EgoX: Egocentric Video Generation from a Single Exocentric Video

Anonymous CVPR submission

Paper ID 3508

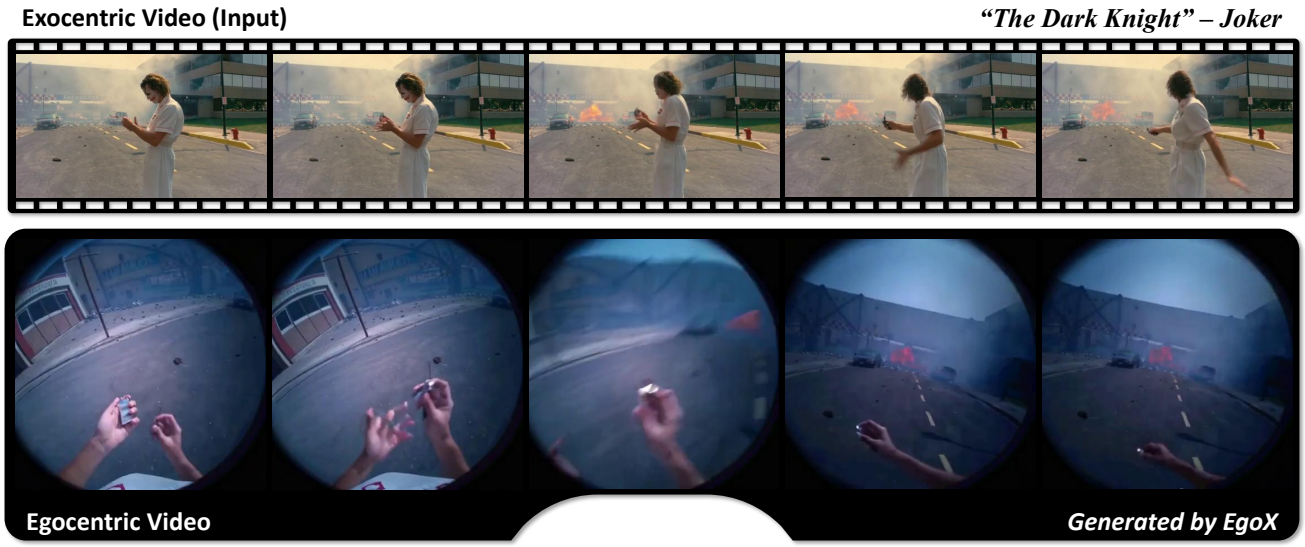

Figure 1. Given a single exocentric video, **EgoX** generates what the scene would look like from the actor's eyes. Shown with an in-the-wild clip from *The Dark Knight*, our approach achieves realistic and generalizable egocentric generation.

Abstract

Egocentric perception enables humans to experience and understand the world directly from their own point of view. Translating exocentric (third-person) videos into egocentric (first-person) videos opens up new possibilities for immersive understanding but remains highly challenging due to extreme camera pose variations and minimal view overlap. This task requires faithfully preserving visible content while synthesizing unseen regions in a geometrically consistent manner. To achieve this, we present ***EgoX****, a novel framework for generating egocentric videos from a single exocentric input. EgoX leverages the pretrained spatio–temporal knowledge of large-scale video diffusion models through lightweight LoRA adaptation and introduces a unified conditioning strategy that combines exocentric and egocentric priors via width- and channel-wise concatenation. Additionally, a geometry-guided self-attention mechanism selectively attends to spatially relevant regions, ensuring geometric coherence and high visual fidelity. Our approach achieves coherent and realistic egocentric video generation while demonstrating strong scalability and robustness across unseen and in-the-wild videos.*

1. Introduction

Don't you wish you could experience iconic scenes from films like *The Dark Knight* as if you were the *Joker* yourself? Exocentric-to-egocentric video generation makes this possible by converting a third-person scene into a realistic first-person perspective. This capability opens up new possibilities in the film industry, where viewers are no longer limited to passively watching a scene but can step into it and become the main character. They can become a superhero themselves or experience what it is like to play on the field as an MLB player. Beyond entertainment, egocentric perspectives are crucial in fields such as robotics and AR/VR, where understanding how the world appears from the actor's point of view enables better imitation, reasoning, and interaction [15, 21]. This stems from the fact that humans

perceive and interact with the world through a first-person, egocentric viewpoint.

However, generating such first-person perspectives is challenging, since the model must maintain scene consistency across views by reconstructing visible areas and realistically synthesizing unseen regions. A straightforward way to achieve this is to use a camera control model. Recent advances in camera control video generation models [18, 35, 47] have shown impressive performance in generating consistent views under moderate pose variations. However, these methods primarily focus on modest viewpoint changes, whereas exocentric-to-egocentric video generation requires extreme camera pose translation that drastically alters the visible field of view. This difference introduces two major challenges. First, extreme viewpoint shifts result in large unseen regions that must be plausibly synthesized based on scene understanding rather than direct observation. Second, only a small portion of the exocentric view corresponds to the egocentric perspective, making it crucial for the model to distinguish between view-related information that should be used as conditioning and unrelated content that should be suppressed. As illustrated in Fig. 2, effective generation therefore requires selectively attending to meaningful regions while discarding irrelevant background areas and plausibly synthesizing uninformed regions in a geometrically consistent manner. Therefore existing camera control models do not account for these challenges and thus often fail in exocentric-to-egocentric video generation.

Due to the inherent difficulty of this task, previous approaches often avoid generating the egocentric view from scratch or require additional inputs to simplify the problem. EgoExo-Gen [43] takes both an exocentric video and the first egocentric frame as inputs to generate only the subsequent sequence. Exo2Ego-V [26] utilizes four simultaneous exocentric camera views to capture richer spatial context and reduce the uninformed regions.

