From Synchrony to Sequence: Exo-to-Ego Generation via Interpolation

Mohammad Mahdi[1⋆] Nedko Savov[1] Danda Pani Paudel[1] Luc Van Gool[1]

[1]INSAIT, Sofia University "St. Kliment Ohridski"

Abstract. Exo-to-Ego video generation aims to synthesize a first-person video from a synchronized third-person view and corresponding camera poses. While paired supervision is available, synchronized exo-ego data inherently introduces substantial spatio-temporal and geometric discontinuities, violating the smooth-motion assumptions of standard video generation benchmarks. We identify this synchronization-induced jump as the central challenge and propose Syn2Seq-Forcing, a sequential formulation that interpolates between the source and target videos to form a single continuous signal. By reframing Exo2Ego as sequential signal modeling rather than a conventional condition–output task, our approach enables diffusion-based sequence models, e.g. Diffusion Forcing Transformers (DFoT), to capture coherent transitions across frames more effectively. Empirically, we show that interpolating only the videos, without performing pose interpolation already produces significant improvements, emphasizing that the dominant difficulty arises from spatio-temporal discontinuities. Beyond immediate performance gains, this formulation establishes a general and flexible framework capable of unifying both Exo2Ego and Ego2Exo generation within a single continuous sequence model, providing a principled foundation for future research in cross-view video synthesis. The code will be released at `https://github.com/insait-institute/Syn2Seq`.

1 Introduction

Given temporally synchronized exocentric (third-person) and egocentric (first-person) videos of the same scene, denoted as the source video x and the target video g, along with their corresponding per-frame camera poses $p = [p_x, p_g]$, the Exo2Ego generation task aims to learn a parametric mapping f_θ that synthesizes the target egocentric observation conditioned on the available source view and camera geometry:

$$\min_\theta \ \|f_\theta(x, p) - g\|.$$

Exo2Ego generation is practically significant in applications such as robotics, augmented reality, and virtual reality [12, 18, 24, 25, 31, 48], where systems must

⋆ `firstname.lastname@insait.ai`

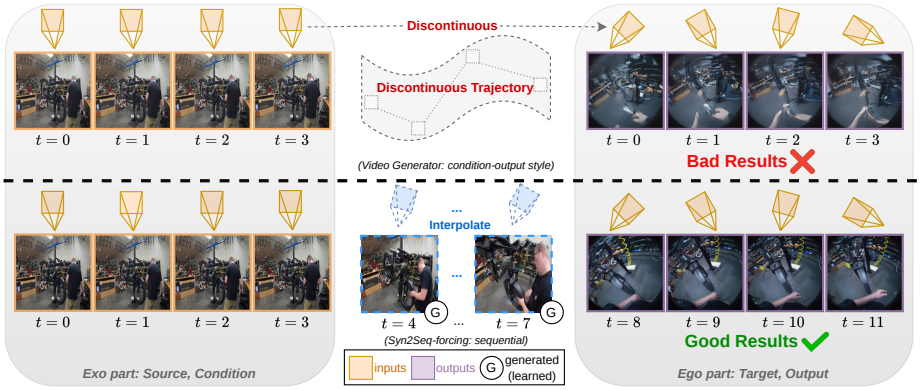

Fig. 1: Top: Standard video generators struggle with discontinuous camera poses and missing transitions from exo to ego. Bottom: Syn2Seq-Forcing (ours) interpolates between views and camera poses to form a single continuous sequence, enabling smooth exo2ego generation; the same framework naturally extends to ego2exo.

reconstruct or predict a first-person experience from third-person observations to enable accurate perception, interaction, and immersive feedback. The problem is inherently challenging: it not only requires transferring viewpoint information but also necessitates capturing fine-grained spatial and temporal cues, including body dynamics, hand-object interactions, and scene elements that may be only partially observable from the exocentric perspective. Moreover, maintaining temporal coherence and visual fidelity across frames further increases the complexity of the task.

While temporal synchronization provides paired supervision, it also introduces two fundamental challenges that make Exo2Ego generation qualitatively different from standard video generation. First, from a frame-coherence perspective, the viewpoint gap between x and g is often substantial, resulting in a large spatio-temporal discontinuity between the end of the source video and the beginning of the target video. Second, from a geometric perspective, the corresponding camera poses can exhibit significant discontinuities: the configuration of p_x at the end of the source sequence may differ sharply from p_g at the start of the target sequence.

In contrast, most conventional video generation benchmarks [2, 51] assume smooth motion and continuous trajectories. Consequently, models trained under these assumptions, e.g., [2], often struggle to handle the abrupt spatial and geometric transitions inherent in synchronized exo-ego video pairs. These challenges highlight the need for methods that explicitly account for both *viewpoint jumps* and *pose discontinuities* to produce temporally coherent and visually accurate egocentric outputs.

Prior work has largely followed a condition–output paradigm, exploring different strategies to condition on x and p [26,30] within various architectures [31], or leveraging multiple exocentric views to enrich spatial cues [24]. In this paper,

however, we take a step back to address the fundamental challenge introduced by synchronization. Our key idea is to replace the traditional *synchronized but discontinuous* formulation with a *sequential* one: we interpolate between x and g to generate a single continuous signal that effectively bridges the spatio-temporal and geometric gaps, as illustrated in Fig. 1.

