Tab. 3. We do not conduct experiments on Bike category since it is only captured on 4 different scenes, which are too few to generalize to new scenes. Fig. 5 visualizes the qualitative comparison of our Exo2Ego-V over SOTA approaches on unseen scenes. Our Exo2Ego-V achieves the best performance on most metrics in Tab. 3 and substantially better egocentric videos quality than other approaches as shown in Fig. 5. PixelNeRF [60] still renders very blurry results but gets good PSNR values since PSNR favors blurry images [38]. We also find that it is very challenging for all methods to evaluate on the unseen scenes due to the significant variance of new environments compared to the training set, and the lack of large-scale scene diversity from the training data.

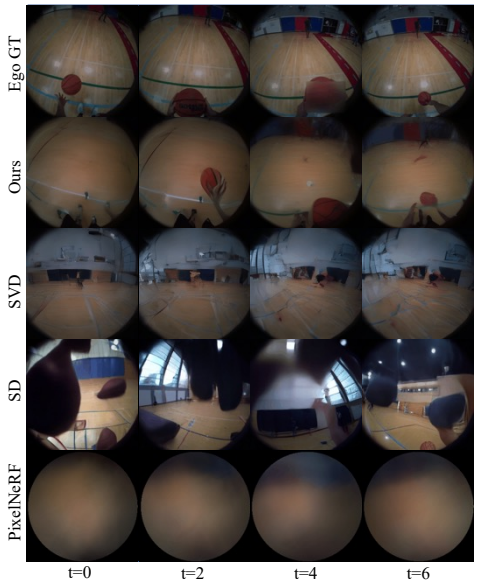


Figure 5: Qualitative comparisons of our method against SOTA approaches on **unseen scenes**.

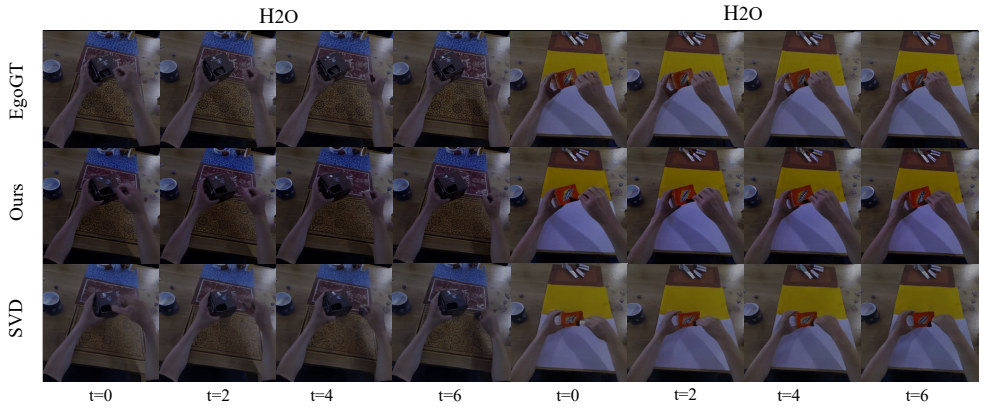


Figure 6: Qualitative comparisons of our method against SOTA approaches on H2O dataset.

