EgoX: Egocentric Video Generation from a Single Exocentric Video

Supplementary Material

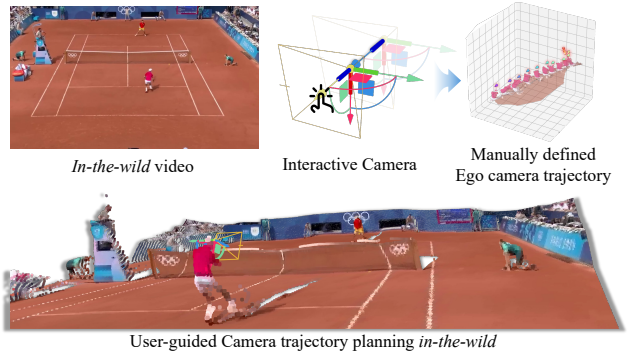


Figure 8. **In-the-wild Ego camera**. The ego camera for the in-the-wild example was obtained by interactively determining its extrinsic parameters using Viser [48].

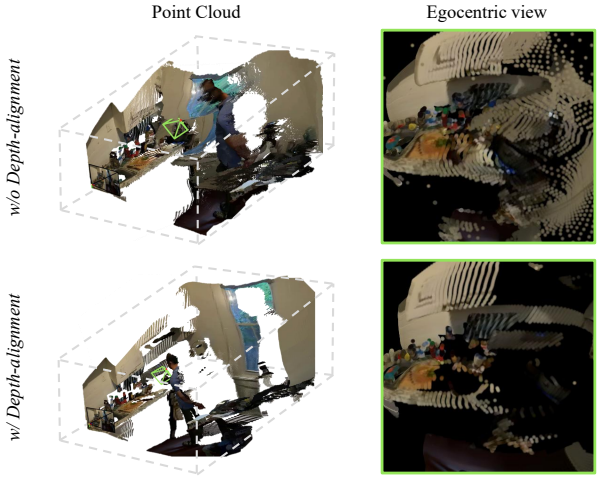


Figure 9. **Depth align comparison.** The above egocentric view is rendered from 3D point clouds across all frames. Without depth alignment, the inconsistent depth values between frames lead to unstable and unexpected camera movements.

F. Implementation Detail

F.1. GGA Implementation Detail

Applying Geometry-Guided Self-Attention (GGA) directly in pixel space is not feasible because the diffusion model operates in the latent space. Therefore, we compute 3D direction vectors at the pixel level and downsample them by averaging over each $4\times16\times16$ patch, matching the VAE downsampling factor include temporal dimension. The resulting patch-level direction vectors are used as geometric cues in the latent-space attention.

These geometric terms are precomputed once before the model inference to avoid runtime overhead. Additionally, applying the geometry-guided bias to all attention layers simultaneously would significantly increase memory usage and computational cost. To address this, we separately apply attention kernels for ego-to-exo and exo-to-ego attention, enabling efficient integration of geometric bias without exceeding memory constraints.

F.2. Ego Camera Pose for In-the-wild Example

Unlike the EgoExo4D [13] dataset, where ground-truth egocentric camera poses are provided, our in-the-wild examples do not include any ego camera pose annotations. To obtain the required egocentric poses for rendering, we manually determined the camera extrinsics using the 3D visualization toolkit Viser [48]. Specifically, we lifted the exocentric video into a 3D point cloud and interactively selected the camera pose that best matches the expected egocentric viewpoint, as illustrated in Fig. 8. As mentioned in Sec. 5, incorporating an automatic head-pose estimation module would be a valuable future extension. Potential options include video-based head-pose trackers [49] or SMPL [27]-based pose estimators, which could eliminate manual intervention and enable fully automatic exocentric-to-egocentric generation.

F.3. Evaluation Detail

In this section, we detail the evaluation procedure used to compute the Object Criteria, leveraging SAM2 [35] for object segmentation and DINOv3 [37] for appearance-based object matching.

Object Criteria

For each video, we perform object segmentation using SAM2 to obtain all valid object regions. For every detected object, we extract its bounding box and corresponding contour mask. Each object region is then cropped according to its bounding box and encoded into a feature vector $\mathbf{f} \in \mathbb{R}^d$ using a pretrained DINOv3 model.

To establish correspondences between the ground-truth egocentric video and the generated output, we compute cosine similarities for all possible pairs of object embeddings:

$$s_{i,j} = \frac{\mathbf{f}_i^{\text{GT}} \cdot \mathbf{f}_j^{\text{model}}}{\left\|\mathbf{f}_i^{\text{GT}}\right\|_2 \left\|\mathbf{f}_j^{\text{model}}\right\|_2}. \tag{8}$$

A pair (i, j) is considered a valid correspondence only if it satisfies a high-confidence appearance threshold $s_{i,j} \geq \tau_{\text{sim}}$, where we set $\tau_{\text{sim}} = 0.9$. These high-confidence

Method	Image Criteria				Object Criteria			Video Criteria			
	PSNR ↑	SSIM ↑	LIPIS ↓	CLIP-I ↑	Location Error ↓	IoU ↑	Contour Accuracy ↑	FVD ↓	Temporal Flickering ↑	Motion Smoothness ↑	Dynamic Degree ↑
EgoX (Ours)	**14.38**	**0.457**	**0.552**	0.877	**149.93**	**0.092**	**0.481**	**440.64**	**0.9813**	**0.9923**	**0.989**
w/o GGA	13.27	0.432	0.587	**0.880**	154.27	0.089	0.400	522.67	0.9812	0.9921	0.955
w/o Ego prior	13.01	0.401	0.581	0.855	171.95	0.059	0.351	523.00	0.9742	0.9908	0.843
w/o Clean latent	14.06	0.426	0.571	0.828	169.20	0.063	0.328	695.01	0.9811	0.9917	0.876

Table 3. **Ablation Study Results on Unseen Scenes.** The performance trends on unseen scenes are consistent with those observed on seen scenes. **Best** results are highlighted in bold.

matched object pairs form the basis for all downstream object-level metrics.

