RECOG: REIFYING ENDOGENOUS CORRESPONDENCE FOR SUBJECT-AWARE CAMERA-CONTROLLABLE VIDEO BACKGROUND GENERATION

Anonymous authors
Paper under double-blind review

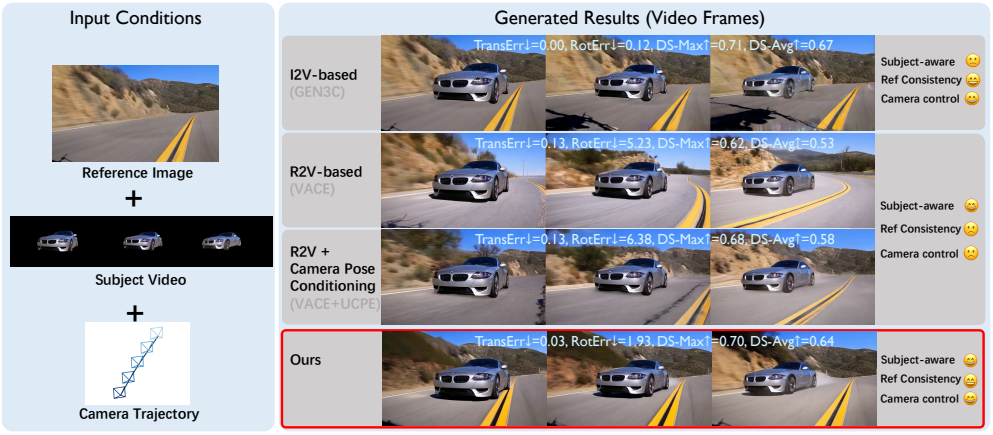

Figure 1: We present ReCoG for subject-aware camera-controllable video background generation. Given a subject video, camera trajectory, and reference image, ReCoG synthesizes a coherent background that adapts to the subject while faithfully following the prescribed camera motion.

ABSTRACT

Subject-aware camera-controllable video background generation aims to synthesize a background video from a reference image that flexibly adapts to the subject while following a prescribed camera trajectory. Existing Reference-to-Video (R2V) methods provide flexible scene conditioning, but lack explicit reference-to-video correspondence, forcing the model to implicitly infer reference geometry and often resulting in inaccurate camera control and reference inconsistency. We reveal that such correspondence, although absent from the R2V formulation, emerges internally during diffusion. Building on this finding, we propose ReCoG, which reifies endogenous reference-to-video correspondence into explicit geometric guidance. ReCoG first locates informative correspondence within the Transformer, then exploits region-level coherence to refine noisy early-denoising matches, and materializes the reliable correspondences into sparse anchors that are warped along the target camera trajectory. Ground-truth-derived anchors that approximate the inference-time sparse-anchor distribution further stabilize training. Across synthetic and real-world evaluations, ReCoG achieves the best performance in both reference consistency and camera control accuracy among the compared methods. These results demonstrate that endogenous correspondence provides an effective route to explicit geometric control while retaining the flexibility of R2V generation.

1 INTRODUCTION

Recent advances in video generation (Wan et al., 2025; Yang et al., 2024; Seedance et al., 2026; Ali et al., 2025) have greatly expanded the possibilities of controllable content creation. However, existing efforts have largely focused on foreground control, while controllable background generation remains

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

less explored. An important problem is **subject-aware camera-controllable video background generation**: given a subject video with its camera trajectory and a reference image specifying the target environment, the goal is to synthesize a background video that naturally accommodates the subject while following the prescribed camera motion, as shown in Figure 1. Such a capability is valuable for digital filmmaking and visual effects, where captured actors or rendered 3D assets need to be integrated into new environments. Recent methods typically adopt Reference-to-Video (R2V) models (Pan et al., 2024) for flexible scene conditioning, and further inject target camera trajectories through pose embeddings (Yao et al., 2025). Despite this progress, generated backgrounds can still deviate from the reference scene or fail to evolve consistently with the prescribed camera motion.

