

PE-Field 4D: Video Generation Models as Canvas

YUNPENG BAI*, University of Texas at Austin, USA

HAOXIANG LI, Pixocial Technology, USA

QIXING HUANG, University of Texas at Austin, USA

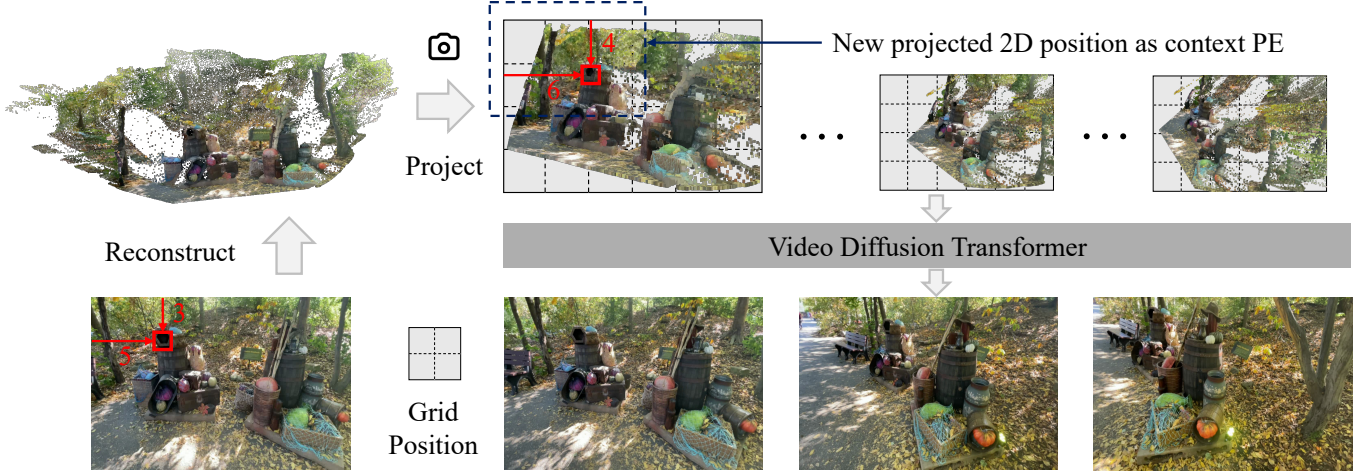


Fig. 1. Overview of our geometry-aware video generation framework. Given an input image/video, we reconstruct its scene geometry and project reference content onto target camera views to obtain position-aligned context. The video diffusion transformer then uses this structured context to synthesize videos that follow the desired camera trajectory while preserving scene appearance and spatial structure.

Diffusion Transformers have recently achieved strong performance in video generation, yet controlling scene geometry under viewpoint changes and camera motion remains challenging. In this work, we revisit the role of positional encoding in video diffusion transformers and show that it provides a useful spatial bias for geometry-aware control. Specifically, if reference tokens are encoded according to their projected locations in the target view, the denoising model is encouraged to retrieve content from position aligned regions of the input video. Building on this observation, we introduce a geometry-aware cross-attention mechanism that enables target video latent tokens to attend to structured context tokens derived from reference images or frames. To establish correspondence between the reference content and the target camera trajectory, we equip the context tokens with a projected positional encoding scheme that combines target-view 2D reprojection with depth-aware disambiguation. At the same time, we preserve the original spatiotemporal positional encoding of the generated video latent, allowing geometric guidance to be injected while maintaining consistency with the video model’s native latent structure. The resulting framework provides a simple and effective approach for controllable video generation. It improves spatial controllability in viewpoint-dependent editing tasks, including camera re-trajectory, novel-view video synthesis, and geometry-aware video editing, while preserving the generative prior of the underlying video diffusion model.

CCS Concepts: • **Computing methodologies** → **Animation**.

Additional Key Words and Phrases: Video Generation, Diffusion Models, Positional Encoding, Spatial Editing

*This work was partly conducted during an internship at Pixocial Technology.

Authors’ addresses: Yunpeng Bai, University of Texas at Austin, USA; Haoxiang Li, Pixocial Technology, USA; Qixing Huang, University of Texas at Austin, USA.

1 INTRODUCTION

Diffusion transformers [Peebles and Xie 2023] have recently emerged as a powerful foundation for visual generation, achieving strong performance in both image and video synthesis. In particular, video diffusion transformers [Gao et al. 2025; Kong et al. 2024; Zhu et al. 2024] inherit the expressive modeling capacity of large-scale image generators while extending it to spatiotemporal data, allowing for increasingly realistic video generation from text, images, and other conditions. Despite this progress, controlling the spatial structure of generated videos remains challenging. Existing video generation models can often synthesize visually plausible motion, yet they still struggle to faithfully preserve scene geometry under viewpoint changes, camera motion, or spatial editing. As a result, 4D video editing tasks such as camera re-trajectory, view-consistent generation, and geometry-aware video editing remain difficult for current transformer-based diffusion models. Existing camera re-trajectory methods either use camera poses as generation conditions [Bai et al. 2025b] or first perform 4D reconstruction and then inpaint the re-rendered results [Ren et al. 2025; YU et al. 2025]. These approaches either lack explicit geometric structure, leading to inaccurate control, or are limited by the quality of the reconstructed scene, and they generally do not generalize well to broader spatial editing tasks.

In this paper, we study 4D generative video editing [Lee et al. 2025] under the video transformers paradigm. Our goal is to leverage rich information in pre-trained models while achieving efficient video editing. Our approach takes inspiration from recent work on *PE-Field* [Bai et al. 2025a], which shows that positional encoding

can be used as a structured interface for geometry-guided generation. Rather than treating positional encoding as a fixed indexing mechanism, PE-Field demonstrates that modifying the positional codes of source tokens according to their projected locations in a target view, i.e. a *reference image*, can effectively guide a diffusion transformer to perform novel view synthesis. This result is conceptually significant: it suggests that positional encoding not only organizes token identity but also actively shapes how attention transports information across views. However, PE-Field has only been validated in image generation models, where the target is a single image and the geometry is relatively simple.

Generalizing the PE-field idea to video editing is difficult, as compared with image generation, video generation requires the model to jointly reason about space, time, and motion. In addition, many practical architectures also rely on temporally compressed latent representations. These characteristics introduce new challenges for PE-based geometry control. First, the model must preserve temporal coherence while incorporating spatially structured context from reference frames. Second, the image editing models used by PE-Field naturally contain an attention pathway between the reference branch and the generated latent, whereas mainstream video generation and editing models typically do not provide such a structure that can be directly reused. Third, temporal compression in video VAEs may blur the correspondence between latent tokens and original frames, making precise positional reassignment ambiguous.