To address the limitations of previous approaches, we propose EgoX, a novel framework that generates egocentric video from a single exocentric video, achieving practical and generalizable egocentric generation from a single exocentric input. Our method leverages the pretrained spatio–temporal knowledge of large-scale video diffusion models with minimal modification, enabling the model to plausibly synthesize unseen regions in a geometrically consistent manner. Specifically, we design a unified conditioning strategy that combines exocentric views and egocentric priors through width-wise and channel-wise integration with clean latent representations, requiring only lightweight LoRA-based adaptation. Furthermore, a geometry-guided self-attention allows the model to focus on spatially relevant regions while suppressing unrelated areas, leading to coherent and high-fidelity egocentric video generation. By effectively leveraging pretrained weights, our approach produces high-quality egocentric videos and demonstrates strong generalization across diverse environments, including challenging in-the-wild scenarios, as illustrated in Fig. 1.

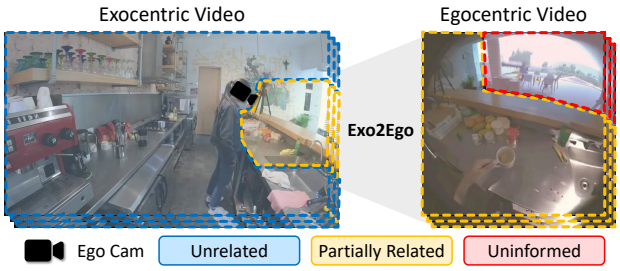

Figure 2. **Exo-to-Ego view generation example.** The model has to preserve view-related content from the exocentric input, generate uninformed regions realistically, and ignore unrelated areas for consistent egocentric synthesis.

To summarize, the major contributions of our paper are as follows:

- We propose a novel framework **EgoX** for synthesizing high-fidelity egocentric video from a *single* exocentric video by effectively exploiting the pretrained video diffusion models.
- We design a unified conditioning strategy that jointly combines exocentric video and egocentric priors through width-wise and channel-wise integration, achieving robust geometric consistency and high-quality generation.
- We introduce a geometry-guided self-attention and clean latent representations that selectively focuses on view-relevant regions and enhances accurate reconstruction, leading to more coherent egocentric synthesis.
- Extensive qualitative and quantitative experiments demonstrate that **EgoX** outperforms previous approaches by a large margin, achieving *state-of-the-art* performance on diverse and challenging exo-to-ego video generation benchmarks.

2. Related Work

2.1. Exo-to-Ego View Generation

Prior works on exo-to-ego view generation have explored various conditioning mechanisms and task formulations to bridge the significant viewpoint gap. Some approaches [26, 27, 31] incorporate exocentric features by concatenating them channel-wise with the egocentric representation. However, this method struggles with the fundamental lack of pixel-wise correspondence between the two viewpoints. This spatial misalignment makes it difficult for the model to effectively leverage the conditioning information, often leading to a poor understanding of the scene geometry, which can result in overfitting or a degradation in output quality. Other works, such as 4Diff [10], employ cross-attention mechanisms to condition the generation on exo-

centric views. This approach, however, prevents the utilization of powerful pretrained diffusion weights, limiting its generalizability and resulting in lower-quality synthesis.

To address these limitations, other methods utilize reference frames or multi-view conditions. For instance, EgoExo-Gen [43] require the first egocentric frame to generate the rest of the sequence. Exo2Ego-V [26] performs full video translation but relies on four exocentric video inputs and separately trained spatial and temporal modules, which limits its generalization and fails to fully exploit spatio–temporal priors. In contrast, our model generalizes effectively using pretrained video diffusion weights while requiring only a single exocentric input.

2.2. Video Diffusion Models

Recent advancements in video diffusion models [1, 5, 6, 14, 39, 45] have led to significant improvements in generative quality, producing highly realistic and coherent video sequences. This has spurred a wide range of research exploring how to utilize these powerful generative capabilities in various applications [9, 19, 20, 32, 49]. A key area of this research focuses on conditional video generation, where the synthesis process is guided by specific inputs. Many works [7, 19, 22, 44, 49] have demonstrated successful control using conditions such as depth maps or static images.

Building on this, several methods have been proposed for camera-controlled video generation [4, 28, 47]. These approaches can be broadly categorized into two main groups. The first group [3, 4, 29, 42, 46] conditions the diffusion model directly on camera extrinsic parameters, often represented as raw matrices or Plücker coordinates. The second group [18, 25, 35, 41, 47] first lifts the input video into an intermediate 3D representation, such as a point cloud. This 3D scene is then rendered from a new, user-specified camera pose, and the resulting image is used as a strong spatial condition to guide the final video generation.

However, existing methods for camera control are primarily designed for modest changes in viewpoint. They struggle to handle the extreme camera pose differences, a challenge that becomes particularly significant in exocentric-to-egocentric video generation. Our work addresses this critical gap by proposing a model capable of generating coherent egocentric videos from a significantly different exocentric perspective.

3. Method

Given an exocentric video sequence $X = \{X_i\}_{i=0}^{F}$ and egocentric camera pose $\phi = \{\phi_i\}_{i=0}^{F}$, the goal is to generate a corresponding egocentric video sequence $Y = \{Y_i\}_{i=0}^{F}$ that depicts the same scene from a first-person viewpoint. The key challenge is to preserve the visible content in the exocentric view while synthesizing unseen regions in a geometrically consistent and realistic manner. To this end, the exocentric sequence X is first lifted into a 3D representation and rendered from the target egocentric viewpoint (Sec. 3.1), which becomes an egocentric prior video P. Both P and the original exocentric video X are then provided as inputs to a video diffusion model (Sec. 3.2). In addition, a geometry-guided self-attention (Sec. 3.3) is proposed to adaptively focus on view-consistent regions and enhance feature coherence across perspectives.