Location Error. For a valid matched pair, spatial alignment is measured using the Euclidean distance between the centers of the two bounding boxes. Let $\mathbf{c}_i^{\text{GT}}$ and $\mathbf{c}_j^{\text{model}}$ denote their centers. The location error is computed as:

$$\mathcal{E}_{i,j}^{\text{loc}} = \left\| \mathbf{c}_i^{\text{GT}} - \mathbf{c}_j^{\text{model}} \right\|_2 . \tag{9}$$

Lower values indicate better spatial consistency.

Bounding Box IoU. To measure coarse geometric consistency, we compute the Intersection over Union (IoU) between the two bounding boxes:

$$\text{IoU}_{i,j} = \frac{\text{Area}(B_i^{\text{GT}} \cap B_j^{\text{model}})}{\text{Area}(B_i^{\text{GT}} \cup B_j^{\text{model}})}. \tag{10}$$

Higher IoU indicates closer agreement in object position and scale.

Contour Accuracy. To evaluate fine-grained geometric consistency, we measure contour-level similarity using the object contours extracted by SAM2. For each matched object pair, SAM2 produces a contour mask, which we denote as C_i^{GT} and C_j^{model} for the ground-truth and generated frames, respectively. The contour IoU is then computed as:

$$\text{IoU}_{i,j}^{\text{contour}} = \frac{\left| C_i^{\text{GT}} \cap C_j^{\text{model}} \right|}{\left| C_i^{\text{GT}} \cup C_j^{\text{model}} \right|}. \tag{11}$$

This metric captures whether the object shape is preserved beyond the coarse bounding-box alignment.

F.4. Text Prompts

Since our method builds on the pretrained diffusion model [40], text prompts are required to condition the model. We generate these text prompts using a vision–language model (GPT-4o). The system prompt used for generating these descriptions is provided in Tab. 6, and examples of the resulting text prompts can be found in Fig. 18.

Figure 10. **GGA benefits example.** Without GGA, events occurring outside the visible region are attended to, leading to the generation of unwanted events in the ego view. With GGA, the model effectively focuses only on the visible region, thereby preventing the generation of these unwanted events.

G. In-depth Ablation Study

G.1. Ablation on Unseen Scene

To further evaluate the generalization capability of each component, we additionally conduct ablation experiments on unseen scenes. As shown in Tab. 3, the overall trends closely follow those observed in the seen-scene setting: removing any single component leads to noticeable degradation in geometric consistency, fidelity, or temporal coherence. These results confirm that all three components, geometry-guided attention, the egocentric prior, and the clean latent strategy, are essential for achieving coherent, high-fidelity egocentric video generation, even in challenging unseen environments.

G.2. Point cloud rendering

To construct accurate egocentric prior frames, we employ monocular depth estimation [41] combined with depth alignment from ViPE [16]. To validate the importance of depth alignment, we compare point cloud rendering with and without the alignment module. As shown in Fig. 9, without depth alignment, monocular depth predictions exhibit frame-wise scale inconsistencies, causing even static background regions to shift across frames. Although the ego camera remains fixed, misaligned depth introduces artificial camera motion, which can confuse the generative model and degrade viewpoint consistency. In contrast, applying depth alignment corrects these temporal inconsistencies by ensuring that the depth scale is coherent across frames. As a result, the rendered point clouds remain stable over time,

Method	Image Criteria				Object Criteria			Video Criteria			
	PSNR ↑	SSIM ↑	LIPIS ↓	CLIP-I ↑	Location Error ↓	IoU ↑	Contour Accuracy ↑	FVD ↓	Temporal Flickering ↑	Motion Smoothness ↑	Dynamic Degree ↑
EgoX (Ours)	**16.05**	**0.556**	**0.498**	0.896	**61.81**	**0.363**	**0.546**	**184.47**	**0.977**	**0.989**	**0.974**
w/o GGA	14.77	0.539	0.530	**0.897**	64.30	0.326	0.538	254.08	0.969	0.987	0.877
Prior width, Exo Channel	13.83	0.471	0.594	0.736	83.08	0.213	0.501	274.14	0.964	0.986	0.915
Prior width, Exo width	14.85	0.499	0.545	0.876	71.93	0.261	0.501	242.83	0.953	0.982	0.910
GGA only for inference	15.23	0.540	0.521	0.895	64.34	0.324	0.540	193.82	0.967	0.985	0.899

Table 4. **Additional Ablation Results.** The results from the conditioning strategy ablation and the GGA Training ablation are shown. These comparisons confirm that our integrated approach achieves the highest performance across all evaluated metrics. **Best** results are highlighted in bold.

providing a reliable egocentric prior for downstream video generation.

G.3. Conditioning Strategy Ablation

We evaluate how different conditioning strategies affect model performance by altering how the exocentric latent and the egocentric prior latent are combined. Conceptually, the exocentric view, whose spatial alignment with the egocentric target is not pixel-consistent and requires implicit warping, should be conditioned in a way that preserves its global spatial structure, making width-wise concatenation a natural choice. Conversely, the egocentric prior provides pixel-aligned viewpoint information, so channel-wise concatenation is better suited for injecting this fine-grained correspondence into the model.