This transformation, moving from condition-output modeling to sequential signal modeling, not only produces measurable improvements in performance but also opens up new avenues for flexible and controllable video generation. By treating the source, interpolated, and target sequences as a unified signal, where past frames condition future frames, our approach enables the network to capture smooth transitions, maintain temporal coherence, and better handle abrupt viewpoint and pose changes inherent in exo–ego video pairs.

Exo2Ego generation requires strong depth understanding, which prior work typically addresses using explicit 3D priors. In contrast, we adopt interpolation as a simple proxy for depth reasoning. Rather than explicitly estimating geometry, we guide the model toward the target viewpoint through intermediate frames—essentially "*if the depth is unknown, reach the target via interpolation.*"

We demonstrate that even interpolating only the video frames without performing interpolation on the camera poses, already leads to substantial gains, emphasizing that the primary challenge in Exo2Ego generation stems from the spatio-temporal discontinuities introduced by synchronization. Beyond these performance improvements, the sequential formulation offers significant conceptual flexibility: in principle, it could be extended to generate longer sequences, such as transitioning from egocentric back to exocentric views, thereby enabling both Exo2Ego and Ego2Exo generation within a single unified framework. Although we only briefly explore this extended flexibility in the present work, the results highlight the broader potential of our approach as a general framework for exo/ego-centric video generation.

Contributions.

1. We study and identify synchronization-induced discontinuities as the core obstacle in Exo2Ego video generation.
2. We resolve this obstacle by reframing the task as sequential modeling with interpolation, yielding significant performance gains even without pose interpolation.
3. We introduce a general-purpose paradigm that can unify both Exo2Ego and Ego2Exo generation within a single continuous sequence model.

2 Related Works

Diffusion-based Video Generation.

Recent advances in diffusion models and large-scale datasets have rapidly improved the quality and diversity of video generation [4,13,14,16,27,28,33,37,39,50,52,53]. Representative systems such as Tune-A-Video [46], Video Diffusion

Model	gt Cnd.	Text Cnd.	Exo Views	HOI Cnd.	3D Prior	Transitional
Trj-Crafter [51]	No	Yes	1	No	Yes	No
PMYS [30]	No	No	1	Yes	No	No
Exo2Ego-V [25]	No	No	4	No	Yes	No
EgoExo-Gen [48]	Yes	Yes	1	Yes	No	No
Exo2EgoSyn [31]	No	No	4	No	No	No
Ours	No	No	1	No	No	Yes

Table 1: Comparison of structural assumptions in different Exo2Ego methods. *gt Cnd.* indicates use of any signal, e.g., the first frame, from ego ground-truth signals, *Text Cnd.* indicates text-based conditioning, *HOI Cnd.* denotes use of hand–object interaction information, and *Transitional* shows if the model can generate exo-to-ego transitions.

Models [14], Stable Video Diffusion [3], AnimateDiff [9], LTX 2 [10], Hunyan-Video [20] and WAN2.2 [44] demonstrate the ability of diffusion models to generate temporally coherent and visually realistic videos. Recently, long-horizon video generation has been advanced by autoregressive and diffusion-based approaches, including Diffusion Forcing Transformers [41], SelfForcing [17], and SelfForcing++ [5], which enable stable generation over extended sequences.

To offer fine-grained control beyond an initial text prompt, recent work has introduced mechanisms for explicit camera control in video generation. Methods such as CameraCtrl [11], MotionCtrl [45], Direct-a-Video [49], and CamTrol [15] enable controllable camera trajectories during generation. ReCamMaster [2] achieve fine-grained camera control by directly manipulating temporal attention within large video generators. World models enable interactive environments with real-time camera control: WorldPlay [42], Genie 3 [36], and LingBot [43]. Long-horizon video generators, such as Diffusion Forcing Transformers [41] and MotionStream [40] have been demonstrated to follow camera controls as well. SPMem [47], VMem [23], Gen3C [38], and TrajectoryCrafter [51] maintain 3D representations during video generation, to produce consistent content throughout viewpoints.

These camera-controlled video generation methods synthesize videos by following a continuous camera motion trajectory during generation. However, they do not address the task of transforming the viewpoint of an existing video, leaving cross-view camera pose transfer largely unexplored.

Ego-Exo Cross-View Learning. Cross-view learning studies how models can relate, interpret, and synthesize observations of the same scene captured from different viewpoints, such as egocentric and exocentric perspectives. Exocentric understanding has been advanced by the emergence of ego–exo dual-view datasets - Ego-ExoLearn [18], Ego-Exo4D [8]. Existing work has largely focused on perception and reasoning tasks, including ego–exo object correspondence [6,7,32,34] and ego–exo visual question answering [12, 22]. In this work, we instead address ego–exo video generation [26], which poses a substantially more challenging problem that requires synthesizing realistic egocentric observations rather than only understanding them.