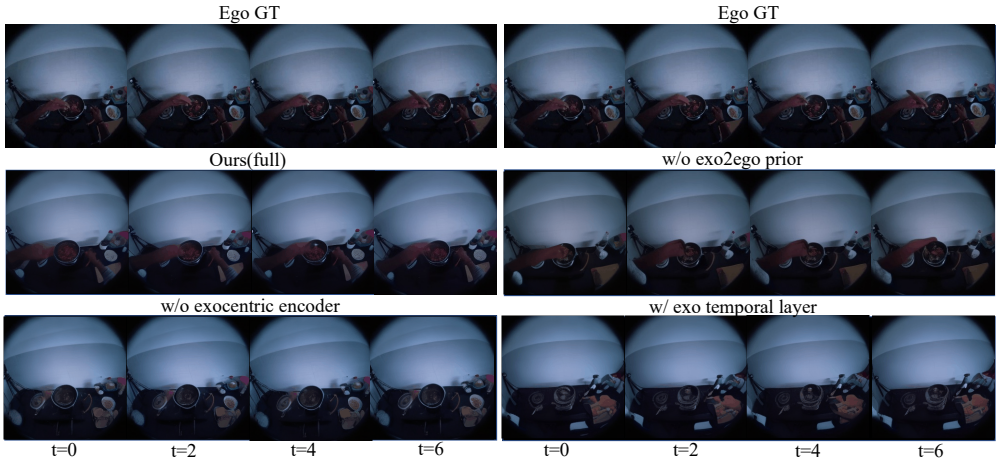


Figure 7: Qualitative ablation results of our method for cooking category on unseen actions.

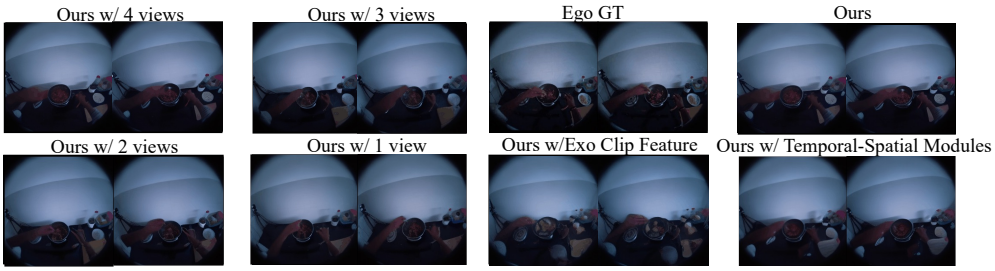


Figure 8: More ablation results of our method for cooking category on unseen actions.

### 4.3 Ablation study

We conduct ablation studies on the cooking category from Ego-Exo4D dataset [15]. We ablate on the proposed multi-view exocentric encoder, Exo2Ego view translation prior, and the temporal layers of multi-view exocentric encoder. As shown in Tab. 4, our full model achieves the best performance in terms of PSNR and SSIM. In addition, we provide the qualitative results of our ablations in Fig. 7 , which further demonstrates the effectiveness of our designs. Removing exocentric encoder results in inferior performance than full model, which clearly proves its capability in extracting dense and multi-scale exo features for ego video generation. Although removing exo2ego prior achieves the best LPIPS, it results in a cleaner but inaccurate egocentric video due to the lack of egocentric guidance, which improves the LPIPS but gets a lower PSNR. As shown in Fig. 7, removing exo2ego prior results in the missing of right arm. In addition, we ablate on adding temporal layers for our multi-view exocentric encoder and evaluate the temporal consistency by computing the CLIP image embeddings on our generated ego clips and report the average cosine similarity between all pairs of clip frames. Adding the exo temporal layers achieves a higher averaged temporal score of $0.924$ compared to $0.918$ of full model for unseen actions, demonstrating higher temporal consistency but at the expense of inferior image-level quality in Tab. 4. Thus, we disgard the exo temporal layer in final model.

We ablate on the number of exo views and replacing our exocentric feature encoder with CLIP features in Tab. 4 and Fig. 8. Our model with 4 exo views achieves the best performance, and our method achieves much better performance compared to the one using CLIP features. We also ablate on first performing temporal attention and then spatial attention for our model. The spatial-temporal model is slightly better than the temporal-spatial model in terms of PSNR and SSIM, and slightly worse for LPIPS. We follow the spatial-temporal attentions [16, 5].