To understand this limitation, it is instructive to compare R2V with camera-controlled Image-to-Video (I2V) generation (Ren et al., 2025). In I2V, the reference image is inherently the first frame of the target video, establishing deterministic reference-to-video correspondence. This allows the reference image to be lifted into 3D using estimated depth and rendered along the target camera trajectory, providing explicit geometric guidance for precise camera control. However, this formulation is ill-suited to subject-aware background generation: an arbitrary background reference is not necessarily a valid first frame for the prescribed foreground, as its viewpoint or spatial layout may be incompatible and it typically lacks subject-induced content such as shadows. R2V avoids this restriction by treating the reference image as a flexible appearance condition, allowing the generated scene to selectively inherit reference content and adapt to the foreground context (Pan et al., 2024). However, this flexibility comes without deterministic reference-to-video correspondence. Even when camera poses are provided through Plücker embeddings (Yao et al., 2025; Sitzmann et al., 2021), the model is only given the target camera configuration, without an explicit mapping of where structures in the reference image should appear in each generated frame. The reference geometry must therefore be inferred implicitly, leading to inaccurate camera-induced motion, structural drift, or reference inconsistency. This raises a central question: *Can we preserve the flexibility of R2V while recovering explicit reference-to-video correspondence for precise camera control?*

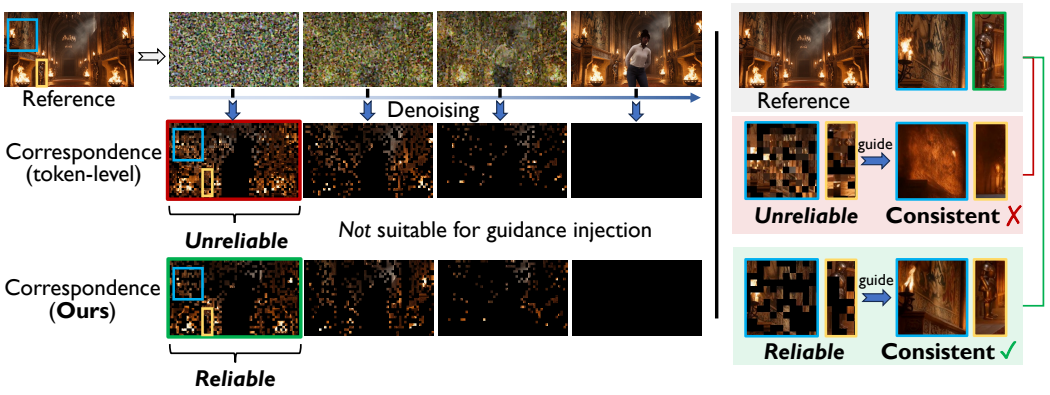

Figure 2: **Early denoising video frames contain implicit geometric correspondences with the reference image.** *Left:* visualization of the correspondence map during the denoising stage. *Right:* our ReCoG enables consistent synthesis with reliable geometric guidance.

Our key insight is that the missing reference-to-video correspondence is not entirely absent from R2V models: it emerges internally during the diffusion process. By inspecting the video-to-reference attention within an intermediate Transformer block, we find that meaningful correspondences already appear at early denoising stages, as shown in Figure 2. This observation suggests a new route to explicit camera control in R2V, but exploiting these endogenous correspondences introduces three challenges. First, selecting the Transformer block from which correspondence is extracted requires balancing correspondence richness against the number of subsequent blocks available for guidance injection. Second, token-level correspondences in early-denoising are noisy and discontinuous, with erroneous matches easily crossing semantic boundaries, making them unreliable for direct geometric guidance. Third, these correspondences are produced by the model itself and therefore evolve during training, creating non-stationary geometric guidance that can destabilize training.

To address these challenges, we propose **ReCoG**, which **R**eifies **e**ndogenous reference-to-video **Co**rrespondences into explicit geometric **G**uidance. For the first challenge, we introduce **Anchor Block Selection (ABS)**, which analyzes the richness of video-to-reference correspondence across Transformer blocks in the pretrained model and selects a block that balances correspondence quality with sufficient subsequent blocks for guidance injection. For the second challenge, we observe that although early-denoising correspondences are noisy at the token level, they remain locally coherent at the region level. Based on this observation, we propose **Noise-Resilient Matching (NRM)** to aggregate token-wise matches into locally consistent regions and suppress noise-induced mismatches. The resulting high-confidence correspondences are materialized into a sparse anchor image and propagated along the prescribed camera trajectory through depth warping, producing an anchor video that explicitly guides the background geometry. For the third challenge, we introduce **Anchor Distribution Alignment (ADA)**, which constructs stable training anchors from ground-truth frames with stochastic masking to mimic the inference-time sparse-anchor distribution, enabling stable optimization and robust background synthesis from partial geometric guidance. Extensive experiments on both synthetic and real-world datasets demonstrate that ReCoG substantially improves camera-control accuracy and reference fidelity while preserving flexible adaptation to the foreground.