Despite its importance, ego–exo video generation has received relatively limited attention. Luo et al. [30] propose a two-stage pipeline (PMYS) that first predicts hand trajectories and subsequently generates egocentric videos conditioned on exocentric inputs. EgoWorld [35] reconstructs egocentric views using exocentric depth, 3D hand pose estimates, and textual descriptions through point-cloud reprojection followed by diffusion-based inpainting. EgoExo-Gen [48] conditions egocentric video synthesis on action descriptions together with the first egocentric frame. Complementary work studies the inverse direction of generation: intention-driven approaches focus on synthesizing exocentric videos from egocentric inputs [29]. In our approach for exo-to-ego video synthesis, we only rely on camera poses, and do not require extra labels or building explicit representations, as demonstrated on Tab. 1. Exo2Ego-V [26] instead leverages multi-view exocentric observations and camera poses within a PixelNeRF-based diffusion framework to generate egocentric videos. In contrast, our method only requires a single exo view and no 3D prior, as shown on Tab. 1.

Finally, Exo2EgoSyn [31] demonstrates how large pretrained video generators can be repurposed for the Exo2Ego task. While benefiting from strong pretrained representations, their approach relies on a single predicted egocentric frame to guide generation and applies per-frame camera control despite the model's temporally coupled attention. In contrast, we train a model better suited for per-frame pose conditioning and make use of full exocentric video inputs, while still leveraging large-scale pretrained knowledge through pose interpolation during data collection.

3 Methodology

3.1 Preliminaries

Diffusion Forcing Transformer. Diffusion Forcing Transformer (DFoT) is a video diffusion framework designed to support flexible history conditioning, *i.e.*, the ability to condition generation on an arbitrary subset or length of previous frames. Unlike standard video diffusion pipelines that assume fixed conditioning layouts, DFoT applies the diffusion process with independent noise levels for each frame. This generalizes the *noising-as-masking* principle: heavily noised frames effectively act as masked (removed) conditioning inputs, whereas lightly noised frames retain usable historical information, allowing the model to dynamically leverage past context.

This design enables a classifier-free guidance (CFG)-style formulation over the history during sampling. Specifically, the *unconditional* branch corresponds to fully noising (masking) the historical frames, while the *conditional* branch preserves selected history frames. Combining these score estimates provides a controllable trade-off between fidelity, temporal consistency, and diversity, now extended to handle variable-length and variable-structure histories. DFoT can be implemented using transformer-based video backbones (e.g., DiT or U-ViT) and is fully compatible with fine-tuning from pretrained video diffusion models,

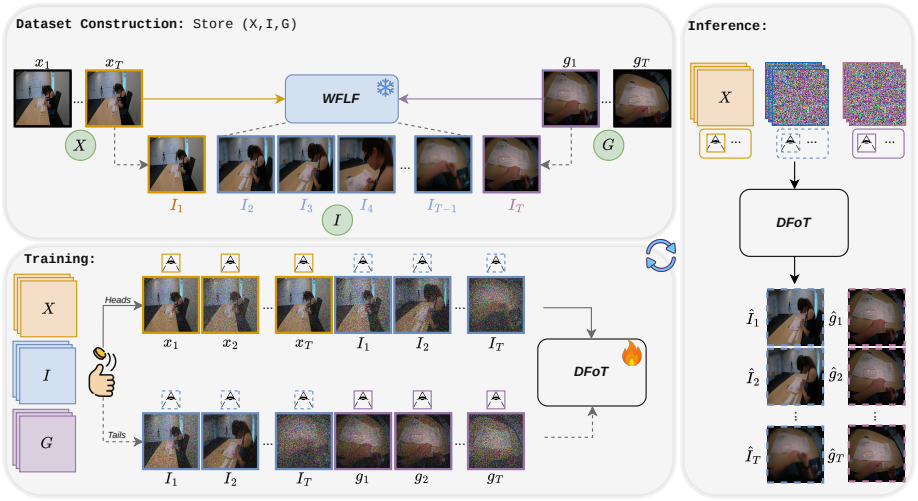

Fig. 2: Syn2Seq-Forcing framework. Top-left: Using the last exo frame and first ego frame, we generate and cache pseudo-ground-truth interpolations with the WFLF model. Bottom-left: During training, a random pair of either (Exo–Interpolated) or (Interpolated–Ego) transitions is selected to train a DFoT conditioned on the corresponding camera poses. Right: During inference, the model receives the exo video and all camera poses to generate both the exo-to-ego transition and the final ego video.

making it a flexible and practical choice for complex video generation tasks that require adaptive temporal conditioning.

Built on top of DFoT, History Guidance (HG) introduces structured mechanisms to exploit flexible history conditioning: *Vanilla History Guidance (HG-v):* applies CFG with an arbitrary history window. *Temporal History Guidance (HG-t):* combines score estimates from multiple history windows across time. *Fractional History Guidance (HG-f):* uses partially noised history to modulate information by frequency content (e.g., low-pass or band-pass effects). *HG-tf:* a joint time-frequency guidance strategy that combines HG-t and HG-f.

In Exo2Ego generation, we identify the primary challenge as the severe discontinuity between synchronized exocentric and egocentric sequences, both in visual content and camera poses. This abrupt spatio-temporal and geometric jump significantly degrades generation quality, as it violates the smooth-transition assumptions underlying standard video diffusion models. By explicitly bridging this gap through interpolation, we transform the synchronized pair into a single continuous signal. Such a formulation naturally aligns with DFoT, which is designed to flexibly process arbitrary history subsets within a unified sequence.