Table 4: Ablation results of our method.

| | Cooking | | | | | |
|---|---|---|---|---|---|---|
| | Unseen Action | | | Unseen Take | | |
| | PSNR↑ | SSIM↑ | LPIPS ↓ | PSNR↑ | SSIM↑ | LPIPS ↓ |
| w/ Exo temporal layer | 16.762 | 0.473 | 0.507 | 16.784 | 0.489 | 0.584 |
| w/o Exocentric encoder | 16.777 | 0.455 | 0.441 | 16.734 | 0.450 | 0.499 |
| w/o Exo2ego prior | 17.268 | 0.493 | 0.364 | 17.668 | 0.502 | 0.401 |
| w/ 3 Views | 17.230 | 0.486 | 0.383 | 17.400 | 0.489 | 0.428 |
| w/ 2 Views | 16.930 | 0.474 | 0.399 | 17.020 | 0.479 | 0.445 |
| w/ 1 Views | 17.020 | 0.478 | 0.395 | 17.200 | 0.476 | 0.439 |
| w/ Exo CLIP | 16.540 | 0.456 | 0.425 | 16.410 | 0.445 | 0.480 |
| w/ Temporal-spatial | 17.000 | 0.484 | 0.402 | 17.290 | 0.490 | 0.443 |
| Ours (full) | 17.367 | 0.493 | 0.408 | 17.712 | 0.504 | 0.456 |

## 5 Conclusion

We introduced a novel framework of Exo2Ego-V, the first work to achieve Exo2Ego video generation for daily-life real-world skilled human activities. To tackle the challenges, we first proposed the new diffusion-based multi-view exocentric encoder to extract the dense multi-scale exocentric features as the appearance conditions. Then, we introduced an Exo2Ego view translation prior to provide coarse yet spatially aligned egocentric features as a concatenation guidance. Finally, we inserted temporal layers into Ego Unet for improved temporal consistency across ego frames. Exo2Ego-V produced significant improvements on challenging Ego-Exo4D dataset [15] over SOTA approaches.

**Limitations and future work.** Exo2Ego-V focuses on Exo2Ego video generation on several categories of skilled human activities. It remains challenging but is worthwhile researching on more general activities. Exploring Gaussian Splatting as translation prior is also a promising direction.

## 6 Acknowledgment

This project is supported by the Mike Zheng Shou's Start-Up Grant from NUS. Jia-Wei Liu is also supported by NUS IDS-ISEP scholarship. Thanks to Zihang Xia for helpful discussions.

## References


[1] Michael Abrash. Creating the future: Augmented reality, the next human-machine interface. In *2021 IEEE International Electron Devices Meeting (IEDM)*, pages 1–2. IEEE, 2021.

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

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

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

[5] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[6] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023.

[7] Joya Chen, Difei Gao, Kevin Qinghong Lin, and Mike Zheng Shou. Affordance grounding from demonstration video to target image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6799–6808, 2023.

[8] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *arXiv preprint arXiv:2303.04137*, 2023.

[9] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736, 2018.

[10] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, pages 1–23, 2022.

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[12] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5712–5721, 2021.

[13] Difei Gao, Lei Ji, Luowei Zhou, Kevin Qinghong Lin, Joya Chen, Zihan Fan, and Mike Zheng Shou. Assistgpt: A general multi-modal assistant that can plan, execute, inspect, and learn. *arXiv preprint arXiv:2306.08640*, 2023.

[14] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022.

[15] 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.

[16] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[18] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020.

[20] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022.

[21] Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. *arXiv preprint arXiv:2311.17117*, 2023.

[22] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. In *European Conference on Computer Vision*, pages 402–418. Springer, 2022.

[23] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

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

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

[26] 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*, pages 10138–10148, 2021.

[27] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6498–6508, 2021.

[28] 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.

[29] 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, 2021.

[30] Jia-Wei Liu, Yan-Pei Cao, Weijia Mao, Wenqiao Zhang, David Junhao Zhang, Jussi Keppo, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Devrf: Fast deformable voxel radiance fields for dynamic scenes. *arXiv preprint arXiv:2205.15723*, 2022.