Our starting point is a simple reinterpretation of why PE warping works. We argue that the success of PE-Field can be understood as an *attention-guidance mechanism*: when source tokens are assigned positional encodings corresponding to their projected locations in the target view, attention naturally becomes biased toward retrieving position aligned content. From this perspective, PE warping is especially well suited for cross-attention. If a video diffusion transformer is allowed to query reference tokens whose positional codes already reflect target-view geometry, then the model can directly retrieve the source content that should appear at each target token location. This converts positional encoding from a passive indexing scheme into an active geometric routing signal.

Motivated by this view, we propose a geometry-aware extension of PE-Field for video diffusion transformers. To efficiently handle video data, our method augments a standard video diffusion transformer with a cross-attention mechanism that jointly attends to global target latent memory and geometry-aware context memory. The target video latent tokens keep the model’s original spatiotemporal positional encoding, preserving the pretrained video prior. In contrast, context tokens extracted from reference images or frames are assigned positional encodings based on their projected target-view coordinates. To further disambiguate tokens that project to similar 2D positions, we introduce a depth-aware positional encoding strategy that injects normalized depth offsets into the temporal axis of the context encoding. In addition, to address ambiguity caused by temporal compression in video latent representations, we modify the context construction process so that each latent frame corresponds to a single source frame, enabling accurate positional reassignment.

Our method provides a simple but effective way to 4D generative video generation based on designing positional encoding fields. By preserving the original denoising dynamics of the video model while equipping context tokens with geometrically meaningful positional encodings, the transformer can use reference frames as structured spatial context rather than merely appearance hints. This leads to improved controllability for tasks such as camera re-trajectory and view-consistent video synthesis (see Fig. 1), and more broadly suggests that positional encoding can serve as a general interface for injecting geometric structure into transformer-based generative models.

Our contributions are summarized as follows:

- We extend the PE-Field paradigm from image diffusion to video diffusion transformers, showing that warped positional encoding can be used to guide geometry-aware video generation.
- We first inject geometric structure directly into the video generation Transformer architecture itself and propose a geometry-aware cross-attention design that combines global target latent memory with grouped local context memory.
- We propose a projected positional encoding strategy for context tokens that incorporates target-view 2D reprojection, depth-aware disambiguation, and frame-level correspondence, resolving the temporal compression ambiguity in video latent representations and enabling precise positional reassignment.
- Experiments show that our method outperforms existing approaches on camera re-trajectory tasks and further generalizes to broader 4D spatial editing applications.

2 RELATED WORKS

2.1 Large-Scale Video Synthesis

Recent progress in video generation has been largely driven by scaling both model capacity and video training data [Gao et al. 2025; Kong et al. 2024; Zhu et al. 2024]. Diffusion-based video transformers that model a clip as a joint spatiotemporal sequence [Gao et al. 2025; Kong et al. 2024; Wan et al. 2025; Yang et al. 2024] have shown impressive synthesis quality and have become a widely adopted design for text-to-video and image-to-video generation. However, because the entire sequence must be processed within a single denoising model, memory consumption grows quickly with video length and resolution. This makes long-duration generation difficult and often requires long videos to be produced as separate clips, where cross-clip consistency is hard to guarantee. Auto-regressive video generators [Chen et al. 2024; Gu et al. 2025a; Huang et al. 2025a; Song et al. 2025; Zhang and Agrawala 2025] mitigate this issue by generating future segments conditioned on previously generated content. Although this strategy extends the reachable temporal span, the model can only access a limited history window, so information from earlier parts of the scene may still be gradually lost.

2.2 Spatial Memory and Camera Control

Maintaining a stable scene representation over long time horizons is closely related to camera-controlled generation, since both require

the model to synthesize consistent observations of the same scene across changing viewpoints. In long-form generation, drift in object appearance, layout, or geometry is commonly observed when the accessible context is limited [Decart et al. 2024; Kanervisto et al. 2025; Song et al. 2025]. One family of methods addresses this issue by building an explicit geometric memory from past frames [Wu et al. 2025]. These methods estimate 3D structures such as point clouds and integrate them with volumetric fusion techniques such as TSDF [Zeng et al. 2017], which can help preserve coarse scene layout for subsequent generation. However, explicit geometry is less convenient for open-ended scenes and may lose detailed appearance or thin structures during reconstruction and fusion.

Camera trajectory control faces a similar alignment problem. Existing methods typically either construct explicit guidance from an input image or video, such as point-based renderings [Cao et al. 2025; Feng et al. 2024; Li et al. 2025b; Ma et al. 2025; You et al. 2024; YU et al. 2025; Yu et al. 2024; Zhai et al. 2025], tracked correspondences [Gu et al. 2025b], or optical flow [Burgert et al. 2025], or inject camera information into pretrained video diffusion models through additional trainable modules and conditioning representations, including pose parameters [Bai et al. 2025b; Wang et al. 2024], Plücker ray embeddings [Bahmani et al. 2025; He et al. 2024, 2025], and relative camera encodings [Zhang et al. 2025]. In contrast to these approaches, our work focuses on explicit token-wise spatial alignment between generated tokens and reference observations. We use this alignment as a common mechanism for camera control and spatial memory, leading to a unified formulation based on time-aware positional encoding warping.

2.3 Positional Encoding Designs of DiT

Several recent papers have studied how to modify RoPE for various 3D vision tasks, examples include CaPE [Li et al. 2025a] and Ray-RoPE [Wu et al. 2026]. Closest related to this work is PE-field [Bai et al. 2025a], which introduces the idea of fine-tuning a pre-trained diffusion transformer based on a novel positional encoding scheme for novel-view synthesis. [Gao et al. 2026] applied PE-field to synthesize novel-views for reconstructing 3D Gaussians. Two concurrent papers [Xu et al. 2026; Yu et al. 2026] extend PE-field to videos. They are mostly suitable for videos of camera motions of static scenes. Our approach differs from them in terms of specific PE strategies, focusing on leveraging information in pre-trained models, efficiency in training and inference, and modeling dynamic scenes.