To validate this intuition, we experiment with alternative fusion layouts. One variant reverses the two conditioning directions, applying channel-wise concatenation to the exocentric latent and width-wise concatenation to the egocentric prior. Another variant concatenates both inputs width-wise. We do not test the setting where both inputs are concatenated channel-wise, as this requires adding extra network modules such as zero-convs. When both inputs are concatenated width-wise, their combined latent becomes too large to fit in memory. Therefore, we resize the fused tensor to match the original exocentric latent shape. Additionally, when the exocentric view is concatenated channel-wise, geometry-guided attention cannot operate because the spatial structure needed for warping is lost, so this variant is evaluated without GGA.

As shown in Tab. 4 and Fig. 13, across all comparisons, our proposed conditioning strategy consistently delivers the strongest results. When the exocentric latent is fused channel-wise, the model fails to learn the necessary warping behavior and cannot properly utilize the exocentric conditioning. Similarly, width-wise concatenation of both latents diminishes the influence of the pixel-aligned prior and leads to quality degradation caused by confusion between global and local information. In contrast, our design, width-wise concatenation for exocentric latents and channel, wise fusion for egocentric priors, achieves the best geometric alignment, the most reliable conditioning behavior, and the highest visual quality.

Method	Ours	-GGA	-Ego Prior	-Clean Latent
Runtime	$\sim$ 10.5 min	$\sim$ 6.5 min	$\sim$ 6.5 min	$\sim$ 6.5 min

Table 5. **Comparison of runtime for each component.** Runtime for each component was measured on an NVIDIA H200 GPU

H. Additional Results

H.1. GGA Training Ablation

To understand the role of geometry-guided attention (GGA) during learning, we compare two settings: applying GGA only at inference time versus applying it during both training and inference. Because GGA serves as a guidance mechanism rather than a learnable module, one might expect it to be sufficient as an inference-only operation. However, when GGA is introduced solely at inference time, the model encounters an attention distribution it has never been trained to interpret. As shown in Tab. 4 and Fig. 13, this mismatch leads to a noticeable drop in visual fidelity and weaker geometric alignment. In contrast, enabling GGA during training allows the model to learn attention patterns that naturally incorporate geometric bias. As a result, the model produces sharper details, more stable reconstructions, and significantly more accurate geometry during egocentric generation.

H.2. Runtime

We measure runtime based on the denoising stage, which is the most computationally intensive component of our pipeline. Generating the egocentric prior takes less than 10 seconds and represents only a small fraction of the total processing time. To quantify the overhead of each component, we evaluate variants of our system that disable individual modules.

GGA introduces a moderate overhead due to the additional attention bias computation required at every attention layer. However, this cost is deemed highly reasonable

and necessary for the significant overall performance improvements observed both qualitatively and quantitatively, particularly in areas like geometry and appearance. Crucially, the GGA provides essential guidance to the model. As illustrated in Fig. 10, when generating the ego-view, the model without GGA may inadvertently attend to events occurring outside the visible region. This leads to the generation of unwanted events within the final ego-view. In contrast, GGA effectively guides the attention mechanism not to attend to these irrelevant regions, thereby preventing the generation of these undesirable events. This critical ability to ensure clean, accurate, and relevant ego-view generation makes the additional computational cost of GGA a worthwhile and necessary investment. Using the egocentric prior incurs a similar cost to the difference between an image-to-video and text-to-video diffusion model, as it increases the input conditioning dimensionality without modifying the model architecture. The clean latent strategy, however, adds no computational overhead, since it only modifies the noise scheduling during denoising without adding extra operations.

H.3. User Study

To further evaluate the generalization capability of our method, we conducted a user study involving 20 unseen-scene videos and 10 in-the-wild videos. A total of 19 participants were asked to choose the best video among the five methods, our method and four baselines, for each of the following criteria:

- **Reconstruction Accuracy** Which result best preserves the content visible in the exocentric video?
- **Motion/Camera Consistency** Which result best follows the motion and camera trajectory observed in the exocentric view?
- **Overall Quality** Which result provides the highest overall egocentric video quality?

As shown in Fig. 11, our method received the highest number of selections across all questions, significantly outperforming all baselines. These results demonstrate that our approach not only reconstructs view-relevant content more faithfully but also generalizes effectively to challenging unseen and in-the-wild scenarios.

H.4. Additional qualitative results

We include time-axis visualizations in Figs. 14, 15 and 17, which allow a clearer examination of temporal dynamics and overall video consistency. Consistent with the quantitative metrics, our method produces natural, high-fidelity egocentric videos with accurate geometry and stable motion. In contrast, Wan VACE often generates overly static videos with minimal dynamics, while other baselines either fail to properly incorporate the exocentric conditioning or exhibit noticeable artifacts and distortions. We also include additional in-the-wild examples for time-axis visualizations in Fig. 16.

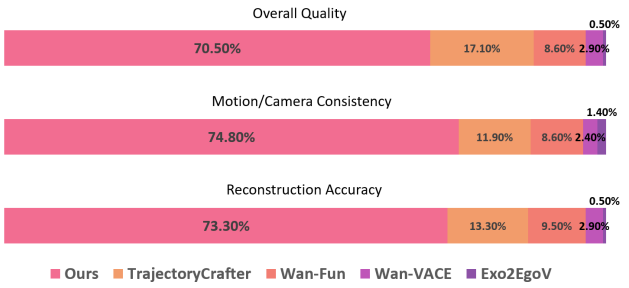


Figure 11. **User study results.** Our method received the highest number of selections across all questions, significantly outperforming all baselines.