Our main contributions are summarized as follows:

- We study subject-aware camera-controllable video background generation and identify the lack of explicit reference-to-video correspondence as a key bottleneck for precise camera control in flexible R2V models. We further reveal that such correspondence emerges endogenously during early diffusion.
- We propose **ReCoG**, which reifies endogenous correspondence into explicit geometric guidance through correspondence-aware block selection, region-level noise-resilient matching and sparse anchoring, and distribution-aligned training for stable optimization.
- Extensive experiments show that ReCoG reduces translation/rotation errors over the strongest camera-control baseline by **28.4%/48.9%** on synthetic data and **15.9%/39.5%** on real-world data, while also achieving the best reference consistency among compared methods.

2 Related work

Reference-to-video generation. Reference-to-Video (R2V) generation aims to synthesize videos conditioned on one or more reference images, with prior works primarily focusing on transferring subject identity or scene appearance (Jiang et al., 2025; Liu et al., 2025; Fei et al., 2025; Hu et al., 2025a; Deng et al., 2025; Hu et al., 2025b; Li et al., 2025; Zhou et al., 2025; Zhang et al., 2025b). Its flexible reference conditioning makes R2V particularly suitable for subject-aware background generation, where the reference scene must adapt to a prescribed foreground rather than serve as a fixed video frame. ActAnywhere (Pan et al., 2024) first explores this setting by generating backgrounds conditioned on foreground subjects and reference scenes, while DynaScene (Yao et al., 2025) further incorporates camera trajectories through pose embeddings. However, these methods treat the reference primarily as an appearance condition and do not explicitly establish reference-to-video geometric correspondence. As a result, scene geometry and camera-induced motion must be inferred implicitly. In contrast, our method reveals that such correspondence emerges endogenously within R2V diffusion and reifies it into explicit geometric guidance for precise camera control.

Camera-control video generation. Camera control has been widely studied in video diffusion models. Early Text-to-Video (T2V) methods (He et al., 2024; Zhang et al., 2025a) typically encode camera trajectories with pose conditions such as Plücker embeddings, leaving the model to implicitly translate camera poses into scene motion. Such conditioning, however, does not explicitly constrain how scene structures evolve across frames. Recent Image-to-Video (I2V) (Ren et al., 2025; Li et al., 2026a) and Video-to-Video (V2V) (Kim et al., 2026; Chen et al., 2025; Yu et al., 2025; Bai et al., 2025a; Xie et al., 2026; Cao et al., 2026; Li et al., 2026a) methods instead exploit the predefined correspondence provided by the first frame or source video. By lifting visual content into 3D with estimated depth and reprojecting it along the target trajectory, they obtain explicit geometric guidance for more precise camera control. R2V lacks such correspondence by construction, as the reference image is not tied to any generated frame. Our method bridges this gap by recovering model-internal reference-to-video correspondence and reifying it into explicit geometric anchors, bringing correspondence-based camera control to flexible R2V generation.

3 METHOD

3.1 PRELIMINARY

Latent video diffusion models. A typical video diffusion model consists of a 3D Variational Autoencoder (VAE) $\{\mathcal{E}, \mathcal{D}\}$ and a denoising network f_θ. The encoder $\mathcal{E}$ compresses a normalized video Z into a spatiotemporal latent $z = \mathcal{E}(Z) \in \mathbb{R}^{T\times H\times W\times C}$. We adopt a rectified flow formulation to construct the noisy latent at timestep $t \in [0, \mathcal{T}]$:

$$z_t = (1 - t/\mathcal{T})\, z + t\epsilon/\mathcal{T}, \tag{1}$$

where $\mathcal{T}$ denotes the total timesteps and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ is Gaussian noise. The network f_θ is trained to predict the flow velocity by minimizing:

$$\mathcal{L} = \mathbb{E}_{z,\epsilon,t}\left[\|f_\theta(z_t, \tau, t) - (\epsilon - z)\|_2^2\right], \tag{2}$$

where τ denotes the text prompt. During inference, a clean latent z_0 is iteratively recovered from Gaussian noise $z_\mathcal{T}$, and mapped back to pixel space via the decoder $Z = \mathcal{D}(z_0)$.