3 METHOD

3.1 PE-Field Design Principles for Video Diffusion Transformers

The key insight of PE-Field is that positional encoding does not merely index token identity but also directly shapes how attention routes information. By assigning source tokens with positional codes that correspond to their projected locations in a target view, the model is encouraged to retrieve content from position-aligned regions during denoising. In this work, we extend this paradigm from image diffusion to video diffusion transformers, where generation

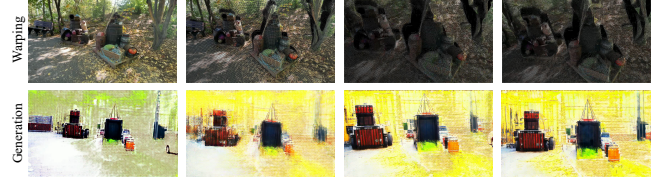


Fig. 2. Visualization of position-aligned attention in a pretrained video diffusion transformer. We insert cross-attention into the pretrained video model and assign new positional codes to the reference tokens shown in the top row. During generation, these tokens are attended by the video latent tokens according to their assigned positions, leading to the corresponding content appearing at the aligned spatial regions in the generated results shown in the bottom row.

must simultaneously model spatial structure, temporal coherence, and camera motion.

To validate this hypothesis in pretrained video DiTs, we inject a cross-attention mechanism into the pretrained model and manually reassign position context tokens. As shown in Fig. 2, the generated video tends to place the referenced content at the spatial locations specified by these reassigned positional encodings. This provides direct evidence that pretrained video diffusion transformers already exhibit a useful spatial bias: target latent tokens preferentially attend to reference tokens from position-aligned regions. Based on this observation, we augment a standard video diffusion transformer with a geometry-aware cross-attention mechanism. The target video latent tokens preserve the model’s original spatiotemporal positional encoding, while context tokens are assigned positional encodings according to their reprojected target-view coordinates and depth-aware ordering. This allows the model to use reference frames as spatially grounded context for camera re-trajectory and spatial aware video generation, while also providing a practical way to address the memory challenge in video modeling.

3.2 Geometry-Aware Cross-Attention

Let $x \in \mathbb{R}^{B \times N \times D}$ denote the noisy latent tokens of the target video to be denoised, and let $y \in \mathbb{R}^{B \times M \times D}$ denote context tokens extracted from reference images or frames. In a standard video diffusion transformer, attention is performed only among the target latent tokens, with positional encoding defined over the three video axes (t, h, w) . We preserve this encoding for the target latent tokens and modify only the encoding of the context tokens.

Our attention block augments the original self-attention with an additional geometry-aware context branch. Queries are always computed from the target latent tokens x :

$$Q = W_q x, \quad (1)$$

while two key-value branches are constructed:

$$K_x = W_k x, \quad V_x = W_v x, \quad (2)$$

$$K_y = W_k y, \quad V_y = W_v y. \quad (3)$$

where K_x and V_x come from latent target video tokens and K_y and V_y come from latent reference frame tokens. After RMS normalization, rotary positional encoding [Su et al. 2024] is applied to Q and K_x

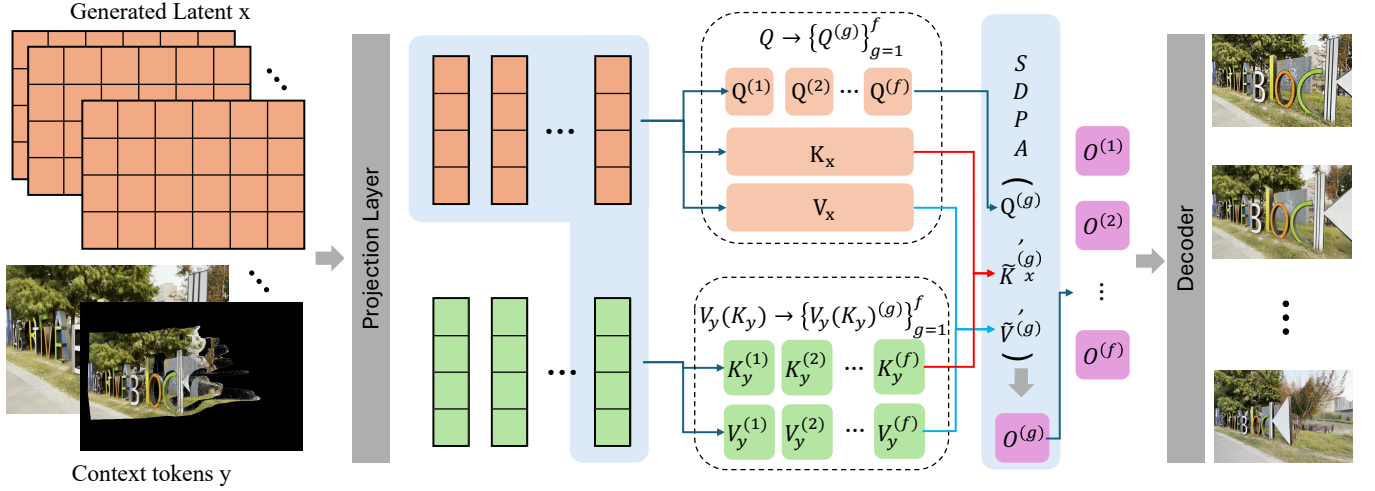


Fig. 3. Overview of the proposed geometry-aware cross-attention mechanism for video diffusion transformers. Given the noisy target video latent tokens x and the reference context tokens y , we compute queries from the target tokens and construct two key-value branches: a global self-attention memory (K_x, V_x) from the target video latent and a geometry-aware context memory (K_y, V_y) from the reference tokens. The target queries are divided into groups, while the self-attention memory is shared globally across all groups and the context memory is consumed in a grouped local manner. For each group g , attention is computed over the concatenated memory $[\tilde{K}^{(g)}, \tilde{V}^{(g)}] = [K_x, K_y^{(g)}; V_x, V_y^{(g)}]$, allowing each target token to jointly attend to global target-video information and position-aligned reference content. The context keys are encoded with projected target-view positional encodings, so the reference tokens are retrieved according to where they should appear in the generated view.

using the original video latent PE, while K_y is encoded using the projected PE to be introduced in Sec. 3.3. The final memory used by attention is

$$K = [K_x; K_y], \quad V = [V_x; V_y], \quad (4)$$

where $[\cdot; \cdot]$ denotes concatenation along the sequence dimension.

This design turns cross-attention into a geometry-aware retrieval process. The target noisy tokens still interact globally through the original self-attention branch, while also querying context tokens whose positional encodings already indicate where they should appear in the target view. As a result, the model can directly select reference content according to the position of the target-view, rather than relying purely on content similarity.

Global self-attention with grouped local context attention. A practical issue is that dense attention over all target and context tokens becomes expensive for long video sequences. We therefore use a structured attention pattern in which the self-attention branch remains the same, while the context branch is consumed in a grouped local manner.