3.1. Egocentric Point Cloud Rendering

For this stage, we render an egocentric prior video $P \in \mathbb{R}^{F\times 3\times H\times W}$ via point cloud rendering from the exocentric view. This prior provides both explicit pixel-wise RGB information and implicit camera trajectory cues that guide viewpoint alignment. Specifically, we first estimate a monocular depth map $D^m \in \mathbb{R}^{F\times H\times W}$ for each frame using a single-image depth estimator [40], and a video-based depth map $D^v \in \mathbb{R}^{F\times H\times W}$ using a temporal depth estimator [8]. Because D^m is estimated independently per frame, depth values often exhibit slight inconsistencies across time. In contrast, D^v produces a temporally smooth yet affine-invariant depth estimate. To combine the advantages of both, we temporally align D^v with D^m. Following [16], we optimize affine transformation parameters α, β using a momentum-based update strategy, yielding $\hat{\alpha} = \{\hat{\alpha}_f\}_{f=0}^{F}$ and $\hat{\beta} = \{\hat{\beta}_f\}_{f=0}^{F}$, which represent the per-frame affine transformations. The final aligned depth is computed as:

$$D^f = \frac{1}{\hat{\alpha}/D^v + \hat{\beta}}, \tag{1}$$

where D^f denotes the final aligned depth map. Dynamic objects are masked out so that only static background regions are used during both alignment and rendering. For further details, please refer to [16].

After obtaining the aligned depth map D^f, we convert it into a 3D point cloud representation using the corresponding camera intrinsics. We then render the egocentric prior frames using a point cloud renderer [33]:

$$P = \text{render}(X, D^f, \phi), \tag{2}$$

where $X \in \mathbb{R}^{F\times 3\times H\times W}$ is the exocentric RGB video and ϕ is egocentric camera poses.

3.2. Exo-to-Ego View Generation with VDM

As illustrated in Fig. 3, the model takes an exocentric video $X \in \mathbb{R}^{F\times 3\times H\times W}$ and the egocentric prior video $P \in \mathbb{R}^{F\times 3\times H\times W'}$ as conditioning inputs. Both inputs are encoded by a frozen VAE encoder, producing latent features $x_0 \in \mathbb{R}^{f\times c\times h\times w}$ and $p_0 \in \mathbb{R}^{f\times c\times h\times w'}$, respectively. These latents are then concatenated with the noisy latent $z_t \in \mathbb{R}^{f\times c\times h\times w'}$ to form the input of the diffusion model.

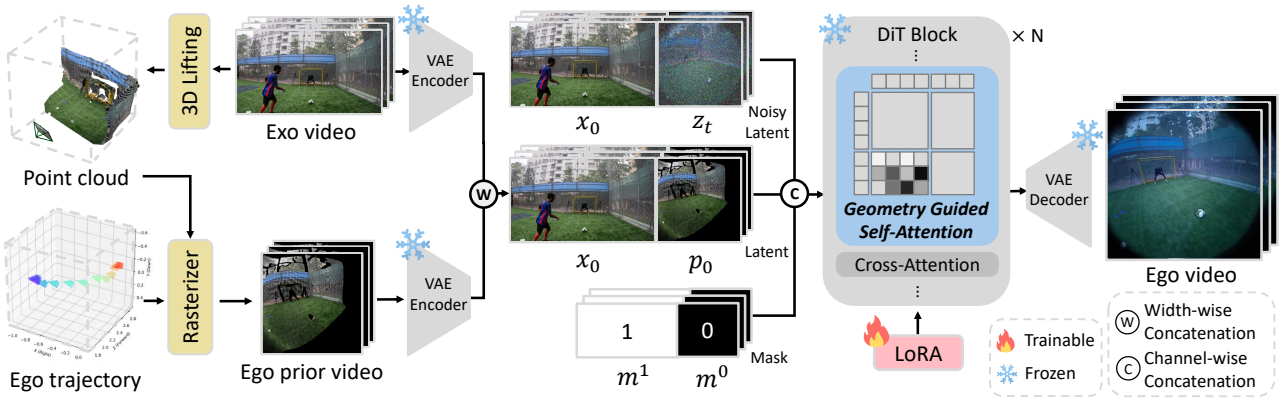

Figure 3. **Overall pipeline.** Given an exocentric video input, we first lift it into a 3D point cloud and render the scene from the egocentric viewpoint to obtain the egocentric prior video. The clean exocentric video latent and the egocentric prior latent are combined via width-wise and channel-wise concatenation in the latent space, and then fed into a pretrained video diffusion model equipped with the proposed geometry-guided self-attention.

The egocentric prior latent p_0 shares the same viewpoint as the target egocentric video and therefore preserves pixel-wise correspondence. We concatenate p_0 with z_t along the channel dimension, providing viewpoint-aligned and temporally coherent guidance during generation. Although p_0 offers explicit geometric cues for the regions visible in the rendered ego view, it remains noisy and lacks substantial portions of the scene. To complement the missing information in the rendered egocentric view, we further use the exocentric video latent x_0 to provide broader scene context. Since the viewpoint of x_0 differs from that of the noisy egocentric latent z_t, their features are not pixel-wise aligned. Therefore, we concatenate x_0 with z_t along the width dimension, encouraging the model to infer cross-view correspondences and perform spatial warping implicitly. Unlike [17], which utilizes SDEdit [30] by concatenating a noisy conditioning latent with a noisy target latent for conditional generation, our method concatenates the clean latent x_0 with the noisy z_t throughout all denoising timesteps, while only z_t is updated and x_0 remains fixed. This design encourages the model to consistently reference fine-grained details from x_0, enabling more accurate and reliable spatial warping.

The overall relation between inputs and outputs is defined as:

$$z_{t-1} = f_\theta(x_0, z_t | x_0, p_0 | m^1, m^0), \tag{3}$$

where f_θ denotes a single-step denoising function of the VDM, x_0 is the exocentric video latent, p_0 is the egocentric prior latent, and m is the binary mask specifying whether each spatial region is used for conditioning or for synthesis. Once the sampling is complete, we remove the exocentric part of the latent and decode only the egocentric part to obtain the final result.

3.3. Geometry-Guided Self-Attention

As mentioned in Sec. 1, the exocentric video condition includes irrelevant regions that can distract the model during exo-to-ego view generation. To address this, we introduce a Geometry-Guided Self-Attention (GGA) that adaptively emphasizes spatially corresponding regions between exocentric and egocentric representations. When egocentric query tokens $q_{\text{ego}} \in \mathbb{R}^{l \times c}$ attend to exocentric key tokens $k_{\text{exo}} \in \mathbb{R}^{l' \times c}$, the attention should jointly account for semantic similarity (i.e., appearance) and 3D spatial alignment. Ideally, tokens that are both semantically similar and geometrically aligned with the egocentric viewpoint should receive higher attention weights, while unrelated or misaligned regions are suppressed to ensure geometric consistency and realism in the generated views.