WAN2.2 First/Last Frame (WFLF). WAN2.2 is a powerful foundational video Latent Diffusion Model(LDM) designed for high-quality video generation conditioned on text and/or images. Wan2.2-Lightning is a distilled Wan2.2 variant designed for efficient sampling with very few diffusion steps and without

requiring CFG. WFLF is the boundary-conditioned baseline built on Wan2.2-Lightning 8-step LoRA, while following the FLF2V setup (conditioning on the first and last frames) to preserve global structure and temporal consistency. This design makes WFLF a compact and efficient frame interpolator, capable of generating temporally coherent transitions with minimal diffusion steps.

3.2 Task Definition

Given an exocentric video sequence $X = \{x_1, \ldots, x_T\} \in \mathbb{R}^{T\times C\times H\times W}$, with T frames, per-frame exocentric camera poses $P_x = \{p_1^x, \ldots, p_T^x\} \in \mathbb{R}^{T\times v}$, and per-frame egocentric camera poses $P_g = \{p_1^g, \ldots, p_T^g\} \in \mathbb{R}^{T\times v}$, our goal is to synthesize a temporally synchronized egocentric video $G = \{g_1, \ldots, g_T\} \in \mathbb{R}^{T\times C\times H\times W}$, while preserving temporal alignment and scene-level consistency across viewpoints.

This cross-view generation problem is intrinsically challenging due to the large appearance and geometric gap between exocentric and egocentric observations, especially at the exo-to-ego transition (e.g., from x_T to g_1). The difficulty is further amplified by substantial viewpoint change and camera-motion mismatch between the two domains. In our setting, this becomes particularly severe because the exocentric camera is often static, i.e., $p_1^x = \cdots = p_T^x$, which provides limited multi-view geometric cues and weak 3D observability. Consequently, the model must infer missing scene structure and plausible ego-view dynamics from highly under-constrained exocentric evidence.

3.3 Our Framework

Instead of learning a direct cross-view mapping

$$G = f_\theta(X, P_x, P_g)\,,$$

we cast the task as *diffusion-based sequential modeling* and propose *Syn2Seq-Forcing*, as shown in Fig. 2. We build a unified temporal stream S by concatenating exocentric frames, an intermediate interpolated segment, and egocentric frames. The interpolated segment is introduced to bridge the large exo-to-ego transition gap and enforce a smoother appearance and motion trajectory. In parallel, we interpolate camera poses to form a unified pose sequence

$$P = (P_x, P_i, P_g),$$

where P_i denotes interpolated transition poses. Following DFoT, we apply per-frame independent noise levels and feed the model the corrupted sequence S^n, while conditioning on P:

$$\hat{\epsilon} = \mathrm{S2S}_\phi(S^n, P)\,.$$

This design offers three practical benefits. First, ego-view generation is conditioned not only on exo observations but also on transition-aware interpolated

visual and pose cues, which is difficult to obtain in a direct one-shot formulation. Second, during sampling, the model jointly synthesizes both the transition segment and the target ego sequence in a temporally coherent manner. Third, because the formulation is sequence-centric rather than direction-specific, it naturally extends to the reverse setting (Ego→Exo) within the same framework.

Video Interpolation. To construct the unified stream $S = [X, I, G]$, we synthesize an intermediate segment I that bridges the last exocentric frame x_T and the first egocentric frame g_1. We use the frozen WFLF model for this purpose. Managing to require $|I| = T$ (matching the lengths of X and G), we query WFLF to generate only the interior transition frames:

$$I' = \mathrm{WFLF}(x_T, g_1, \mathrm{prompt}), \qquad |I'| = T - 2,$$

and then explicitly attach the boundary frames:

$$I = [x_T,\ I',\ g_1], \qquad |I| = T.$$

Here, *prompt* is a dataset-specific text prompt used to steer generation.

The resulting I is treated as a pseudo-interpolation signal: WFLF is not updated during training, and its outputs serve as supervisory transition targets within our S2S training pipeline. The details for querying WFLF is provided in supplementary materials.

Pose Interpolation. Given $P_x = \{p_t^x\}_{t=1}^T$ and $P_g = \{p_t^g\}_{t=1}^T$, with $p_t \in \mathbb{R}^v$, we define an intermediate pose trajectory

$$P_i = \{p_j^i\}_{j=1}^T,$$

that connects the boundary poses p_T^x and p_1^g. Each pose is decomposed as

$$p = (k,\, [R \,|\, \mathbf{t}]),$$

where k denotes camera intrinsics, $R \in SO(3)$ is rotation, and $\mathbf{t} \in \mathbb{R}^3$ is translation. For normalized time $\tau_j = \frac{j-1}{T-1}$, we interpolate

$$R_j^i = \mathrm{Slerp}(R_T^x, R_1^g; \tau_j)\,, \qquad \mathbf{t}_j^i = (1-\tau_j)\mathbf{t}_T^x + \tau_j \mathbf{t}_1^g,$$

and form

$$p_j^i = \big(k_j^i,\, [R_j^i \,|\, \mathbf{t}_j^i]\big), \quad j = 1, \ldots, T.$$

In practice, k_j^i is fixed to the exocentric intrinsics for all j. We also enforce boundary consistency:

$$p_1^i = p_T^x, \qquad p_T^i = p_1^g.$$

The final conditioning pose sequence is

$$P = (P_x,\, P_i,\, P_g),$$

which provides a smooth camera-geometry transition between the exocentric and egocentric segments. Details of $\mathrm{Slerp}(\cdot)$ are provided in the supplementary material.