[31] Jia-Wei Liu, Yan-Pei Cao, Jay Zhangjie Wu, Weijia Mao, Yuchao Gu, Rui Zhao, Jussi Keppo, Ying Shan, and Mike Zheng Shou. Dynvideo-e: Harnessing dynamic nerf for large-scale motion-and view-change human-centric video editing. *arXiv preprint arXiv:2310.10624*, 2023.

[32] Jia-Wei Liu, Yan-Pei Cao, Tianyuan Yang, Eric Zhongcong Xu, Jussi Keppo, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Hosnerf: Dynamic human-object-scene neural radiance fields from a single video. *arXiv preprint arXiv:2304.12281*, 2023.

[33] Hongchen Luo, Kai Zhu, Wei Zhai, and Yang Cao. Intention-driven ego-to-exo video generation. *arXiv preprint arXiv:2403.09194*, 2024.

[34] Mi Luo, Zihui Xue, Alex Dimakis, and Kristen Grauman. Put myself in your shoes: Lifting the egocentric perspective from exocentric videos. *arXiv preprint arXiv:2403.06351*, 2024.

[35] Priyanka Mandikal and Kristen Grauman. Dexvip: Learning dexterous grasping with human hand pose priors from video. In *Conference on Robot Learning*, pages 651–661. PMLR, 2022.

[36] 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. *Communications of the ACM*, 65(1):99–106, 2021.

[37] 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.

[38] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021.

[39] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021.

[40] 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.

[41] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9054–9063, 2021.

[42] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021.

[43] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. *arXiv preprint arXiv:2303.09535*, 2023.

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[45] 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.

[46] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10674–10685. IEEE, 2022.

[47] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023.

[48] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[49] Springer. *U-net: Convolutional networks for biomedical image segmentation*, 2015.

[50] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12959–12970, 2021.

[51] 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.

[52] 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*, pages 3446–3455, 2021.

[53] Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16210–16220, 2022.

[54] Benita Wong, Joya Chen, You Wu, Stan Weixian Lei, Dongxing Mao, Difei Gao, and Mike Zheng Shou. Assistq: Affordance-centric question-driven task completion for egocentric assistant. In *European Conference on Computer Vision*, pages 485–501. Springer, 2022.

[55] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Weixian Lei, Yuchao Gu, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. *arXiv preprint arXiv:2212.11565*, 2022.

[56] Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P Srinivasan, Dor Verbin, Jonathan T Barron, Ben Poole, et al. Reconfusion: 3d reconstruction with diffusion priors. *arXiv preprint arXiv:2312.02981*, 2023.

[57] 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.

[58] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. *arXiv preprint arXiv:2311.16498*, 2023.

[59] Zihui Sherry Xue and Kristen Grauman. Learning fine-grained view-invariant representations from unpaired ego-exo videos via temporal alignment. *Advances in Neural Information Processing Systems*, 36, 2024.

[60] 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*, pages 4578–4587, 2021.

[61] 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*, 45(6):6631–6646, 2020.

[62] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation, 2023.

[63] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022.

# A Supplemental material

## A.1 Implementation details

**Exo2Ego translation prior details.** Inspired by ReconFusion [56], our Exo2Ego translation prior $\Psi$ is based on a light PixelNeRF with 6-layer MLPs (Fig. 9) for higher efficiency together with a ResNet34 [17] pretrained on the ImageNet dataset to extract exo image features. For each synchronized timestep of the 4 exocentric videos and 1 egocentric video, we extract 4 exo frames and 1 ego frame and randomly sample 128 pixel rays from these 5 images and sample 3D points $\mathbf{x}$ for each iteration. Then, we add positional embedding to these points $\gamma(\mathbf{x})$ and query their latent features $\mathbf{f}(\mathbf{x})$ by projecting them on the latent images. Then we concatenate them with viewing direction as input to our translation prior as in Fig. 9. We conduct volumetric rendering at the third layer for the latent features, and then input the latent features to the last 3 layers for final pixel color and latent features.