Diffusion Transformer (DiT) architecture. In DiT-based video models, f_θ consists of N stacked Transformer blocks, each integrating spatiotemporal information and textual conditions through self-attention and cross-attention. In our R2V framework, the reference image latent z_{ri} and the video latent z_v are concatenated along the sequence dimension:$z = z_{ri} \oplus z_v \in \mathbb{R}^{(1+T)HW\times C}$. Consequently, the self-attention weights can be decomposed into:

$$\mathbf{W}_{\text{sa}} = \text{Softmax}\left(\tilde{Q}(z)\tilde{K}(z)^T/\sqrt{C}\right) = \begin{bmatrix} \mathbf{W}_{\text{r2r}} & \mathbf{W}_{\text{r2v}} \\ \mathbf{W}_{\text{v2r}} & \mathbf{W}_{\text{v2v}}, \end{bmatrix} \tag{3}$$

where $\tilde{(\cdot)}$ denotes Rotary Positional Embedding (RoPE) (Su et al., 2024) to encode spatiotemporal positions, $\mathbf{W}_{\text{r2r}} \in [0,1]^{HW\times HW}$ captures intra-reference relationships and $\mathbf{W}_v \in [0,1]^{THW\times(1+T)HW}$ (combining $\mathbf{W}_{\text{v2r}}$ and $\mathbf{W}_{\text{v2v}}$) characterizes how video tokens attend to both the reference image and themselves. This inherent interaction serves as the foundation for our method.

3.2 OVERVIEW

Given a foreground subject video Z_{fg} with its mask M_{fg}, a camera trajectory Φ, a textual description τ, and a background reference image I_{bg}, our goal is to synthesize a holistic video Z that preserves the foreground content in Z_{fg} and the background appearance of I_{bg} while faithfully following the prescribed camera motion. Figure 3 illustrates the ReCoG pipeline. Our core idea is to extract reliable endogenous video-to-reference correspondences from a pretrained R2V model and reify them into explicit geometric guidance. This involves three challenges: selecting a Transformer block that provides rich correspondence while retaining sufficient subsequent blocks for guidance injection, recovering reliable correspondence from noisy early-denoising attention, and avoiding unstable training caused by correspondence that evolves with the model.

To address these challenges, we first systematically analyze video-to-reference attentions across Transformer blocks to identify an Anchor Block, dividing the backbone into Pre-anchor blocks, the Anchor Block, and Post-anchor blocks (Section 3.3). At the onset of inference, our Sparse Anchoring Strategy (SAS) extracts video-to-reference correspondences from the Anchor Block and materializes them into a sparse anchor image, where Noise-Resilient Matching (NRM) aggregates noisy token-wise matches at the region level to recover reliable correspondences (Section 3.4). For training, Anchor Distribution Alignment (ADA) constructs a stable GT-derived sparse anchor image that mimics the inference-time anchor distribution, avoiding non-stationary endogenous guidance (Section 3.5). Finally, Sparse Anchor Injection propagates this image along Φ through depth warping and injects the resulting anchor video into the Post-anchor blocks as geometric guidance (Section 3.6).

3.3 ANCHOR BLOCK SELECTION

At the onset of denoising, extracting correspondence between the reference image and generated video is non-trivial, since the video latent is still heavily dominated by Gaussian noise. Nevertheless, we observe that meaningful reference-to-video associations already emerge within the DiT backbone

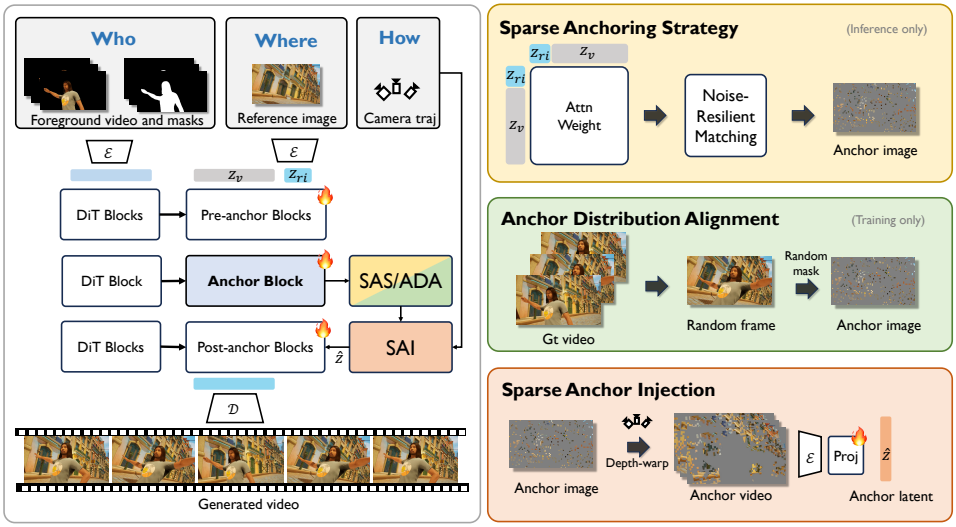

Figure 3: Pipeline overview. ReCoG extracts endogenous video-to-reference correspondence from a selected Anchor Block, constructs reliable sparse anchors through SAS with NRM, and uses ADA to provide stable training anchors. SAI then propagates the anchor along the target camera trajectory and injects the resulting anchor latent into Post-anchor blocks as explicit geometric guidance.

before visually clean video content is formed, as shown in Figure 2. Since reference and video tokens interact directly through joint self-attention, the video-to-reference attention provides a natural signal for probing such endogenous correspondence.