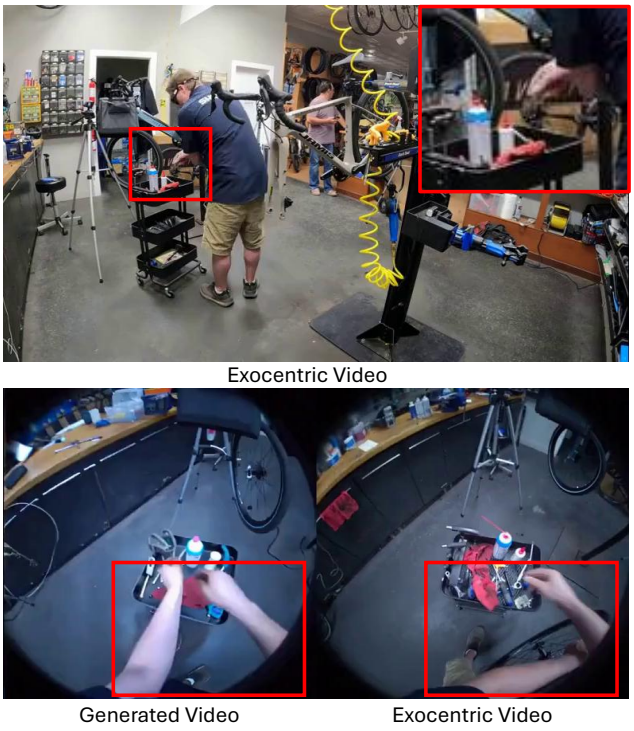


Figure 12. **Failure case due to task ambiguity.** The model's action misinterpretation occurs when it focuses on a small, subtle cue. This is not strictly a model failure, but rather a limitation imposed by the task's high ambiguity, where even a human observer might struggle to correctly infer the action based on such sparse visual evidence.

H.5. Failure example

Although our method performs robustly across diverse scenes, challenging real-world scenarios from datasets such as EgoExo4D [13] can still lead to occasional failure cases. These scenes often involve subjects facing away from

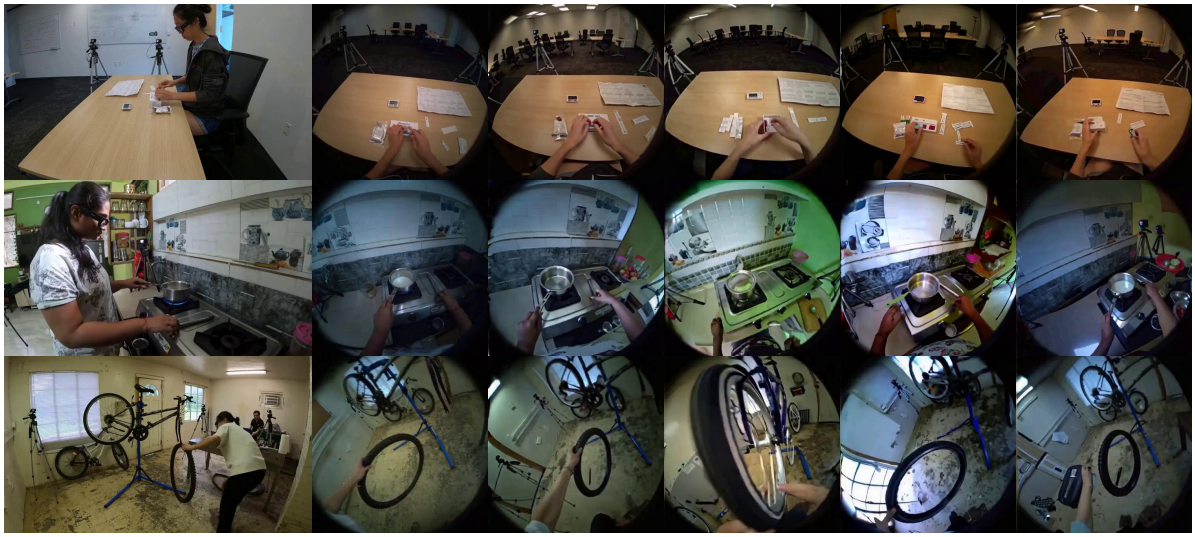


Figure 13. **Additional ablation qualitative comparison.** The qualitative results from the conditioning strategy ablation and the GGA Training ablation are shown. Model variants show limitations in geometric fidelity and detail reproduction, whereas our model consistently demonstrates the highest quality output.

the camera, rapid or complex body movements, or low-resolution details, making accurate cross-view reasoning extremely difficult. As illustrated in Fig. 12, when an exocentric frame contains ambiguous actions, such as a person bending one arm while the other arm is partially occluded, the model may misinterpret the configuration and generate an egocentric view with both arms extended. Such failure cases arise from inherent ambiguities in the exocentric input and the extreme viewpoint transformation required by the task.