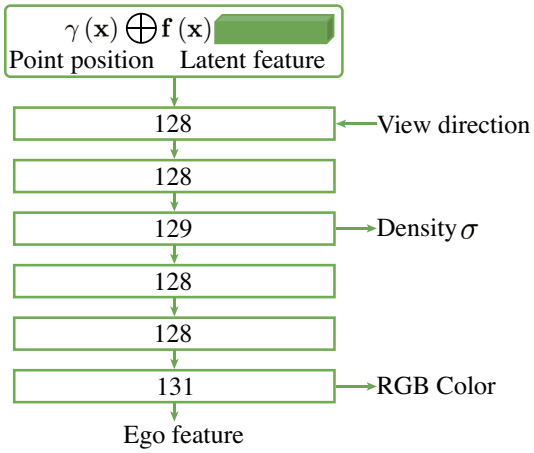


Figure 9: Network details for our Exo2Ego translation prior.

For egocentric video generation, we only sample rays from $32 \times 32$ ego frames so that the rendered ego features are spatially aligned with the noisy ego latent. Therefore, we concatenate the rendered ego features with the noisy ego latent as the input to our egocentric video diffusion pipeline.

**Egocentric camera unprojection.** The ego camera from Ego-Exo4D dataset [15] utilizes the FisheyeRadTanThinPrism (Fisheye624) model, which accounts for thin-prism distortion. This model includes four additional coefficients: $s_0$, $s_1$, $s_2$, $s_3$. The projection function is:

$$
\begin{aligned}
&\mathbf{u} = \mathbf{fx} * (\mathbf{u_r} + \mathbf{t_x}(\mathbf{u_r}, \mathbf{v_r}) + \mathbf{tp_x}(\mathbf{u_r}, \mathbf{v_r})) + \mathbf{c_x}, \\
&\mathbf{v} = \mathbf{fy} * (\mathbf{v_r} + \mathbf{t_x}(\mathbf{u_r}, \mathbf{v_r}) + \mathbf{tp_x}(\mathbf{u_r}, \mathbf{v_r})) + \mathbf{c_y}, \\
&\mathbf{tp_x}(\mathbf{u_r}, \mathbf{v_r}) = \mathbf{s_0}\mathbf{r}(\theta)^{\mathbf{2}} + \mathbf{s_1}\mathbf{r}(\theta)^{\mathbf{4}}, \\
&\mathbf{tp_y}(\mathbf{u_r}, \mathbf{v_r}) = \mathbf{s_2}\mathbf{r}(\theta)^{\mathbf{2}} + \mathbf{s_3}\mathbf{r}(\theta)^{\mathbf{4}}, \\
&\mathbf{r}(\theta) = \sqrt{(\mathbf{u} - \mathbf{c_x})^2/\mathbf{f_x^2} + (\mathbf{v} - \mathbf{c_y})^2/\mathbf{f_y^2}}, \\
&\phi = \arctan((\mathbf{u} - \mathbf{c_x})/\mathbf{f_x}, (\mathbf{v} - \mathbf{c_y})/\mathbf{f_y})
\end{aligned}
\tag{7}
$$

$\mathbf{u}, \mathbf{v}$ are the camera pixel coordinates and $\phi$, $\theta$ are the world point. $\mathbf{fx}$, $\mathbf{fy}$ are the focal lengths. Its parametrization contains 4 additional coefficients: $\mathbf{s_0}, \mathbf{s_1}, \mathbf{s_2}, \mathbf{s_3}$. Firstly we use the Newton method to calculate the $\mathbf{u_r}$, $\mathbf{v_r}$ and then we calculate $\phi$, $\theta$ using the above unprojection method. Finally, we sample 3D points along the calculated directions $\phi$, $\theta$.