Specifically, let f denote the number of latent frames after temporal compression. We divide the sequence into f groups, each corresponding to one latent frame and containing $G = N/f$ tokens. We reshape the query tensor into groups,

$$Q \rightarrow \{Q^{(g)}\}_{g=1}^f, \quad Q^{(g)} \in \mathbb{R}^{B \times G \times D}, \quad (5)$$

The self-attention memory (K_x, V_x) is treated as global and shared across all groups, while the context memory (K_y, V_y) is partitioned

group-wise:

$$K_y \rightarrow \{K_y^{(g)}\}_{g=1}^f, \quad V_y \rightarrow \{V_y^{(g)}\}_{g=1}^f. \quad (6)$$

For each group g , attention is computed against the concatenated memory

$$\tilde{K}^{(g)} = [K_x; K_y^{(g)}], \quad \tilde{V}^{(g)} = [V_x; V_y^{(g)}]. \quad (7)$$

The group output is then

$$O^{(g)} = \text{SDPA}(Q^{(g)}, \tilde{K}^{(g)}, \tilde{V}^{(g)}), \quad (8)$$

where $\text{SDPA}(\cdot)$ denotes scaled dot-product attention. In this way, the cross-attention to context tokens is restricted to the content corresponding to the same latent frame. The outputs of all groups are finally merged back into the original token order.

This hybrid formulation preserves long-range global reasoning through the target latent tokens, while keeping the geometry conditioned context lookup more spatially localized and efficient. Importantly, this is not a separate module from cross-attention, but the cross-attention mechanism itself: each target query attends jointly to global target latent memory and grouped geometry-aware context memory within a single attention operator.

3.3 Projected Positional Encoding for Context Tokens

Another main design lies in how context tokens are positionally encoded. In a standard video diffusion transformer, positional encoding is defined over the latent video grid and factorized along time and space. We leave this encoding unchanged for the target latent tokens x . However, for context tokens y , we replace their original source-view coordinates with projected target-view coordinates.

For each context token, we compute its 2D projection onto the target camera plane using the source-view geometry. Denote the projected position by $(\tilde{h}, \tilde{w})$. Instead of assigning the token a positional code based on its original source-frame location, we encode it using this projected target-plane location. Thus, context tokens are indexed according to where they should appear in the generated target view. This follows the central PE-Field principle, now adapted to video generation. By changing only the positional code while leaving the token content unchanged, we bias attention to associate target noisy tokens with context tokens that are position aligned in the target view.

Depth-aware positional disambiguation. Projected 2D coordinates alone are not always sufficient, since multiple tokens may land near the same target-plane position while lying at different depths. To distinguish such cases, we additionally encode depth into the positional code. Rather than introducing a separate depth axis, we incorporate depth into the same positional encoding framework by adding a depth-dependent offset to the temporal axis of the context token. Let t denote the frame index of the context token and d its depth. We define an effective temporal coordinate

$$t' = t + \Delta(d), \quad (9)$$

where $\Delta(d)$ is a normalized depth offset. In practice, we normalize depth to the interval $[0, 0.1]$. As a result, tokens originating from frame t occupy a temporal encoding range within $[t, t + \frac{1}{10}]$, which preserves frame-level separation while introducing an additional front-to-back ordering among tokens within the same frame.

The final positional encoding of a context token is, therefore, constructed from

$$(t + \Delta(d), \tilde{h}, \tilde{w}). \quad (10)$$

This design allows the model to distinguish tokens that are close in projected 2D position but belong to different depth layers, while still keeping the overall temporal structure compatible with the original video transformer.

3.4 Resolving Temporal Compression Ambiguity in Context Latents

A practical complication arises from the latent representation used in video diffusion models. Many video VAEs temporally compress the input sequence, for example by merging several consecutive RGB frames into a single latent frame. In such cases, the exact contribution of each original frame to a compressed latent token is unknown. This poses a problem for our method, since a single compressed latent frame would correspond to multiple projected positional encodings from different original frames, making the positional assignment inherently ambiguous.

To resolve this ambiguity, we modify the context construction procedure. Instead of feeding distinct consecutive frames directly into the video encoder, we repeat each source frame multiple times before compression, such that each resulting latent frame corresponds to only one original frame. After encoding, each latent frame therefore contains information from a single original frame rather than a mixture of multiple frames.

This design also requires aligning the temporal coordinates of context and generated latents during cross-attention. While each generated latent corresponds to a compressed temporal interval, e.g., four RGB frames, our repeated context latents correspond to individual source frames within that interval. We therefore normalize the context-frame indices to the same latent-time range as the generated tokens. For example, the four source frames associated with one generated latent step are assigned fractional temporal coordinates 1, 1.25, 1.5, and 1.75. This fractional indexing preserves the temporal order of individual source frames while ensuring that their positional encodings remain compatible with the compressed latent timeline of the video model.

4 EXPERIMENTAL RESULTS

4.1 Implementation Details

We perform our experiments mainly on the 14B versions of the Wan2.1 T2V models [Wan et al. 2025]. To adapt the model to the newly inserted cross-attention modules, we perform LoRA fine-tuning. Specifically, LoRA adapters are inserted into the major projection and feed-forward layers, including the q , k , v , o , and ffn layers. We set the LoRA rank to 32 for all experiments. The model is trained with a learning rate of 1×10^{-4} .

Our training data are primarily constructed from the MultiCam dataset [Bai et al. 2025b], which provides multi-view video sequences suitable for learning viewpoint-dependent generation and camera re-trajectory control. Since the data set does not provide ground-truth depth maps, we estimate the depth and corresponding camera parameters using ViPE [Huang et al. 2025b], and use them as input of geometric conditions to construct the projected positional encodings and geometry-aware context tokens.

4.2 Generation Results

Following the evaluation protocol of recent camera-controlled video generation methods [Park et al. 2025b], we build a test benchmark from 90 videos in the DAVIS dataset and use target camera trajectories provided by ReCamMaster [Bai et al. 2025b]. The selected videos contain diverse object motions, scene layouts, and fine geometric structures, allowing us to evaluate both visual fidelity and geometry-aware camera control under challenging trajectory transformations. For each input video, we generate novel-view results conditioned on the target camera motion and compare our method with ViewCrafter [Yu et al. 2024], EX-4D [Hu et al. 2025], Uni3C [Cao et al. 2025], Vista4D [Lin et al. 2026], ReCamMaster [Bai et al. 2025b], GEN3C [Ren et al. 2025], TrajectoryCrafter [YU et al. 2025], and ReDirector [Park et al. 2025b].