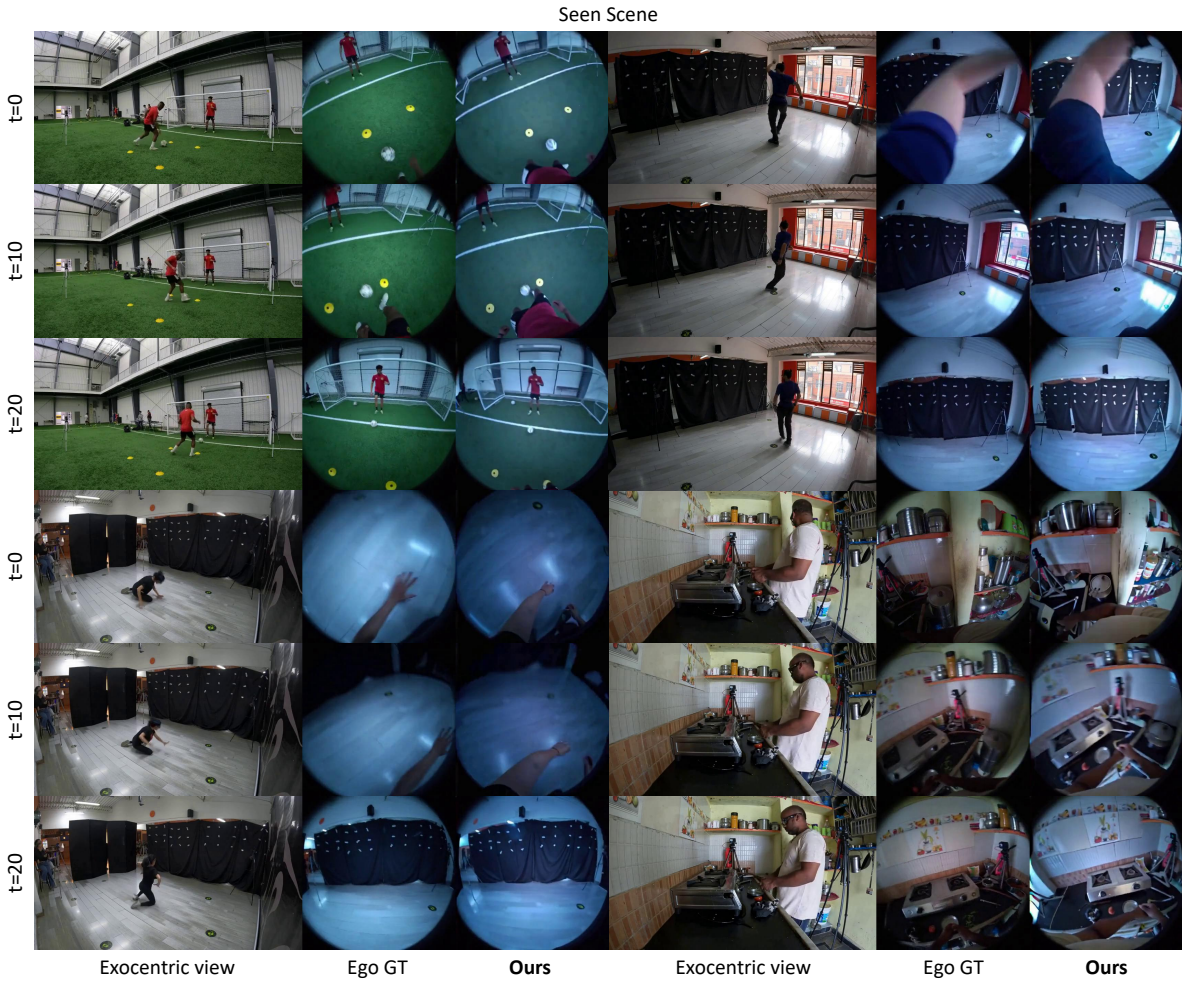


Figure 14. **Qualitative results for time sequence.** Our model accurately and seamlessly generates the entire time sequence.

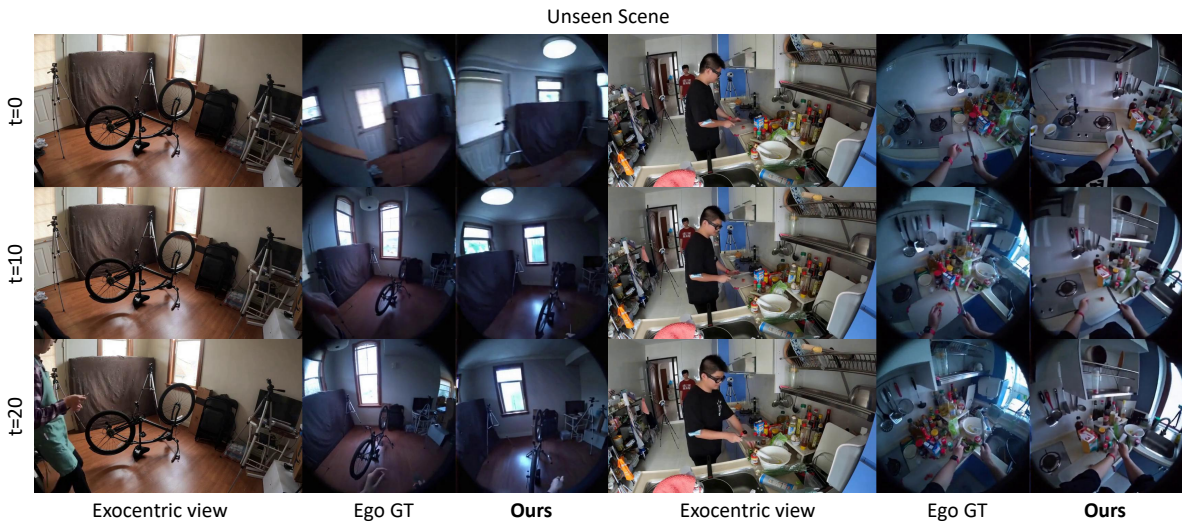


Figure 15. **Qualitative results for time sequence.** Our model accurately and seamlessly generates the entire time sequence.