## A.2 Training details

We optimize our Exo2Ego-V using Adam optimizer [24]. We set the learning rate of the multi-view exocentric encoder and the egocentric diffusion model as 0.00001, and we set the learning rate of view translation prior as 0.0001. We first train the translation prior with 500K iterations on a single A100 GPU for 36 hours, and then optimize our Exo2Ego spatial appearance translation with 500K iterations on 8 A100 GPUs for 48 hours, and finally finetune our temporal motion module with 100K iterations on 8 A100 GPUs for 40 hours, all using the PyTorch [40] deep learning framework.

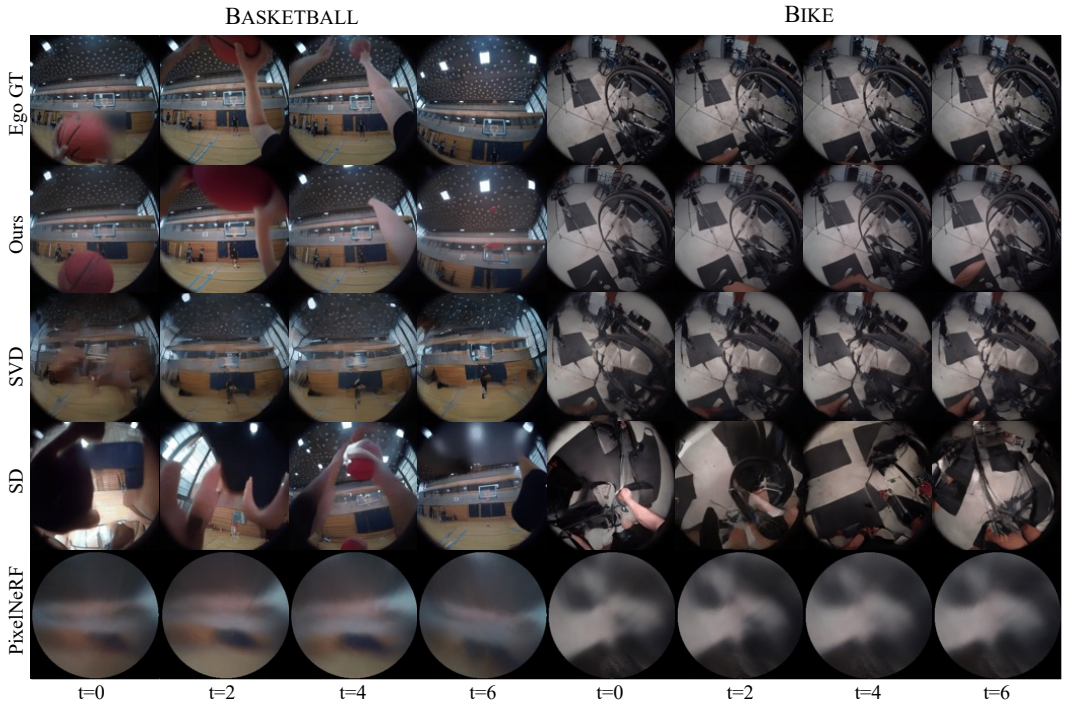


Figure 10: More qualitative comparisons of our method against SOTA approaches on **unseen actions**.

### A.3 Additional qualitative results

We provide more qualitative comparisons of Exo2Ego-V over SOTA approaches on unseen actions in Fig. 10. Our method achieves the best performance across all approaches.

### A.4 Feature Visualization

In Fig. 11, we present our ego feature visualization results using the Exo2Ego prior. The visualization results clearly represent the contents of the ego views. This indicates that our Exo2Ego prior can effectively extract and transmit the important information from ego views to the multi-view exocentric encoder.

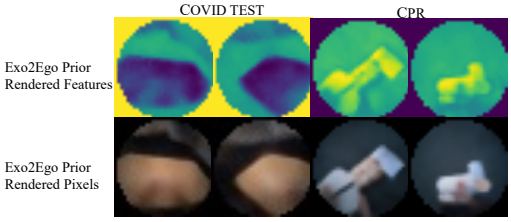


Figure 11: Exo2Ego prior feature visualization.