Following the evaluation setting of ReDirector [Park et al. 2025b], we assess the generated videos from three complementary aspects: visual quality, geometric consistency, and camera-control accuracy. Specifically, we adopt VBench [Huang et al. 2024] metrics to evaluate perceptual and temporal quality, use Dyn-MET3R [Park et al. 2025a] to measure the geometric consistency of the generated retargeted videos, and compute per-frame MET3R [Asim et al. 2025] to evaluate consistency between the generated results and the input video. In addition, we report translation and rotation errors, denoted as



Fig. 4. Qualitative comparison on the *bmx-trees* video from DAVIS [Pont-Tuset et al. 2017]. Under the transformed camera trajectory, our method preserves more accurate scene geometry and fine structures, such as the bicycle frame, wheels, and thin foreground/background elements. Compared with TrajectoryCrafter, ReCamMaster, and GEN3C, our results show better geometric consistency and fewer distortions in detailed regions during viewpoint changes.

Table 1. Quantitative evaluation on DAVIS videos using camera trajectories from ReCamMaster. We compare against representative camera-controlled video generation and retargeting methods across visual quality, geometric consistency, and camera accuracy. Our method achieves the best overall performance, demonstrating stronger appearance preservation, more stable geometry under viewpoint changes, and more accurate adherence to the target camera motion.

Method	Visual Quality↑						Geometric Consistency		Camera Accuracy	
	Subject Consistency	Background Consistency	Aesthetic Quality	Imaging Quality	Temporal Flickering	Motion Smoothness	Dyn-MEt3R↑	MEt3R↓	TransErr↓	RotErr↓
ViewCrafter	0.8427	0.8716	0.4628	0.5485	0.8842	0.9137	0.6124	0.4869	0.1047	12.638
EX-4D	0.8713	0.8984	0.4869	0.5842	0.9156	0.9468	0.6947	0.3915	0.0736	7.842
Uni3C	0.8896	0.9118	0.5017	0.6214	0.9398	0.9635	0.7462	0.3467	0.0469	4.913
Vista4D	0.8991	0.9175	0.5109	0.6538	0.9536	0.9771	0.8123	0.3059	0.0167	1.984
ReCamMaster	0.8983	0.9182	0.5052	0.6448	0.9543	0.9752	0.7721	0.3585	0.0301	2.412
GEN3C	0.9011	0.9202	0.5097	0.6259	0.9421	0.9673	0.7426	0.3462	0.0548	6.728
TrajectoryCrafter	0.8824	0.9146	0.5031	0.6013	0.9319	0.9605	0.7315	0.3328	0.0684	8.943
ReDirector	0.8977	0.9164	0.5078	0.6521	0.9527	0.9784	0.8041	0.3186	0.0189	2.107
w/o Temporal Dis.	0.8946	0.9138	0.5061	0.6417	0.9435	0.9688	0.7564	0.3479	0.0317	3.284
w/o Depth	0.9069	0.9228	0.5171	0.6675	0.9661	0.9864	0.8126	0.3047	0.0161	1.982
Ours	0.9098	0.9247	0.5196	0.6731	0.9704	0.9892	0.8235	0.2968	0.0142	1.887

TransErr and RotErr, which quantify how well the generated video follows the target camera motion based on relative pose estimation from ViPE [Huang et al. 2025b].

As shown in Tab. 1, our method consistently outperforms previous approaches in visual quality, geometric consistency, and camera precision. Compared with ReCamMaster and TrajectoryCrafter, our approach better preserves subject and background appearance while



Fig. 5. Qualitative results on two challenging cases, including a non-rigid fire scene (top) and a scene with complex occlusions (bottom), demonstrating the robustness of our method.

reducing geometric distortions during large camera changes. Compared to GEN3C, our method produces a more stable scene structure and more accurate camera follow-up. Although ReDirector provides strong camera control, our method further improves both DynMEt3R and MEt3R, indicating better geometry preservation in the generated retakes. The lower translation and rotation errors also demonstrate that our results follow the target trajectories more faithfully. We further evaluate our method on challenging scenarios, including non-rigid motion (e.g., flames), highly dynamic scenes, and complex occlusions. As shown in Fig. 5, our method remains effective across these challenging cases.

4.3 Applications

Beyond camera trajectory transformation, our framework naturally extends to other spatially aware generation tasks, such as animating a single image into a video. The key reason is that our method does not rely on task-specific supervision but instead injects spatially aligned context through projected positional encodings. When the input is a single reference image, its visual tokens can still be treated as geometry-aware context, allowing the denoising model to retrieve content from spatially corresponding regions while generating temporally coherent motion.

As shown in Fig. 6, our method can animate static images while preserving the spatial layout and fine structures of the reference content. Compared with generation driven only by global image conditioning, the proposed position-aligned context provides more

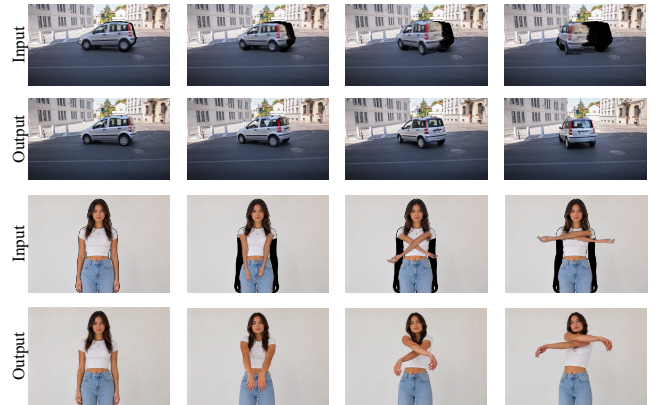


Fig. 6. **General applications beyond camera trajectory transformation.** Our method not only supports controllable camera trajectory transformation, but also naturally extends to other spatially aware generation tasks, such as object motion synthesis and pose-guided animation, by aligning the generated content with structured geometric guidance.

explicit spatial guidance, helping the model maintain object identity, local appearance, and scene structure during motion synthesis. This demonstrates that our approach is not limited to camera-controlled video retargeting but can serve as a general mechanism for spatially grounded video generation.

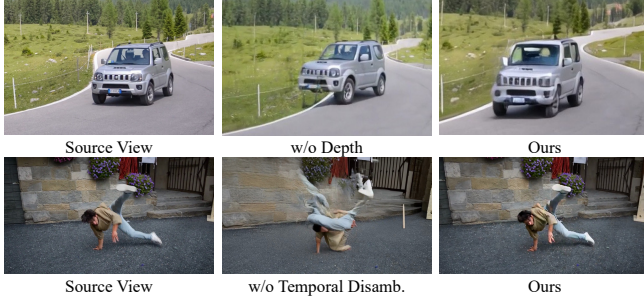


Fig. 7. Ablation study on depth and temporal disambiguation. Compared with variants without depth-aware or temporal disambiguation, our full method preserves more accurate geometry and achieves more faithful generation under camera transformation.