To achieve this, we leverage self-attention augmentation with 3D geometric cues. Using the 3D point cloud obtained in Sec. 3.1, we compute 3D direction vectors from the ego camera centers $c = \{c_i\}_{i=0}^{F}$, $c_i \in \mathbb{R}^3$ in world space to each query and key token position, $\tilde{q}, \tilde{k} \in \mathbb{R}^3$. The unit direction vectors are defined as $\hat{q} = \frac{\tilde{q}-c_i}{\|\tilde{q}-c_i\|_2}$, $\hat{k} = \frac{\tilde{k}-c_i}{\|\tilde{k}-c_i\|_2}$,. We then compute the cosine similarity between the two direction vectors and incorporate it into the attention computation as a multiplicative geometric prior.

Specifically, the modified attention logits are formulated as:

$$s'_{m,n} = s_{m,n} + \log\big(g(\hat{q}_m, \hat{k}_n) \cdot \lambda_g\big), \tag{4}$$

$$g(\hat{a}, \hat{b}) = \cos\_\text{sim}(\hat{a}, \hat{b}) + 1, \tag{5}$$

where $s_{m,n} = \frac{q_m^\top k_n}{\sqrt{c}}$ denotes the standard attention logits [38] and λ_g is a hyperparameter that balances this geometry bias term defined in Eq. (5). We add one to the cosine similarity term to ensure positive values before taking the logarithm.

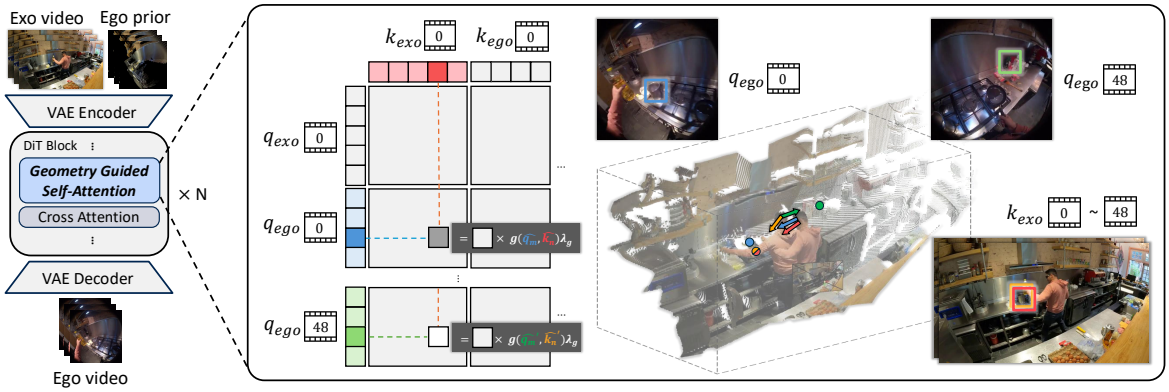

Figure 4. **Geometry-Guided Self-Attention Overview.** 3D direction similarities between egocentric queries and exocentric keys are used as an additive bias in the attention map, guiding the model to focus on geometrically aligned regions. Although the orange and red directions are the same key tokens, their directions differ due to different camera centers. The blue–red pairs have similar directions and thus receive higher scores, whereas the green–orange pairs have opposite directions and obtain lower scores.

Finally, given an egocentric query q_m and an exocentric key k_n, the attention weight $a_{m,n}$ is computed as:

$$a_{m,n} = \frac{\exp(s'_{m,n})}{\sum_{j=1}^{l} \exp(s'_{m,j})} \tag{6}$$

$$= \frac{\exp(s_{m,n})\, g(\hat{q}_m, \hat{k}_n)\, \lambda_g}{\sum_{j=1}^{l} \exp(s_{m,j})\, g(\hat{q}_m, \hat{k}_j)\, \lambda_g}. \tag{7}$$

This formulation allows the attention mechanism to be explicitly guided by geometric alignment between query and key directions, improving spatial consistency and visual coherence across views.

In image generation, spatial relationships can be encoded by multiplying rotation matrices to each query and key before attention, as done in [10, 23, 24, 37]. However, in video generation, the camera center of q_{ego} changes at every frame, making it necessary to compute key directions relative to each query separately. This implies that the geometry bias term should be recomputed for every query–key pair within each frame's attention operation. As illustrated in Fig. 4, even k_{exo} located at the same position (e.g. red) may have entirely different direction vectors (e.g. red and orange) depending on the camera pose. To handle this, we compute all pairwise direction similarities between k_{exo} and q_{ego} and use this term as an additive bias attention mask, allowing us to reuse optimized attention kernels. This formulation provides a precise geometry-guided self-attention that effectively aligns exocentric and egocentric representations.

4. Experiments

In the following sections, we aim to answer the following research questions that guide our experimental evaluation:

- How does our method outperform existing baselines in both qualitative and quantitative evaluations? (Sec. 4.2, Sec. 4.3)
- How accurately does the model reconstruct regions visible in the exocentric view? (Sec. 4.1, Sec. 4.3)
- How well does the model generalize to unseen scenes and challenging in-the-wild videos? (Sec. 4.2, Sec. 4.3)
- How does each proposed component contribute to overall performance and generation quality? (Sec. 4.4)

4.1. Experimental Setup

Implementation Details. To support channel-wise concatenation of noisy latent and ego prior latent, we adopt the inpainting variant of Wan 2.1 (14B) Image-to-Video model [39] as our base model. We fine-tuned the model using LoRA (rank = 256) with a batch size of 1, and a single day on 8 H200 (140 GB) GPUs. For the dataset, we curated 4,000 clips from Ego-Exo4D [12] covering diverse scenes and actions, using 3,600 clips for training and 400 for testing. Additionally, we collected 100 unseen clips that are not included in the training set to evaluate generalization performance. More detailed information can be found in Sec. F.