Training. We adopt a two-stage pretraining–finetuning paradigm, optimizing the following loss function for both stages:

$$\mathcal{L} = \|\epsilon - \mathrm{S2S}_\phi(S^n, P)\|^2 .$$

Pretraining. In the pretraining stage, the model is trained on 356k video clips collected from all available categories. Each sample consists of a direct Exo→Ego pair, without any interpolation. The pretraining runs for 20 epochs (approximately 4 days) on 8 NVIDIA H200 GPUs, providing a strong initialization for subsequent fine-tuning.

Finetuning. Building upon the pretrained weights, we finetune category-specific models using 40k videos per category, enabling interpolation, for 150 epochs (roughly 4 days) on 8 NVIDIA H200 GPUs. Training samples are constructed as paired triplets $(S = (X, I, G), P = (P_x, P_i, P_g)$ with I and P_i representing interpolated pseudo ground-truth visual and pose sequences. At each training step, we randomly select one of two sequential sub-tasks: either

$$D = \big((X, I), (P_x, P_i)\big), \quad \text{or} \quad D = \big((I, G), (P_i, P_g)\big).$$

This stochastic decomposition allows the model to jointly learn both transitions, Exo→Interp and Interp→Ego, within a unified training framework. More details regarding implementation, hyperparameters, and the setup for enabling Ego2Exo generation are provided in the supplementary material.

Inference. During inference, we condition the model on clean exocentric frames and append a noise-only suffix of length $2T$. The first T noise slots are denoised into the transition segment, and the remaining T noise slots are denoised into the egocentric segment. Formally, the generated sequence is

$$\hat{S} = \big[X,\ \hat{I},\ \hat{G}\big], \qquad |\hat{I}| = |\hat{G}| = T,$$

where $\hat{I}$ and $\hat{G}$ are produced in a single denoising process conditioned on X and the corresponding pose stream.

4 Experiments

4.1 Experimental Setup

Dataset. We evaluate our approach in three categories of the Ego-Exo4D [8] benchmark: Bike, Health, and Cooking. For each egocentric video, the dataset provides four exocentric videos (this study uses only one exo video randomly taken from the available ones) along with the corresponding camera poses for the fixed exocentric cameras and for every frame of the egocentric video. Our experiments process 363 videos from the Bike category, 678 videos from the Cooking category, and 397 videos from the Health category. Both egocentric and exocentric videos are spatially resized to a resolution of $H = W = 256$.

Method	Health			Bike			Cooking		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
Trj-Crafter [51]	14.3091	0.4176	0.5623	14.0102	0.3409	0.6001	13.7554	0.3729	0.6003
Wan-FCtrl [1]	13.8993	0.4319	0.5734	13.5699	0.3312	0.6197	13.4200	0.3599	0.6071
Wan VACE [19]	14.1922	0.4299	0.5966	13.2104	0.3329	0.6156	13.3680	0.3689	0.6129
Exo2EgoSyn [31]	15.6234	0.4821	0.4991	15.1319	0.3895	0.5052	13.9011	0.4039	0.6125
Ours	16.7139	0.5728	0.4828	15.6301	0.4721	0.5012	14.3897	0.4531	0.5854

Table 2: Main comparisons: Quantitative evaluation of Exo2Ego generation methods across three categories.

Since the WFLF architecture requires the number of frames to satisfy $4n+1$, we sample 9 frames per video. We use a skip-frame step of 4 during preprocessing, ensuring that sufficient motion is captured in the sampled frames. For testing splits, we follow the partitions released by [8].

Baseline Construction. To benchmark against prior Exo-to-Ego methods and reflect recent progress in controllable video generation, we include comparisons with Trajectory Crafter [51], Wan Fun Control [1], and Wan VACE [19], which utilize different conditioning strategies. Exo2Ego-V [26] is limited to scenarios with four exocentric views, and EgoExo-Gen (X-Gen) [48] lacks a publicly available implementation; therefore, both methods are omitted from our experiments.

Evaluation Metrics. To quantitatively assess our method, we report performance using PSNR, SSIM, and LPIPS (AlexNet [21]), following the evaluation protocol described in [26]. This ensures a consistent and fair comparison with baseline methods while providing complementary measures of both pixel-level fidelity and perceptual similarity for the generated egocentric videos.

4.2 Comparison with Baselines

We present a comparison with the proposed baselines in Tab. 2. Across all categories, Syn2Seq-Forcing(ours) consistently achieves superior results, highlighting the robustness of our approach. To ensure a fair comparison with methods that cannot produce smooth exocentric-to-egocentric transitions, only the Ego-generated portion of Syn2Seq-Forcing is taken in to account for metric computation, excluding the interpolated segments. Qualitative results are shown in Fig. 3, and additional visual examples are available in the supplementary material.