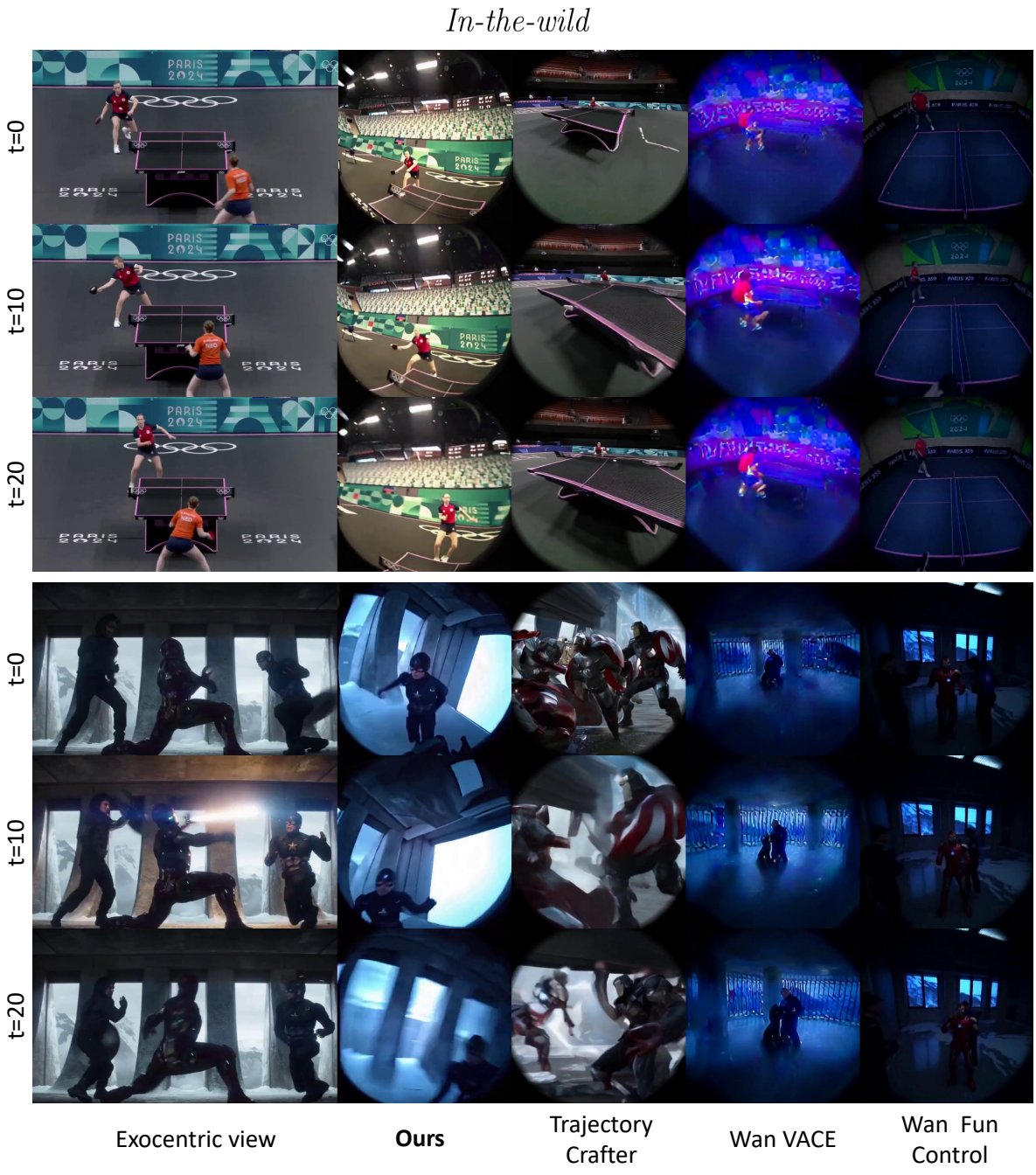


Figure 16. **Qualitative comparison with in-the-wild example.** Our model generates the entire time sequence accurately and seamlessly, even on challenging in-the-wild examples. Baselines struggle to maintain visual quality and accurate camera movements across all frames.