Baselines. Among existing exocentric-to-egocentric video generation approaches, Exo2Ego-V [26] and EgoExo-Gen [43] serve as representative baselines. We adopt Exo2Ego-V as our primary baseline, as EgoExo-Gen does not provide publicly available implementation. With the rapid progress in conditional video generation and camera control models, several recent methods have demonstrated performance comparable to or even surpassing Exo2Ego-V. Therefore, we additionally included Trajectory Crafter [47], a state-of-the-art camera control model, as well as Wan Fun Control [2] and Wan VACE [19], which offer distinct conditioning approach. Wan Fun Control applies channel-wise concatenation for conditioning, and Wan VACE employs an auxiliary conditioning network,

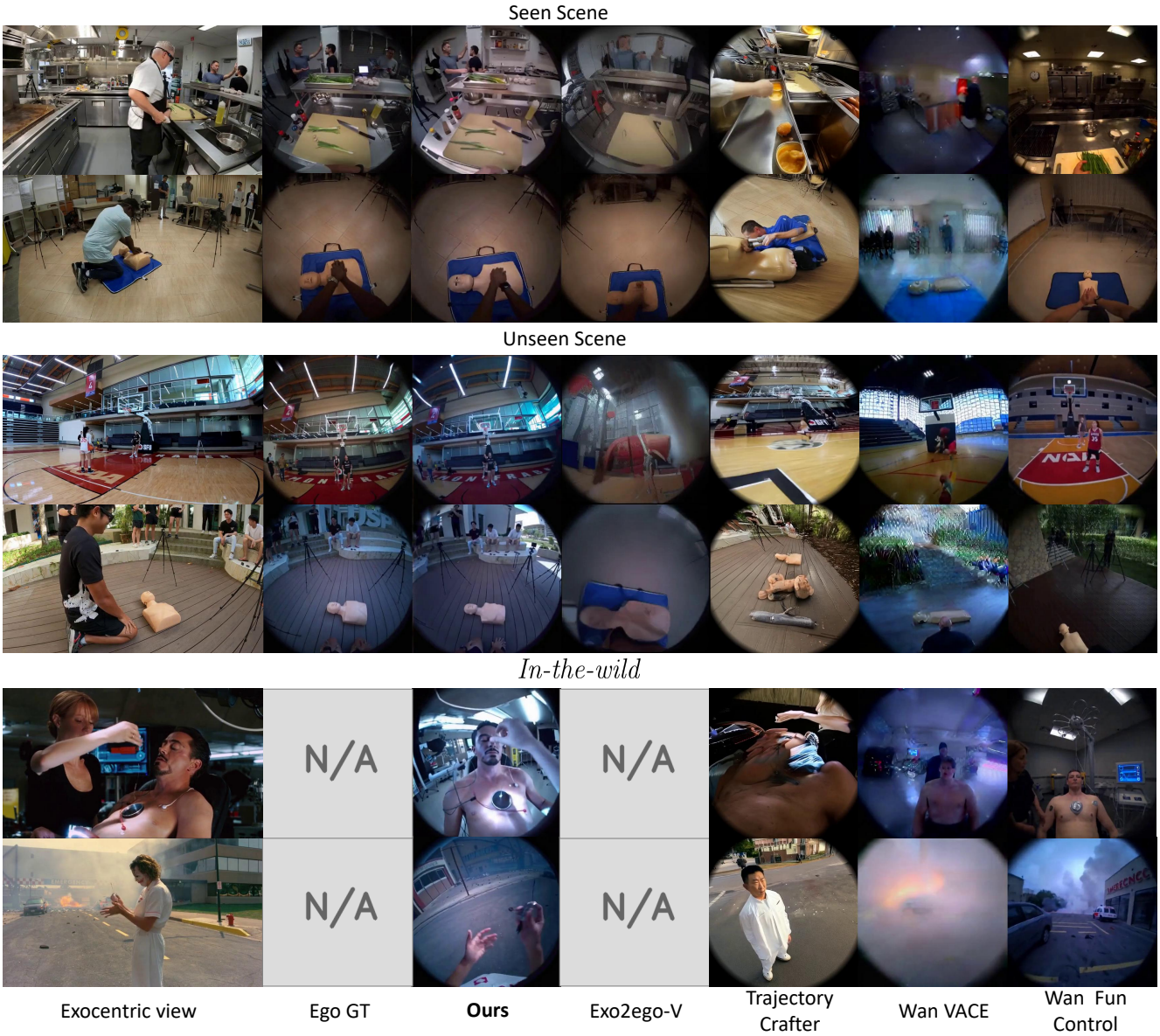

Figure 5. **Qualitative comparison.** Each example shows the exocentric input views and the corresponding generated egocentric views. While other methods fail to reconstruct realistic and coherent videos, our approach produces geometrically accurate and high-quality egocentric generations. N/A indicates that the result is unavailable either due to missing ground truth or the need for additional input views.

providing diverse points of comparison for our method. For the fair comparison, we finetuned these baselines using the same training dataset as ours.

Evaluation Metrics. To evaluate the quality of generated videos, we employed three types of criteria.

- **Image Criteria.** We measured PSNR, SSIM, LPIPS, and CLIP-I to assess how closely each generated frame matches the ground-truth distribution.
- **Object Criteria.** Following the object-level evaluation protocol of Ego-Exo4D [13], we assessed object-level consistency between the generated egocentric video and the ground truth. We used SAM2 [34] to segment and track objects and DINOv3 [36] to establish correspondences. For each matched object, we evaluated center-location error, Intersection-Over-Union(IoU), and Contour Accuracy to measure spatial alignment and boundary fidelity.
- **Video Criteria.** We measured FVD [11] to evaluate how closely the generated video aligns with the ground-truth distribution. In addition, we assessed VBench [48]-Temporal Flickering, Motion Smoothness, and Dynamic Degree to quantify temporal stability and motion quality.