4.3 Ablation Studies

Interpolation. We demonstrate that interpolating only the video frames already provides a substantial performance boost. In this setup, no pose interpolation is performed; we use the egocentric poses directly and fill the remaining camera poses with zero vectors. We further evaluate the effect of including pose interpolation, with results summarized in Tab. 3. Qualitative examples illustrating the impact of both frame-only and frame-plus-pose interpolation are presented in Fig. 4.

Fig. 3: Comparison of different methods against ground-truth. Bettter viewed zoomed.

Model	Health			Bike			Cooking		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
Exo2Ego-Direct	13.5419	0.4176	0.5822	14.0091	0.3552	0.5866	13.1609	0.3831	0.6176
Exo2Ego-FI	16.1003	0.5427	0.4902	15.1245	0.4405	0.5387	14.2131	0.4401	0.5899
Exo2Ego-FPI	16.7139	0.5728	0.4828	15.6301	0.4721	0.5012	14.3897	0.4531	0.5854

Table 3: Interpolation effect: *Direct* predicts the ego view directly from the exo view without interpolation. *FI* applies only frame interpolation, and uses ego camera poses with setting other poses to zero. *FPI* applies both frame and pose interpolation.

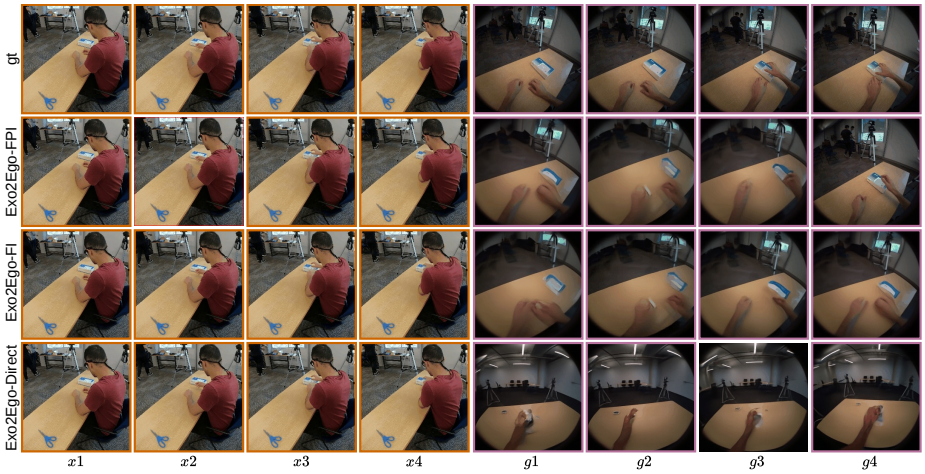

Fig. 4: Comparison illustrating the effect of interpolation on the generated outputs.

Type of Video Interpolation. We, as reported in Tab. 4, investigate the effectiveness of our interpolator, WFLF, in comparison to DFoT's native inference-time interpolation. As mentioned earlier, typical video generator models alone cannot adequately handle the large viewpoint gap between exocentric and egocentric frames, resulting in inferior performance. We report quantitative results separately for the first T generated frames (interpolated frames), the next T frames (egocentric frames), and for both segments combined, providing a comprehensive evaluation of the interpolator's contribution.

Setup	Bike-INT			Bike-EGO			Bike-Both		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
WFLF	15.6933	0.4754	0.5022	15.6301	0.4721	0.5012	15.6522	0.4736	0.5016
DFoT-INTRPL	13.1109	0.3203	0.6327	13.9012	0.3335	0.6192	13.4801	0.3276	0.6258

Table 4: Video interpolator: DFoT-INTRPL denotes DFoT's native interpolation capability during inference.

Type of Pose Embedding We perform an ablation study on the type of pose embedding, comparing three variants, according to Tab. 5: the *Global* version, which represents camera poses as 16-dimensional vectors per frame; *Ray Encoding*, which encodes per-pixel rays in a 180-dimensional space; and *Plücker* embeddings, a per-pixel 6-dimensional representation of camera poses. Across experiments, we consistently observe, that Plücker embeddings yield the best performance, demonstrating superior compatibility with our interpolation framework.

Setup	**Bike-INT**			**Bike-EGO**		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR	SSIM	LPIPS
Global	15.0799	0.4588	0.5116	15.0197	0.4571	0.5184
Ray Encoding	15.6028	0.4739	0.5097	15.5702	0.4729	0.5051
Pluckers(ours)	15.6933	0.4754	0.5022	15.6301	0.4721	0.5012

Table 5: Effect of different methods for embedding camera poses.

5 Conclusion

In this work, we revisited Exo2Ego video generation task from a synchronization-aware perspective and demonstrated that the primary bottleneck is not simply the quality of conditioning, but the abrupt spatio-temporal and geometric discontinuities between synchronized exocentric and egocentric segments. To address this challenge, we replaced the traditional condition–output formulation with a sequential signal approach that explicitly bridges these gaps through interpolation and models the resulting sequence as a continuous signal under diffusion-based forcing. Exo2Ego generation typically depends on explicit depth reasoning, often implemented through 3D priors in prior work. In this work, we instead use interpolation as a lightweight proxy, guiding the model toward the target view via intermediate frames rather than explicit geometric estimation. This reformulation yields clear empirical improvements and remains effective even when only video interpolation is applied, highlighting that mitigating temporal discontinuity is a key factor driving performance.