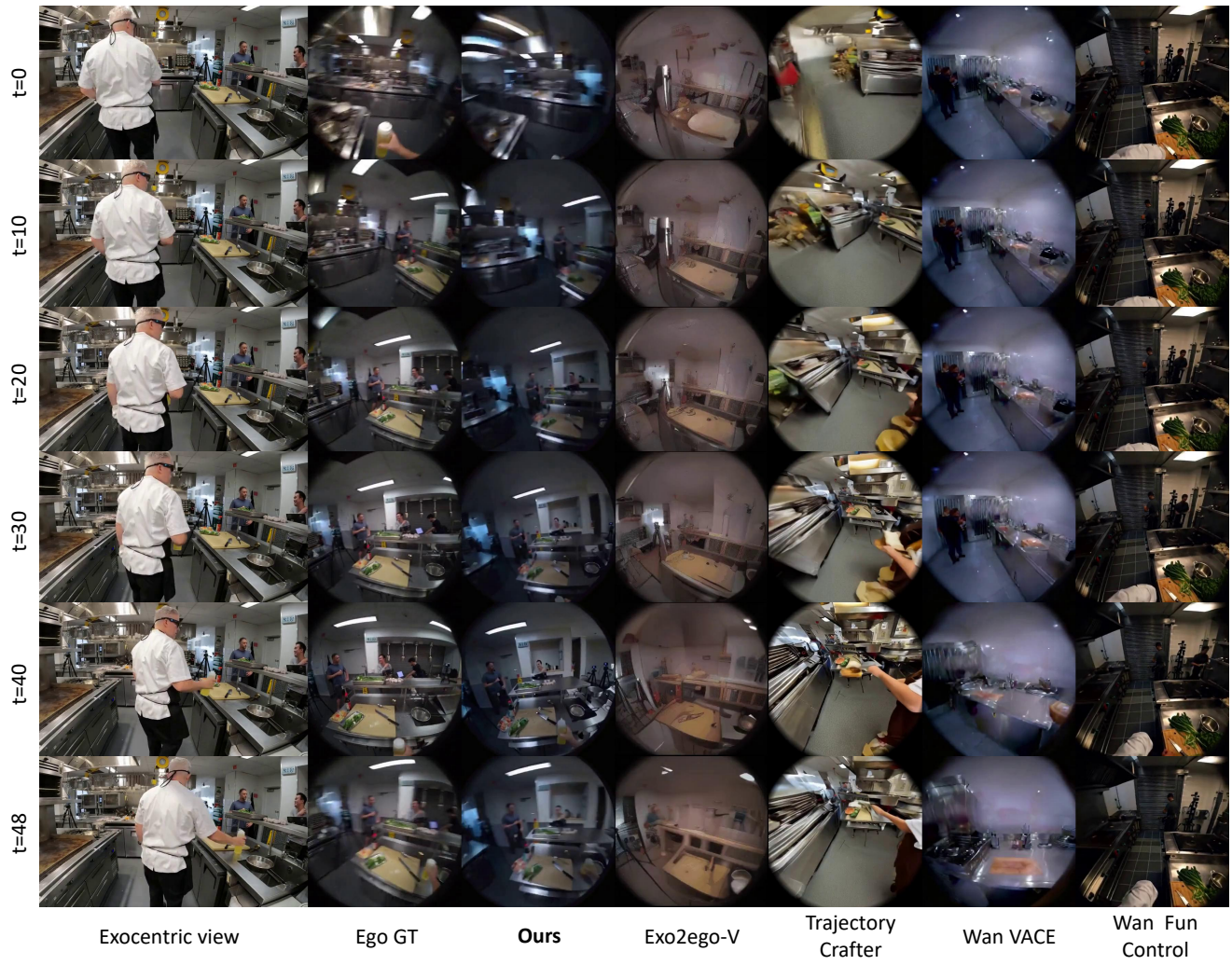


Figure 17. **Qualitative comparison for time sequence.** Our model accurately and seamlessly generates the entire time sequence. In contrast, other baselines struggle with maintaining high visual quality and generating accurate camera movements across all frames.

System Prompt to obtain exo and egocentric video prompt

You are a hyper-realistic scene reconstruction AI. Your task is to analyze a sequence of video frames provided in chronological order and produce a comprehensive, two-part analysis: a static scene overview followed by a dynamic, frame-by-frame action breakdown. Your guiding principle is **strict objectivity**.

— MISSION PROTOCOL —

Phase 1: Scene Establishment

First, analyze all provided frames to establish a detailed, static description of the physical environment. Detail the surfaces (walls, floors), furniture, and all unmoving background items. This is your 'establishing shot'.

Phase 2: Action Transition Analysis

After establishing the scene, provide a detailed description of the action progression and transitions observed across the sequence. Focus on how actions evolve, change, and flow from one moment to the next, maintaining awareness of the overall context established in Phase 1.

— CRITICAL DIRECTIVES —

1. Exhaustive Object Inventory: THIS IS YOUR MOST IMPORTANT TASK.

You must meticulously identify and catalog EVERY visible item.

- **NO GENERIC TERMS**: Do not use vague words like 'tool', 'box', 'utensil', or 'device'.
- **BE SPECIFIC**: Use precise names (e.g., 'smartphone', 'coffee mug', 'wooden spoon', 'cutting board', 'refrigerator', 'laptop computer', 'ceramic bowl', 'stainless steel knife').
- **DESCRIBE PROPERTIES**: Include colors, materials, textures, and positions (e.g., 'a blue ceramic mug on a granite countertop').

2. Focus on Hand-Object Interaction: THE ACTION'S CORE.

- For the '[Exo view]', your primary narrative focus MUST be the person's hands.** Describe their precise posture, movement, and interaction with objects (e.g., 'the person's right hand grasps the knife handle,' 'the left hand's fingertips stabilize the tomato').
- Every action description should revolve around what the hands are doing.

3. Strict Objectivity: DESCRIBE, DO NOT INTERPRET.

- **AVOID JUDGMENT**: Do not use subjective or abstract adjectives (e.g., AVOID 'modern', 'beautiful', 'cluttered', 'well-lit'). Describe only physical, measurable attributes.

4. Transition-Focused Analysis

- Analyze the sequence as a continuous flow of actions
- Describe how movements and interactions transition and evolve
- Focus on the progression and changes rather than individual frame descriptions
- Maintain narrative continuity throughout the sequence

— OUTPUT STRUCTURE —

You MUST follow this exact two-block format:

[Exo view]

Scene Overview: Detailed description of the static background environment from the third-person perspective. List all background objects.

Action Analysis: Describe the progression of actions and transitions observed throughout the sequence. Focus on how movements evolve, interactions change, and the flow of activities from beginning to end. Describe the continuous narrative of what is happening.

[Ego view]

Scene Overview: Detailed description of the static background environment from the first-person perspective. List all background objects.

Action Analysis: Describe the progression of actions and transitions observed throughout the sequence from the first-person perspective. Focus on how movements evolve, interactions change, and the flow of activities from beginning to end. Describe the continuous narrative of what is happening from the ego viewpoint.

{image}

Table 6. **System Prompt for VLM.** This is the system prompt used to generate the input text prompt for our model. Since the exocentric views were width-wise concatenated, the prompt describes both the exocentric and egocentric views.

Exocentric video

Egocentric video

[Exo view]
Scene Overview: The environment is a commercial kitchen. The walls are painted a neutral color, and the floor is a gray tile. The kitchen features stainless steel surfaces, including a large countertop and multiple cooking stations. There is a flat-top grill on the left side of the frame. In the background, there are several stainless steel...
Action Analysis: In the first frame, the person in the white shirt stands at the countertop, holding a bundle of green onions with their right hand while their left hand is positioned near their waist. The person appears to be engaged in conversation with two individuals in the background...
[Ego view]
Scene Overview: From the first-person perspective, the view is focused on the countertop in front of the individual. The cutting board is positioned directly in front, with green onions and a stainless steel knife visible. A bottle of oil and various condiments are arranged to the side...
Action Analysis: In the first frame, the hands are positioned to hold the green onions, with the right hand grasping the bundle while the left hand stabilizes it. The attention is drawn to the conversation occurring in the background. In the second frame, the grip on the green...