Scenarios	Method	Image Criteria				Object Criteria			Video Criteria			
		PSNR ↑	SSIM ↑	LIPIS ↓	CLIP-I ↑	Location Error ↓	IoU ↑	Contour Accuracy ↑	FVD ↓	Temporal Flickering ↑	Motion Smoothness ↑	Dynamic Degree ↑
Seen Scenes	Exo2Ego-V	14.53	0.384	0.569	0.774	156.66	0.074	0.364	622.47	0.960	0.966	**0.985**
	TrajectoryCrafter	13.05	0.375	0.606	0.780	100.74	0.128	0.427	546.09	0.960	0.980	0.947
	Wan Fun Control	12.25	0.463	0.617	0.810	112.57	0.076	0.417	595.07	0.968	0.980	0.901
	Wan VACE	12.95	0.413	0.626	0.829	109.62	0.114	0.376	508.69	**0.989**	**0.994**	0.673
	EgoX (Ours)	**16.05**	**0.556**	**0.498**	**0.896**	**61.81**	**0.363**	**0.546**	**184.47**	0.977	0.990	0.974
Unseen Scenes	Exo2Ego-V	12.70	0.439	0.597	0.679	214.32	0.003	0.296	1283.50	0.971	0.976	0.978
	TrajectoryCrafter	12.24	0.297	0.619	0.778	192.16	0.039	0.301	821.71	0.966	0.984	0.944
	Wan Fun Control	13.59	0.439	0.604	0.799	191.40	0.042	0.329	968.78	0.971	0.985	0.944
	Wan VACE	12.17	0.345	0.638	0.820	191.97	0.038	0.314	1045.45	**0.995**	**0.996**	0.427
	EgoX (Ours)	**14.38**	**0.457**	**0.552**	**0.877**	**149.93**	**0.092**	**0.481**	**440.64**	0.981	0.992	**0.989**

Table 1. **Quantitative Results.** Comparison on image, object, and video metrics. Our method achieves the best overall performance, with Wan VACE showing higher video scores due to static outputs. **Best** results are highlighted in bold, and second-best results are underlined.

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
w/o Ego prior	13.67	0.479	0.573	0.864	90.70	0.417	0.464	211.50	0.974	**0.990**	0.802
w/o clean latent	15.07	0.528	0.540	0.861	70.17	0.376	0.506	343.33	0.963	0.986	0.864

Table 2. **Ablation Study Results.** Performance comparison by removing each core component of our framework. The full model achieves the best results, while excluding geometry-guided self-attention, ego prior, or clean latent conditioning causes performance degradation. **Best** results are highlighted in bold, and second-best results are underlined.

4.2. Qualitative Results

Fig. 5 visualizes the qualitative comparisons between our method and the baselines. Note that in the *in-the-wild* scenario, ground-truth egocentric videos are unavailable, and Exo2Ego-V is also not applicable since only a single exocentric video is provided, which does not meet its four-view input requirement. Exo2Ego-V fails to generate high-fidelity frames even when using four exocentric inputs, whereas our model achieves superior visual quality and generalizes well to unseen scenes from only a single exocentric view. Trajectory Crafter struggles with large camera translations, producing spatial distortions and temporal inconsistencies. Both Wan VACE and Wan Fun Control fail to effectively utilize the exocentric conditioning input, resulting in mismatched geometry, degraded realism, and the inclusion of irrelevant exocentric content in the egocentric view. Overall, these results demonstrate that our model effectively leverages pretrained video diffusion knowledge to generate geometrically accurate, visually coherent, and highly realistic egocentric videos, maintaining strong performance even under challenging in-the-wild conditions. More qualitative results, including temporally aligned visualizations, can be found in Sec. H.

4.3. Quantitative Results

As shown in Tab. 1, our method achieves the best overall performance across both image and object criteria. In particular, we observe a significant performance gap in the object-based criteria, indicating that our approach preserves scene geometry and object consistency more effectively than other baselines. While image-level scores may appear slightly lower due to the inherent challenge of synthesizing unseen regions that differ from the ground-truth egocentric view, our method still achieves the best results across all image metrics. For video-based metrics, Wan VACE records the highest temporal smoothness and flicker scores. However, this is largely attributed to its generation of overly static videos with limited motion, resulting in low dynamic realism. In contrast, our model produces temporally coherent and visually dynamic sequences, demonstrating a better balance between spatial fidelity and motion realism.

4.4. Ablation Study

We conducted ablation studies to evaluate the contribution of each core component in our framework, including the geometry-guided self-attention (GGA), the egocentric prior conditioning, and the clean latent representation. For each ablation variant, one component was removed while keeping all other settings identical. Quantitative evaluations were performed on the seen scene subset to ensure a controlled comparison. As shown in Fig. 6 and Tab. 2, removing any of these components results in a noticeable performance drop, both qualitatively and quantitatively. Without GGA, the model fails to maintain geometric alignment, attending to broad and unrelated regions, which leads to spatial inconsistency. Without the egocentric prior, the model