4.4 Ablation Studies

Depth-aware positional disambiguation. We ablate the depth-aware offset in our positional encoding. Without depth, tokens are distinguished only by projected 2D coordinates, which becomes ambiguous when foreground and background regions overlap in the target view. As shown in Fig. 7, this leads to less accurate geometry and more distortions near fine structures and occlusion boundaries. The quantitative results in Tab. 1 also show a consistent drop after removing depth-aware disambiguation.

Temporal compression ambiguity. We ablate our frame-repetition strategy by directly using the original video VAE compression, where several RGB frames are encoded into one context latent frame. This mixes information from multiple source frames and makes projected positional assignment ambiguous. As shown in Fig. 7, this baseline fails to localize reference content accurately, resulting in weaker geometric alignment. Tab. 1 further shows a clear drop in geometric consistency and camera accuracy, confirming the importance of precise frame-level positional assignment.

5 CONCLUSIONS

We extended the PE-Field paradigm from image diffusion to video diffusion transformers. Our key idea is to interpret warped positional encoding as a geometry-aware attention signal, allowing target video latent tokens to retrieve spatially aligned information from reference frames through cross-attention. Our results demonstrate that positional encoding provides an effective interface for incorporating geometric cues into video diffusion transformers.

Limitations. Our method introduces additional context tokens, particularly when source frames are encoded separately to resolve temporal-compression ambiguity. As shown in Table 2, for generating an 81-frame video at 480p resolution on an NVIDIA H100 GPU, the proposed frame-wise context attention increases computation and memory over the base DiT, while remaining substantially more efficient than global context attention. Moreover, our method relies on estimated depth and camera poses; estimation errors may produce inaccurate projected positional encodings and imperfect spatial alignment.

Table 2. Efficiency comparison for generating an 81-frame video at 480p resolution on an NVIDIA H100 GPU.

Method	Attention Design	FLOPs	Latency	VRAM
Base DiT	No geometry context	167.8 PFLOPs	~8.1 min	~40–45 GB
PE-Field 4D	Frame-wise one-way K/V attention	231.0 PFLOPs	~11.6 min	~50–58 GB
Global context	All-to-all context attention	573.4 PFLOPs	~27.7 min	~51–59 GB

Future work. A promising direction is to replace temporally expanded video latents with more compact geometry-aware context representations that preserve essential spatial and appearance information while reducing token count and attention cost.

Acknowledgements. This project was supported by NSF-2047677, 2413161, 2504906, and 2515626; GIFTs from Adobe, Google, and Chen&Meng Family Foundation; and computing support on the Vista GPU Cluster through the Center for Generative AI (CGAI) and the Texas Advanced Computing Center (TACC) at UT Austin.

REFERENCES

- Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele, and Jan Eric Lenssen. 2025. Met3r: Measuring multi-view consistency in generated images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6034–6044.
- Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B Lindell, and Sergey Tulyakov. 2025. Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 22875–22889.
- Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuoqiu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. 2025b. Recamaster: Camera-controlled generative rendering from a single video. *arXiv preprint arXiv:2503.11647* (2025).
- Yunpeng Bai, Haoxiang Li, and Qixing Huang. 2025a. Positional encoding field. *arXiv preprint arXiv:2510.20385* (2025).
- Ryan Burgert, Yuancheng Xu, Wenqi Xian, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, Yitong Deng, Lingxiao Li, Mohsen Mousavi, et al. 2025. Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 13–23.
- Chenjie Cao, Jinghui Zhou, Shikai Li, Jingyun Liang, Chaohui Yu, Fan Wang, Xiangyang Xue, and Yanwei Fu. 2025. Uni3c: Unifying precisely 3d-enhanced camera and human motion controls for video generation. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*. 1–12.
- Boyuan Chen, Diego Marti Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. 2024. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems* 37 (2024), 24081–24125.
- Etched Decart, Quinn McIntyre, Spruce Campbell, Xinlei Chen, and Robert Wachen. 2024. Oasis: A universe in a transformer. URL: <https://oasis-model.github.io> (2024).
- Wanquan Feng, Jiawei Liu, Pengqi Tu, Tianhao Qi, Mingzhen Sun, Tianxiang Ma, Songtao Zhao, Siyu Zhou, and Qian He. 2024. I2vcontrol-camera: Precise video camera control with adjustable motion strength. *arXiv preprint arXiv:2411.06525* (2024).
- AO Gao, Jingyu Gong, Xin Tan, Zhizhong Zhang, and Yuan Xie. 2026. GSCompleter: A Distillation-Free Plugin for Metric-Aware 3D Gaussian Splatting Completion in Seconds. *arXiv:2604.20155 [cs.CV]* <https://arxiv.org/abs/2604.20155>
- Yu Gao, Haoyuan Guo, Tuyen Hoang, Weilin Huang, Lu Jiang, Fangyuan Kong, Huixia Li, Jiashi Li, Liang Li, Xiaojie Li, et al. 2025. Seedance 1.0: Exploring the Boundaries of Video Generation Models. *arXiv preprint arXiv:2506.09113* (2025).
- Yuchao Gu, Weijia Mao, and Mike Zheng Shou. 2025a. Long-context autoregressive video modeling with next-frame prediction. *arXiv preprint arXiv:2503.19325* (2025).
- Zekai Gu, Rui Yan, Jiahao Lu, Peng Li, Zhiyang Dou, Chenyang Si, Zhen Dong, Qifeng Liu, Cheng Lin, Ziwei Liu, et al. 2025b. Diffusion as shader: 3d-aware video diffusion for versatile video generation control. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Papers*. 1–12.
- Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. 2024. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101* (2024).
- Hao He, Ceyuan Yang, Shanchuan Lin, Yinghao Xu, Meng Wei, Liangke Gui, Qi Zhao, Gordon Wetzstein, Lu Jiang, and Hongsheng Li. 2025. Cameractrl ii: Dynamic scene exploration via camera-controlled video diffusion models. *arXiv preprint arXiv:2503.10592* (2025).