### A.5 Reasoning Efficiency

We provide the inference time comparison as shown in Tab. 5, where the inference time of our method to generate an 8-frame egocentric video is 9.06 second, which is comparable with other baselines. We believe it is feasible to use our model in offline applications to generate egocentric videos from the exocentric videos, such as capturing exocentric cooking videos and generating corresponding egocentric videos offline for cooking skills learning. Improving the inference speed towards real-time is very promising and we leave it as future works.

Table 5: Inference time of our method in comparison with baselines.

| | Ours | SVD [5] | SD [46] | PixelNeRF [60] |
|---|---|---|---|---|
| Inference time (second) | 9.06 | 4.26 | 6.91 | 5.65 |

### A.6 Preliminary on denoising diffusion probabilistic models (DDPMs)

DDPMs [19] are generative frameworks designed to synthesize data by reproducing a consistent forward Markov chain $x_1, \ldots, x_T$. The process begins from a random noise distribution and progres-

sively denoise the noisy contents to the clean data. Considering a data distribution as $x_0 \sim q(x_0)$, the Markov transition $q(x_t|x_{t-1})$ is conceptualized as a Gaussian distribution by a variance $\beta_t \in (0, 1)$:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbb{I}), \quad t = 1, \ldots, T. \tag{8}$$

Under the Bayes and Markov principles, the conditional probabilities can be derived as:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbb{I}),\ q(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}; \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t\mathbb{I}),\ t = 1, \ldots, T, \tag{9}$$

where $\alpha_t = 1 - \beta_t,\ \bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s,\ \tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t, \tilde{\mu}_t(x_t, x_0) = \frac{\sqrt{\bar{\alpha}_t}\beta_t}{1-\bar{\alpha}_t}x_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t$. DDPMs utilize a reverse approach to synthesize the chain $x_1, \ldots, x_T$:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)), \quad t = T, \ldots, 1. \tag{10}$$

The model parameters $\theta$ are optimized to ensure the synthesized reverse sequence aligns with the forward sequence.

### A.7 Broader impacts

Our work can generate egocentric videos from exocentric videos. Since the scale of egocentric dataset is still much less than the exocentric dataset, our method has the potential to improve the egocentric vision such as egocentric perception by generating more egocentric data from the exocentric data. Our Exo2Ego-V also support applications on AI assistant and augmented reality by generating egocentric videos from exocentric videos. We believe our method will not bring negative societal impacts.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main contribution of our paper is clearly described in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations in Conclusion.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: This paper does not contain any theoretical proof.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide dataset details in Sec. 4.1 and implementation details in Sec. A.1. We use publicly released Ego-Exo4D dataset for experiments, and our code and model will be made public if this paper gets accepted.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We use publicly released Ego-Exo4D dataset for experiments, and our code and model will be made public if this paper gets accepted. We also provide dataset details in Sec. 4.1 and implementation details in Sec. A.1.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provide detailed experimental setting in Sec. 4.1, implementation details in Sec. A.1, and training details in Sec. A.2.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We omitted error bars from our analysis due to the excessive computational expense involved in enumerating all experimented categories in the dataset.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We discuss computer resource information in Sec. A.2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We conduct experiments with publicly released dataset Ego-Exo4D that preserve anonymity of participants. Our research conforms with the NeurIPS Code of Ethics in every respect.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss potential societal impacts in Sec. A.7

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our work does not have such risk.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We properly cited papers and sources for existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide asset documentation alongside our code and model when we release them.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We do not involve crowdsourcing nor research with human subjects. We only use publicly released Ego-Exo4D dataset for experiments.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: We do not involve crowdsourcing nor research with human subjects. We only use publicly released Ego-Exo4D dataset that already handled IRB approvals for experiments.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.