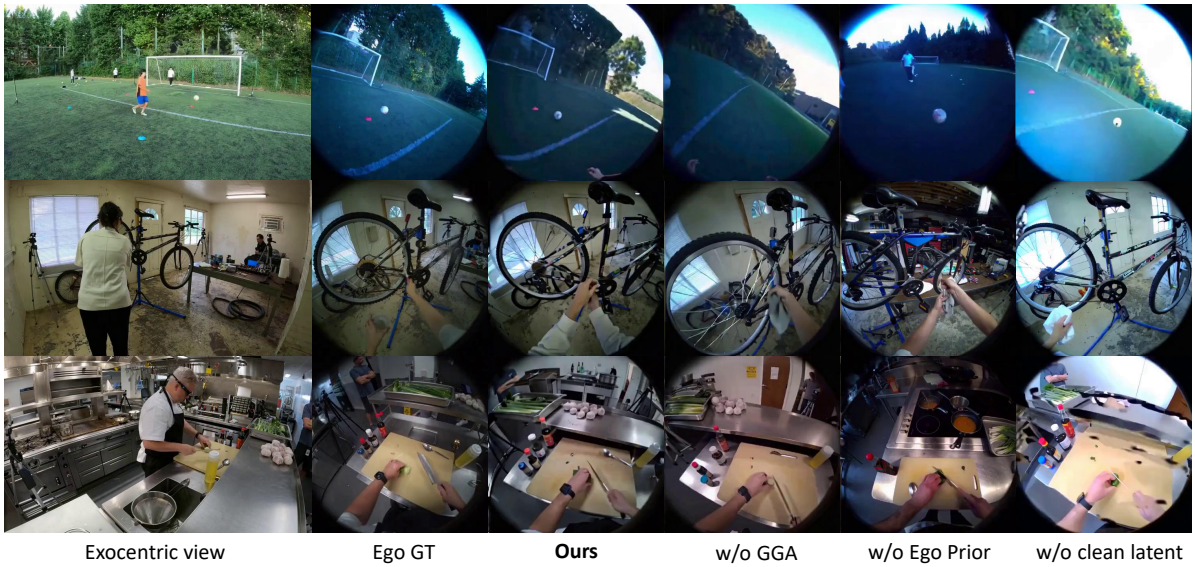

Figure 6. **Ablation qualitative comparison.** Visual results when removing each core component. Removing any single component—GGA, the egocentric prior, or the clean latent representation—results in degraded generation quality and geometric consistency.

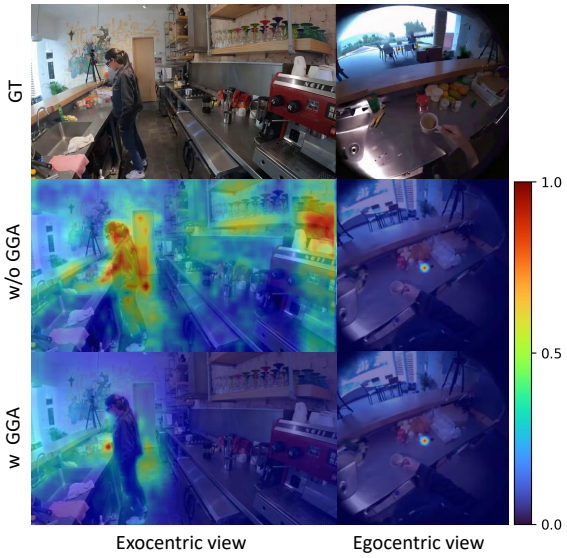

Figure 7. **Attention map visualization.** Visualization of the attention weights when querying the center token of the egocentric view. Without GGA, the model attends to unrelated regions, whereas with GGA, attention is concentrated on related regions, highlighting improved spatial alignment.

lacks explicit pixel-wise and camera trajectory information, thus struggling to follow the correct viewpoint and producing visually implausible frames. Without the clean latent, the exocentric latent is concatenated in a noisy state, which blurs fine-grained details. As a result, the target latent fails to preserve these details, leading to missing or degraded object structures. In the last row of Fig. 6, for instance, the model does not generate the spoon or the small circular ingredients on the cutting board that appear in the ground-truth egocentric view.

To further demonstrate the effectiveness of the geometry-guided self-attention, we visualize the attention maps queried by egocentric tokens. As shown in Fig. 7, without GGA, the model attends to broad irrelevant regions, while with GGA, it sharply focuses on view-relevant areas, reinforcing geometric coherence and stabilizing feature alignment. Additional ablation studies are provided in Sec. G.

5. Conclusion

We introduce **EgoX**, the first framework capable of generating egocentric videos from a single exocentric input while achieving strong generalization across diverse scenes. Our method introduces a unified conditioning strategy that combines exocentric and egocentric priors via width- and channel-wise concatenation for effective global context and viewpoint alignment, while leveraging lightweight LoRA-based adaptation to preserve the pretrained video diffusion model's spatio-temporal reasoning ability. Furthermore, clean latent representations and geometry-guided self-attention enable the model to selectively focus on spatially relevant regions and maintain geometric consistency, resulting in coherent and high-fidelity egocentric generation. Despite its effectiveness, our current framework requires an egocentric camera pose as input. Although this information can be provided interactively by users, incorporating an automatic head-pose estimation module would be a valuable future direction.

References

[1] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025. 3

[2] aigc-apps. VideoX-Fun: A flexible framework for video generation. https://github.com/aigc-apps/VideoX-Fun, 2024. Accessed: YYYY-MM-DD. 5

[3] Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B. Lindell, and Sergey Tulyakov. Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. *Proc. CVPR*, 2025. 3

[4] Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. *arXiv preprint arXiv:2503.11647*, 2025. 3

[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. 3

[6] Junsong Chen, Yuyang Zhao, Jincheng Yu, Ruihang Chu, Junyu Chen, Shuai Yang, Xianbang Wang, Yicheng Pan, Daquan Zhou, Huan Ling, et al. Sana-video: Efficient video generation with block linear diffusion transformer. *arXiv preprint arXiv:2509.24695*, 2025. 3

[7] Liyang Chen, Tianxiang Ma, Jiawei Liu, Bingchuan Li, Zhuowei Chen, Lijie Liu, Xu He, Gen Li, Qian He, and Zhiyong Wu. Humo: Human-centric video generation via collaborative multi-modal conditioning, 2025. 3

[8] Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22831–22840, 2025. 3

[9] Zhaoxi Chen, Tianqi Liu, Long Zhuo, Jiawei Ren, Zeng Tao, He Zhu, Fangzhou Hong, Liang Pan, and Ziwei Liu. 4dnex: Feed-forward 4d generative modeling made easy. *arXiv preprint arXiv:2508.13154*, 2025. 3

[10] Feng Cheng, Mi Luo, Huiyu Wang, Alex Dimakis, Lorenzo Torresani, Gedas Bertasius, and Kristen Grauman. 4diff: 3d-aware diffusion model for third-to-first viewpoint translation. In *European Conference on Computer Vision*, pages 409–427. Springer, 2024. 2, 5

[11] Songwei Ge, Aniruddha Mahapatra, Gaurav Parmar, Jun-Yan Zhu, and Jia-Bin Huang. On the content bias in fréchet video distance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 6

[12] 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. 5

[13] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024. 6

[14] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024. 3

[15] Yuhang Hu, Boyuan Chen, and Hod Lipson. Egocentric visual self-modeling for autonomous robot dynamics prediction and adaptation. *npj Robotics*, 3(1):14, 2025. 1