- Tao Hu, Haoyang Peng, Xiao Liu, and Yuewen Ma. 2025. Ex-4d: Extreme viewpoint 4d video synthesis via depth watertight mesh. *arXiv preprint arXiv:2506.05554* (2025).
- Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, et al. 2025b. Vipe: Video pose engine for 3d geometric perception. *arXiv preprint arXiv:2508.10934* (2025).
- Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. 2025a. Self Forcing: Bridging the Train-Test Gap in Autoregressive Video Diffusion. *arXiv preprint arXiv:2506.08009* (2025).
- Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. 2024. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21807–21818.
- Anssi Kanervisto, Dave Bignell, Linda Yilin Wen, Martin Grayson, Raluca Georgescu, Sergio Valcarcel Macua, Shan Zheng Tan, Tabish Rashid, Tim Pearce, Yuhao Cao, et al. 2025. World and human action models towards gameplay ideation. *Nature* 638, 8051 (2025), 656–663.
- Weiwei Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. 2024. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603* (2024).
- Yao-Chih Lee, Zhoutong Zhang, Jiahui Huang, Jui-Hsien Wang, Joon-Young Lee, Jia-Bin Huang, Eli Shechtman, and Zhengqi Li. 2025. Generative Video Motion Editing with 3D Point Tracks. *arXiv preprint arXiv:2512.02015* (2025).
- Ruilong Li, Brent Yi, Junchen Liu, Hang Gao, Yi Ma, and Angjoo Kanazawa. 2025a. Cameras as Relative Positional Encoding. *Advances in Neural Information Processing Systems* (2025).
- Teng Li, Guangcong Zheng, Rui Jiang, Shuigen Zhan, Tao Wu, Yehao Lu, Yining Lin, Chuanyun Deng, Yapan Xiong, Min Chen, et al. 2025b. Realcam-i2v: Real-world image-to-video generation with interactive complex camera control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 28785–28796.
- Kuan Heng Lin, Zhizheng Liu, Pablo Salamañca, Yash Kant, Ryan Burgert, Yuancheng Xu, Koichi Namekata, Yiwei Zhao, Bolei Zhou, Micah Goldblum, et al. 2026. Vista4D: Video Reshooting with 4D Point Clouds. *arXiv preprint arXiv:2604.21915* (2026).
- Baorui Ma, Huachen Gao, Haoge Deng, Zhengxiong Luo, Tiejun Huang, Lulu Tang, and Xinlong Wang. 2025. You see it, you got it: Learning 3d creation on pose-free videos at scale. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2016–2029.
- Byeongjun Park, Hyojun Go, Hyelin Nam, Byung-Hoon Kim, Hyungjin Chung, and Changick Kim. 2025a. Steerx: Creating any camera-free 3d and 4d scenes with geometric steering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 27326–27337.
- Byeongjun Park, Byung-Hoon Kim, Hyungjin Chung, and Jong Chul Ye. 2025b. ReDirector: Creating Any-Length Video Retakes with Rotary Camera Encoding. *arXiv preprint arXiv:2511.19827* (2025). <https://arxiv.org/abs/2511.19827>
- William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4195–4205.
- Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. 2017. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675* (2017).
- Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. 2025. GEN3C: 3D-Informed World-Consistent Video Generation with Precise Camera Control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. 2025. History-guided video diffusion. *arXiv preprint arXiv:2502.06764* (2025).
- Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing* 568 (2024), 127063.
- Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. 2025. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314* (2025).
- Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. 2024. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*. 1–11.
- Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetstein. 2025. Video World Models with Long-term Spatial Memory. *arXiv preprint arXiv:2506.05284* (2025).
- Yu Wu, Minsik Jeon, Jen-Hao Rick Chang, Oncel Tuzel, and Shubham Tulsiani. 2026. RayRoPE: Projective Ray Positional Encoding for Multi-view Attention. *arXiv preprint arXiv:2601.15275* (2026).
- Tianxing Xu, Zixuan Wang, Guangyuan Wang, Li Hu, Zhongyi Zhang, Peng Zhang, Bang Zhang, and Song-Hai Zhang. 2026. UCM: Unifying Camera Control and Memory with Time-aware Positional Encoding Warping for World Models. *arXiv:2602.22960 [cs.CV]* <https://arxiv.org/abs/2602.22960>
- Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. 2024. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072* (2024).
- Meng You, Zhiyu Zhu, Hui Liu, and Junhui Hou. 2024. Nvs-solver: Video diffusion model as zero-shot novel view synthesizer. *arXiv preprint arXiv:2405.15364* (2024).
- Mark YU, Wenbo Hu, Jinbo Xing, and Ying Shan. 2025. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. *arXiv preprint arXiv:2503.05638* (2025).
- Wei Yu, Runjia Qian, Yumeng Li, Liquan Wang, Songheng Yin, Sri Siddharth Chakravarthy P, Dennis Anthony, Yang Ye, Yidi Li, Weiwei Wan, and Animesh Garg. 2026. MosaicMem: Hybrid Spatial Memory for Controllable Video World Models. *CoRR* abs/2603.17117 (2026). [arXiv:2603.17117](https://arxiv.org/abs/2603.17117) doi:10.48550/ARXIV.2603.17117
- Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. 2024. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048* (2024).
- Andy Zeng, Shuran Song, Matthias Nießner, Matthew Fisher, Jianxiong Xiao, and Thomas A. Funkhouser. 2017. 3DMatch: Learning Local Geometric Descriptors from RGB-D Reconstructions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 199–208. doi:10.1109/CVPR.2017.29
- Shangjin Zhai, Zhichao Ye, Jialin Liu, Weijian Xie, Jiaqi Hu, Zhen Peng, Hua Xue, Danpeng Chen, Xiaomeng Wang, Lei Yang, et al. 2025. Stargen: A spatiotemporal autoregression framework with video diffusion model for scalable and controllable scene generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 26822–26833.
- Cheng Zhang, Boying Li, Meng Wei, Yan-Pei Cao, Camilo Cruz Gambardella, Dinh Phung, and Jianfei Cai. 2025. Unified Camera Positional Encoding for Controlled Video Generation. *arXiv preprint arXiv:2512.07237* (2025).
- Lvmin Zhang and Maneesh Agrawala. 2025. Packing input frame context in next-frame prediction models for video generation. *arXiv preprint arXiv:2504.12626* (2025).
- Zheng Zhu, Xiaofeng Wang, Wangbo Zhao, Chen Min, Bohan Li, Nianchen Deng, Min Dou, Yuqi Wang, Botian Shi, Kai Wang, et al. 2024. Is sora a world simulator? a comprehensive survey on general world models and beyond. *arXiv preprint arXiv:2405.03520* (2024).