Exocentric video

Egocentric video

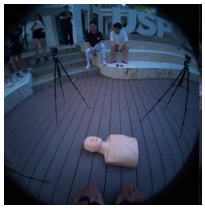
[Exo view]
Scene Overview: The environment is a classroom or training facility with a light-colored tiled floor. The walls are painted in a neutral tone, and there are several large windows covered with beige curtains. On the left side of the room, there is a whiteboard with handwritten notes and diagrams. Adjacent to the whiteboard, there are several cardboard boxes stacked against the wall...
Action Analysis: In the first frame, a person is kneeling on the blue mat, positioned over the CPR mannequin. Their hands are placed on the mannequin's chest, preparing for chest compressions. The individual is wearing a light blue shirt and dark pants ...
[Ego view]
Scene Overview: From the first-person perspective, the view is focused on the CPR training mannequin lying on a blue mat. The mannequin's head is facing towards the observer, and the chest area is clearly visible. The hands are positioned on the mannequin's chest,...
Action Analysis: In the first frame, the observer's hands are positioned on the mannequin's chest, preparing for the chest compressions. The fingers are interlocked, indicating readiness to begin the procedure. As the sequence transitions to the second frame,...

Exocentric video

Egocentric video

[Exo view]
Scene Overview: The scene is set in a basketball gymnasium with a polished wooden floor. The court features a red and gold design with the word "NCAA" prominently displayed in gold on the red surface. The walls are primarily glass, allowing natural light to illuminate the space, and there are large windows along one side of the gym. ...
Action Analysis: The sequence begins with a female player wearing a red jersey (number 35) standing near the free-throw line, holding a basketball. Another player in a black jersey is positioned near the hoop, while a third individual stands off to the side. ...
[Ego view]
Scene Overview: From the first-person perspective, the environment appears as a spacious basketball gymnasium with a polished wooden floor beneath. The court is marked with a red and gold design, prominently displaying "NCAA" in gold. The high ceiling is visible, with fluorescent lights illuminating the area. In front, the basketball hoop is clearly visible...
Action Analysis: From this perspective, the sequence begins with the player (myself) holding the basketball, preparing for a shot. The focus is on the grip of the right hand around the ball, while the left hand stabilizes it. As the second frame progresses, the player maintains the same posture, indicating readiness to shoot...

Exocentric video



Egocentric video

[Exo view]
Scene Overview: The scene is set outdoors on a wooden deck with a smooth, dark brown finish. The area is surrounded by greenery, including plants and trees visible in the background. To the left, there is a stone wall with a light-colored finish. In the foreground, a beige CPR training mannequin is positioned on the ground, lying face up. ...
Action Analysis: The sequence begins with an individual kneeling beside the CPR mannequin, which is positioned on the wooden deck. The person's right hand is poised above the mannequin, indicating an intention to perform a procedure...

[Ego view]
Scene Overview: From the first-person perspective, the view is directed towards a beige CPR training mannequin lying on the wooden deck. The surface of the deck is smooth and dark brown, contrasting with the light-colored stone steps visible in the background. ...
Action Analysis: From this perspective, the individual's hands are positioned above the mannequin, ready to initiate a training procedure. The right hand is likely preparing to make contact with the mannequin, while the left hand may be stabilizing the body or preparing to assist in the demonstration. The focus remains on the mannequin, with the surrounding group of observers visible in the peripheral view. ...

Exocentric video

Egocentric video (Gen)

[Exo view]
Scene Overview: The environment is a clinical or laboratory-like setting characterized by a smooth, gray ceiling with fluorescent lighting fixtures. The floor appears to be a polished concrete surface. In the background, there are various pieces of equipment, including a large, metallic apparatus with multiple arms and a control panel featuring a digital display. ...
Action Analysis: The sequence begins with the female figure holding a pendant-like object above the male subject's chest, with her right hand positioned to manipulate it. The pendant is suspended by a thin string, and the action suggests a focus on the interaction between the object and the subject. ...
[Ego view]
Scene Overview: From my first-person perspective, my right arm is extended forward, with my hand holding a small, silver, circular pendant suspended by a delicate chain. This pendant is positioned directly above the bare chest of a male subject, who is reclined in a chair to my right. On his chest, a larger, circular metallic device is visible, ...
Action Analysis: My hand is holding the pendant, carefully keeping it suspended above the male subject's chest. The initial action involves subtly adjusting the pendant's height and position, ensuring it's directly over the device on his chest. My focus is entirely on this precise placement, ...

Exocentric video

Egocentric video (Gen)

[Exo view]
Scene Overview: The scene is set on a street in front of a hospital, indicated by the large \"EMERGENCY\" sign visible in the background. The ground is asphalt, marked with yellow lines, and appears to have debris scattered across it, including small rocks and larger pieces of material. To the left, a gray sedan is parked, partially obscured by smoke. ...
Action Analysis: The sequence begins with a figure standing in the middle of the street, wearing a white uniform with red trim. The figure's hands are initially positioned in front of them, possibly holding an object that is not clearly visible. In the second frame, the figure raises a small metallic object, which appears to be a lighter or similar item, ...
[Ego view]
Scene Overview: From the inferred first-person perspective, the environment appears chaotic and filled with smoke. The \"EMERGENCY\" sign looms in the background, indicating proximity to a hospital. The ground is uneven with debris scattered around, and the asphalt is hot underfoot. ...
Action Analysis: Initially, the hands are positioned in front of the body, with the lighter held in the right hand. As the lighter is ignited, the flames from the explosion in the background become visible, creating a stark contrast against the smoke-filled scene. The left hand instinctively raises, possibly in a defensive gesture or to shield from the heat. ...

Figure 18. **Used Prompt Example**. This is the input text prompt for our model. Since the exocentric views were width-wise concatenated, the prompt describes both the exocentric and egocentric views.