Beyond these results, our framework provides a unified perspective on cross-view video generation: by treating the source, transitional, and target segments as a single structured sequence, it inherently supports flexible conditioning and points toward bidirectional view translation (Exo2Ego and Ego2Exo) within a single model. We anticipate that this approach will inspire future research on longer-horizon cross-view video synthesis and more comprehensive multi-view sequential modeling.

6 Acknowledgements

This research was partially funded by the Ministry of Education and Science of Bulgaria (support for INSAIT, part of the Bulgarian National Roadmap for Research Infrastructure).

References

1. AIGC-Apps: Videox-fun: A flexible framework for video generation. `https://github.com/aigc-apps/VideoX-Fun` (2024), accessed: 2026-03-05
2. Bai, J., Xia, M., Fu, X., Wang, X., Mu, L., Cao, J., Liu, Z., Hu, H., Bai, X., Wan, P., et al.: Recammaster: Camera-controlled generative rendering from a single video. arXiv preprint arXiv:2503.11647 (2025)
3. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
4. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22563–22575 (2023)
5. Cui, J., Wu, J., Li, M., Yang, T., Li, X., Wang, R., Bai, A., Ban, Y., Hsieh, C.J.: Self-forcing++: Towards minute-scale high-quality video generation. arXiv preprint arXiv:2510.02283 (2025)
6. Fu, Y., Wang, R., Fu, Y., Paudel, D.P., Van Gool, L.: Cross-view multi-modal segmentation @ ego-exo4d challenges 2025 (2025), ego-Exo4D Challenge
7. Fu, Y., Wang, R., Ren, B., Sun, G., Gong, B., Fu, Y., Paudel, D.P., Huang, X., Van Gool, L.: Objectrelator: Enabling cross-view object relation understanding across ego-centric and exo-centric perspectives. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6530–6540 (2025)
8. Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., Ashutosh, K., Baiyya, V., Bansal, S., Boote, B., 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. pp. 19383–19400 (2024)
9. Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv:2307.04725 (2023)
10. HaCohen, Y., Brazowski, B., Chiprut, N., Bitterman, Y., Kvochko, A., Berkowitz, A., Shalem, D., Lifschitz, D., Moshe, D., Porat, E., et al.: Ltx-2: Efficient joint audio-visual foundation model. arXiv preprint arXiv:2601.03233 (2026)
11. He, H., Xu, Y., Guo, Y., Wetzstein, G., Dai, B., Li, H., Yang, C.: Cameractrl: Enabling camera control for text-to-video generation. arXiv preprint arXiv:2404.02101 (2024)
12. He, Y., Huang, Y., Chen, G., Pei, B., Xu, J., Lu, T., Pang, J.: Egoexobench: A benchmark for first-and third-person view video understanding in mllms. arXiv preprint arXiv:2507.18342 (2025)
13. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022)

14. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. Advances in neural information processing systems **35**, 8633–8646 (2022)
15. Hou, C., Chen, Z.: Training-free camera control for video generation. arXiv preprint arXiv:2406.10126 (2024)
16. Huang, S., Gong, B., Feng, Y., Chen, X., Fu, Y., Liu, Y., Wang, D.: Learning disentangled identifiers for action-customized text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7797–7806 (2024)
17. Huang, X., Li, Z., He, G., Zhou, M., Shechtman, E.: Self forcing: Bridging the train-test gap in autoregressive video diffusion. arXiv preprint arXiv:2506.08009 (2025)
18. Huang, Y., Chen, G., Xu, J., Zhang, M., Yang, L., Pei, B., Zhang, H., Dong, L., Wang, Y., Wang, L., et al.: Egoexolearn: A dataset for bridging asynchronous ego- and exo-centric view of procedural activities in real world. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22072–22086 (2024)
19. Jiang, Z., Han, Z., Mao, C., Zhang, J., Pan, Y., Liu, Y.: Vace: All-in-one video creation and editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17191–17202 (2025)
20. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
21. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25** (2012)
22. Lee, I., Park, W., Jang, J., Noh, M., Shim, K., Shim, B.: Towards comprehensive scene understanding: Integrating first and third-person views for lvlms. arXiv preprint arXiv:2505.21955 (2025)
23. Li, R., Torr, P., Vedaldi, A., Jakab, T.: Vmem: Consistent interactive video scene generation with surfel-indexed view memory. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 25690–25699 (2025)
24. Li, Y.M., Huang, W.J., Wang, A.L., Zeng, L.A., Meng, J.K., Zheng, W.S.: Egoexo-fitness: Towards egocentric and exocentric full-body action understanding. In: European Conference on Computer Vision. pp. 363–382. Springer (2024)
25. Liu, G., Tang, H., Latapie, H., Yan, Y.: Exocentric to egocentric image generation via parallel generative adversarial network. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1843–1847. IEEE (2020)
26. Liu, J.W., Mao, W., Xu, Z., Keppo, J., Shou, M.Z.: Exocentric-to-egocentric video generation. Advances in Neural Information Processing Systems **37**, 136149–136172 (2024)
27. Liu, Y., Pan, J., Yang, J., Chen, T., Zhou, P., Zhang, B.: Diverse instance generation via diffusion models for enhanced few-shot object detection in remote sensing images. IEEE Geoscience and Remote Sensing Letters (2025)
28. Liu, Y., Pan, J., Zhang, B.: Control copy-paste: Controllable diffusion-based augmentation method for remote sensing few-shot object detection. arXiv preprint arXiv:2507.21816 (2025)
29. Luo, H., Zhu, K., Zhai, W., Cao, Y.: Intention-driven ego-to-exo video generation. arXiv preprint arXiv:2403.09194 (2024)