[16] Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, et al. Vipe: Video pose engine for 3d geometric perception. *arXiv preprint arXiv:2508.10934*, 2025. 3

[17] Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Huanzhang Dou, Chen Liang, Yutong Feng, Yu Liu, and Jingren Zhou. In-context lora for diffusion transformers. *arXiv preprint arxiv:2410.23775*, 2024. 4

[18] Tianyu Huang, Wangguandong Zheng, Tengfei Wang, Yuhao Liu, Zhenwei Wang, Junta Wu, Jie Jiang, Hui Li, Rynson WH Lau, Wangmeng Zuo, et al. Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation. *arXiv preprint arXiv:2506.04225*, 2025. 2, 3

[19] Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598*, 2025. 3, 5

[20] Zeren Jiang, Chuanxia Zheng, Iro Laina, Diane Larlus, and Andrea Vedaldi. Geo4d: Leveraging video generators for geometric 4d scene reconstruction, 2025. 3

[21] Daekyum Kim, Brian Byunghyun Kang, Kyu Bum Kim, Hyungmin Choi, Jeesoo Ha, Kyu-Jin Cho, and Sungho Jo. Eyes are faster than hands: A soft wearable robot learns user intention from the egocentric view. *Science Robotics*, 4(26): eaav2949, 2019. 1

[22] Kinam Kim, Junha Hyung, and Jaegul Choo. Temporal in-context fine-tuning for versatile control of video diffusion models. *arXiv preprint arXiv:2506.00996*, 2025. 3

[23] Xin Kong, Shikun Liu, Xiaoyang Lyu, Marwan Taher, Xiaojuan Qi, and Andrew J Davison. Eschernet: A generative model for scalable view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9503–9513, 2024. 5

[24] Ruilong Li, Brent Yi, Junchen Liu, Hang Gao, Yi Ma, and Angjoo Kanazawa. Cameras as relative positional encoding. *arXiv preprint arXiv:2507.10496*, 2025. 5

[25] Teng Li, Guangcong Zheng, Rui Jiang, Shuigen Zhan, Tao Wu, Yehao Lu, Yining Lin, and Xi Li. Realcam-i2v: Real-world image-to-video generation with interactive complex camera control. *arXiv preprint arXiv:2502.10059*, 2025. 3

[26] Jia-Wei Liu, Weijia Mao, Zhongcong Xu, Jussi Keppo, and Mike Zheng Shou. Exocentric-to-egocentric video generation. *Advances in Neural Information Processing Systems*, 37:136149–136172, 2024. 2, 3, 5

[27] Mi Luo, Zihui Xue, Alex Dimakis, and Kristen Grauman. Put myself in your shoes: Lifting the egocentric perspective from exocentric videos. In *European Conference on Computer Vision*, pages 407–425. Springer, 2024. 2

[28] Yawen Luo, Jianhong Bai, Xiaoyu Shi, Menghan Xia, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, and Tianfan Xue. Camclonemaster: Enabling reference-based camera control for video generation, 2025. 3

[29] Xiaofeng Mao, Shaoheng Lin, Zhen Li, Chuanhao Li, Wenshuo Peng, Tong He, Jiangmiao Pang, Mingmin Chi, Yu Qiao, and Kaipeng Zhang. Yume: An interactive world generation model. *arXiv preprint arXiv:2507.17744*, 2025. 3

[30] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 4

[31] Junho Park, Andrew Sangwoo Ye, and Taein Kwon. Egoworld: Translating exocentric view to egocentric view using rich exocentric observations. *arXiv preprint arXiv:2506.17896*, 2025. 2

[32] Minho Park, Taewoong Kang, Jooyeol Yun, Sungwon Hwang, and Jaegul Choo. Spherediff: Tuning-free omnidirectional panoramic image and video generation via spherical latent representation. *arXiv preprint arXiv:2504.14396*, 2025. 3

[33] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Daniel Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Pytorch3d: An open-source library for 3d deep learning. In *CVPR Workshops*, 2020. 3

[34] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 6

[35] Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6121–6132, 2025. 2, 3

[36] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. 6

[37] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 5

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[39] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 3, 5

[40] Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details. *arXiv preprint arXiv:2507.02546*, 2025. 3

[41] Zun Wang, Jaemin Cho, Jialu Li, Han Lin, Jaehong Yoon, Yue Zhang, and Mohit Bansal. EPiC: Efficient Video Camera Control Learning with Precise Anchor-Video. *arXiv preprint arXiv:2505.21876*, 2025. 3

[42] Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. Worldmem: Long-term consistent world simulation with memory. *arXiv preprint arXiv:2504.12369*, 2025. 3

[43] Jilan Xu, Yifei Huang, Baoqi Pei, Junlin Hou, Qingqiu Li, Guo Chen, Yuejie Zhang, Rui Feng, and Weidi Xie. Egoexogen: Ego-centric video prediction by watching exo-centric videos. *arXiv preprint arXiv:2504.11732*, 2025. 2, 3, 5

[44] Bowen Xue, Qixin Yan, Wenjing Wang, Hao Liu, and Chen Li. Stand-in: A lightweight and plug-and-play identity control for video generation. *arXiv preprint arXiv:2508.07901*, 2025. 3

[45] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 3

[46] Deheng Ye, Fangyun Zhou, Jiacheng Lv, Jianqi Ma, Jun Zhang, Junyan Lv, Junyou Li, Minwen Deng, Mingyu Yang, Qiang Fu, et al. Yan: Foundational interactive video generation. *arXiv preprint arXiv:2508.08601*, 2025. 3

[47] Mark YU, Wenbo Hu, Jinbo Xing, and Ying Shan. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. *arXiv preprint arXiv:2503.05638*, 2025. 2, 3, 5

[48] Fan Zhang, Shulin Tian, Ziqi Huang, Yu Qiao, and Ziwei Liu. Evaluation agent: Efficient and promptable evaluation framework for visual generative models. *arXiv preprint arXiv:2412.09645*, 2024. 6

[49] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 3