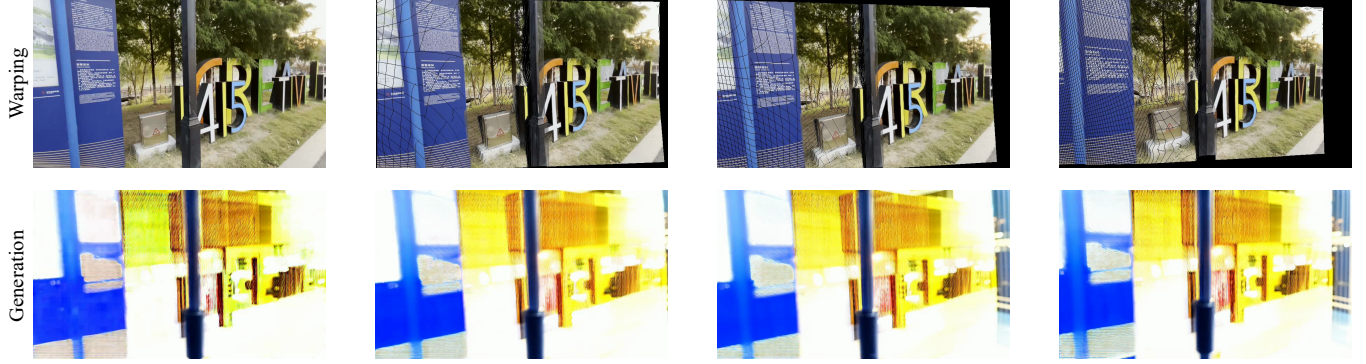


Fig. 8. Additional visualization of position-aligned attention in a pretrained video diffusion transformer. After assigning projected positional codes to context tokens, each target latent attends to the context content according to its projected location in the target view.



Fig. 9. Additional qualitative comparisons. Our method better preserves object structure under large camera changes and produces more position-aligned results than prior methods.

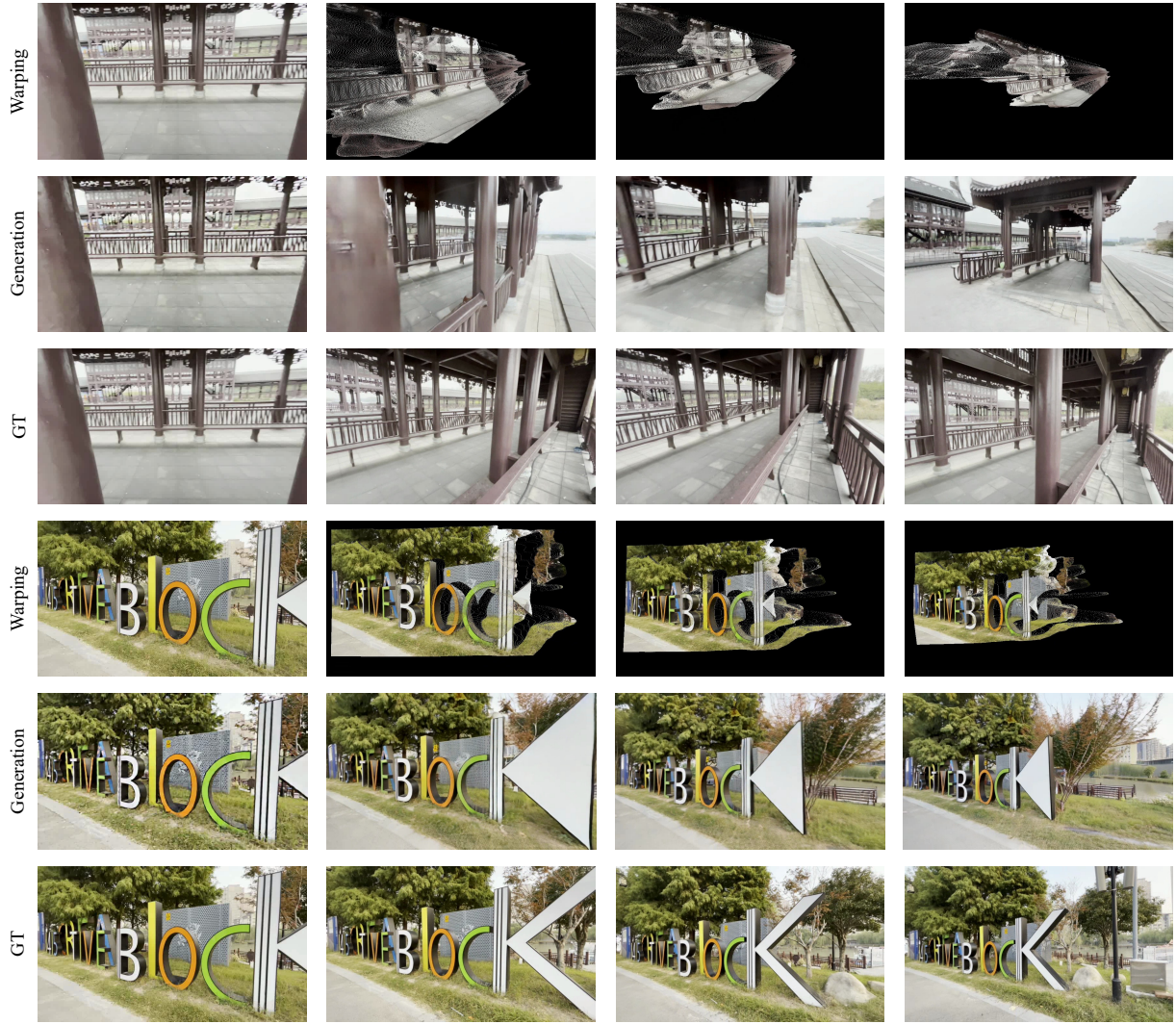


Fig. 10. Image-to-video generation results of our method on the DL3DV dataset. The leftmost image is the input image.

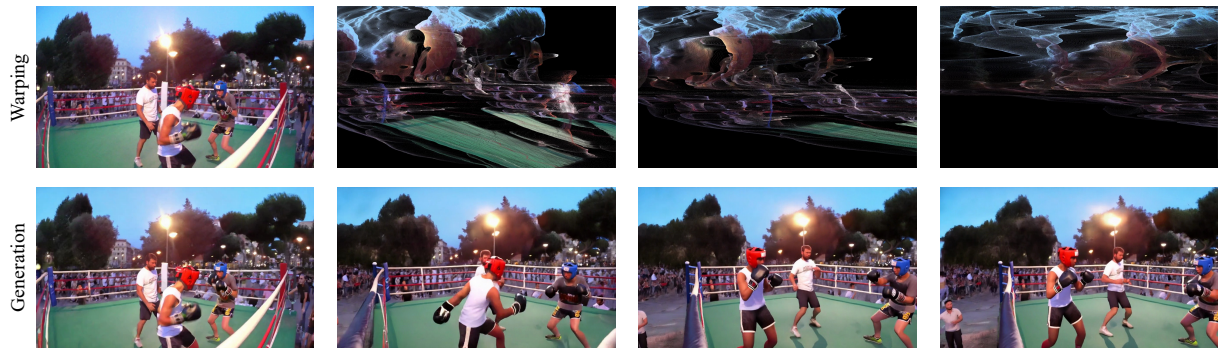


Fig. 11. This example shows a case where the reconstructed geometry is inaccurate. Our method can still generate plausible video results even when the reconstruction is noisy or erroneous.