30. Luo, M., Xue, Z., Dimakis, A., Grauman, K.: Put myself in your shoes: Lifting the egocentric perspective from exocentric videos. In: European Conference on Computer Vision. pp. 407–425. Springer (2024)
31. Mahdi, M., Fu, Y., Savov, N., Pan, J., Paudel, D.P., Van Gool, L.: Exo2egosyn: Unlocking foundation video generation models for exocentric-to-egocentric video synthesis. arXiv preprint arXiv:2511.20186 (2025)
32. Mur-Labadia, L., Santos-Villafranca, M., Bermudez-Cameo, J., Perez-Yus, A., Martinez-Cantin, R., Guerrero, J.J.: O-mama: Learning object mask matching between egocentric and exocentric views. In: ICCV (2025)
33. Pan, J., Lei, S., Fu, Y., Li, J., Liu, Y., Sun, Y., He, X., Peng, L., Huang, X., Zhao, B.: Earthsynth: Generating informative earth observation with diffusion models. arXiv preprint arXiv:2505.12108 (2025)
34. Pan, J., Wang, R., Qian, T., Mahdi, M., Fu, Y., Xue, X., Huang, X., Van Gool, L., Paudel, D.P., Fu, Y.: V^2-sam: Marrying sam2 with multi-prompt experts for cross-view object correspondence. arXiv preprint arXiv:2506.05856 (2025)
35. Park, J., Ye, A.S., Kwon, T.: Egoworld: Translating exocentric view to egocentric view using rich exocentric observations. arXiv preprint arXiv:2506.17896 (2025)
36. Parker-Holder, J., Fruchter, S.: Genie 3: A new frontier for world models. URL https://deepmind. google/discover/blog/genie-3-a-new-frontier-for-world-models/. Blog post (2025)
37. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
38. Ren, X., Shen, T., Huang, J., Ling, H., Lu, Y., Nimier-David, M., Müller, T., Keller, A., Fidler, S., Gao, J.: Gen3c: 3d-informed world-consistent video generation with precise camera control. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6121–6132 (2025)
39. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
40. Shin, J., Li, Z., Zhang, R., Zhu, J.Y., Park, J., Shechtman, E., Huang, X.: Motionstream: Real-time video generation with interactive motion controls. arXiv preprint arXiv:2511.01266 (2025)
41. Song, K., Chen, B., Simchowitz, M., Du, Y., Tedrake, R., Sitzmann, V.: History-guided video diffusion. arXiv preprint arXiv:2502.06764 (2025)
42. Sun, W., Zhang, H., Wang, H., Wu, J., Wang, Z., Wang, Z., Wang, Y., Zhang, J., Wang, T., Guo, C.: Worldplay: Towards long-term geometric consistency for real-time interactive world modeling. arXiv preprint arXiv:2512.14614 (2025)
43. Team, R., Gao, Z., Wang, Q., Zeng, Y., Zhu, J., Cheng, K.L., Li, Y., Wang, H., Xu, Y., Ma, S., et al.: Advancing open-source world models. arXiv preprint arXiv:2601.20540 (2026)
44. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
45. Wang, Z., Yuan, Z., Wang, X., Li, Y., Chen, T., Xia, M., Luo, P., Shan, Y.: Motionctrl: A unified and flexible motion controller for video generation. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–11 (2024)
46. Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7623–7633 (2023)

47. Wu, T., Yang, S., Po, R., Xu, Y., Liu, Z., Lin, D., Wetzstein, G.: Video world models with long-term spatial memory. arXiv preprint arXiv:2506.05284 (2025)
48. Xu, J., Huang, Y., Pei, B., Hou, J., Li, Q., Chen, G., Zhang, Y., Feng, R., Xie, W.: Egoexo-gen: Ego-centric video prediction by watching exo-centric videos. arXiv preprint arXiv:2504.11732 (2025)
49. Yang, S., Hou, L., Huang, H., Ma, C., Wan, P., Zhang, D., Chen, X., Liao, J.: Direct-a-video: Customized video generation with user-directed camera movement and object motion. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–12 (2024)
50. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)
51. Yu, M., Hu, W., Xing, J., Shan, Y.: Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 100–111 (2025)
52. Zhang, D.J., Wu, J.Z., Liu, J.W., Zhao, R., Ran, L., Gu, Y., Gao, D., Shou, M.Z.: Show-1: Marrying pixel and latent diffusion models for text-to-video generation. International Journal of Computer Vision **133**(4), 1879–1893 (2025)
53. Zhou, D., Wang, W., Yan, H., Lv, W., Zhu, Y., Feng, J.: Magicvideo: Efficient video generation with latent diffusion models. arXiv preprint arXiv:2211.11018 (2022)