Exploiting this signal requires determining from which Transformer block the correspondence should be extracted. The selection of the Anchor Block is a strategic decision balancing **correspondence richness** and **geometric guidance capacity**: the selected block should exhibit strong video-to-reference interaction while leaving sufficient subsequent blocks for guidance injection. For the l-th Transformer block, we denote its video-token attention matrix as $\mathbf{W}_v^l$. For each video token i, we identify its dominant source token as:

$$P_i^l = \arg\max_j \mathbf{W}_v^l[i, j]. \tag{4}$$

Since the first HW tokens correspond to the reference image, $P_i^l \leq HW$ indicates that video token i is reference-dominant. We quantify the reference interaction block l by

$$\rho^l = \frac{1}{THW} \sum_{i=1}^{THW} \mathbb{I}\left[P_i^l \leq HW\right], \tag{5}$$

where $\mathbb{I}[\cdot]$ denotes the indicator function. We evaluate ρ^l across the pretrained backbone and select the block with the strongest reference interaction as the **Anchor Block**. Empirically, the selected block lies near the middle of the backbone, leaving sufficient subsequent blocks for geometric guidance injection. Accordingly, we divide the backbone into *Pre-anchor blocks* for initial feature interaction, the *Anchor Block* for correspondence extraction, and *Post-anchor blocks* for geometric guidance injection. The effect of this selection is analyzed in Section 4.3.

3.4 Sparse Anchoring Strategy

Identifying the Anchor Block reveals where endogenous correspondence emerges, but these early-denoising matches remain noisy and fragmented at the token level. Directly converting them into geometric guidance would introduce erroneous structures. We therefore propose **Sparse Anchoring Strategy (SAS)**, which extracts candidate video-to-reference matches, refines them through Noise-Resilient Matching (NRM), and materializes the reliable matches into a sparse geometric anchor. We omit the block index in Equation (4) and denote the dominant source token of video token i as P_i. We retain video tokens with $P_i \leq HW$, whose strongest response originates from the reference image, as initial candidate video-to-reference matches.

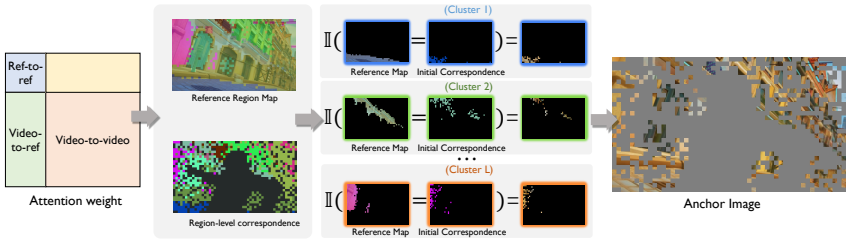

Figure 4: Noise-resilient matching (NRM) groups strongly associated reference tokens into regions and maps noisy token-level video-to-reference matches to region-level correspondence. The frame with the strongest region-preserving alignment is selected to construct the sparse anchor image.

Noise-Resilient Matching (NRM). Our key observation is that, although early-denoising correspondences are noisy at the token level, they still exhibit coherent patterns at the region level. Intuitively, a video token may already recognize that it should correspond to a building in the reference image, yet noise prevents it from locating the exact reference token and instead leads it to another feature-similar token within the same building region. This motivates us to reason about correspondence at the region level rather than requiring exact token-level matches. To this end, we partition the reference image into coherent regions. For efficiency, we exploit the reference-to-reference attention $\mathbf{W}_{\mathrm{r2r}}$, which measures pairwise affinities between reference tokens, with larger values indicating stronger associations. We cluster reference tokens based on these affinities, grouping strongly related tokens into the same region and obtaining a Reference Region Map $S \in \{1, \ldots, L\}^{HW}$, where L denotes the number of regions. Each token-level match P_i can then be mapped to its corresponding reference region S_{P_i}. For each video frame, we compare the matched region S_{P_i} with the region at the same spatial location in the reference, $S_{i \bmod HW}$. We retain only matches that originate from the reference image and remain within the same region:

$$R_i = \mathbb{I}\left[P_i \leq HW \wedge S_{P_i} = S_{i \bmod HW}\right]. \tag{6}$$

This region-level criterion tolerates local token fluctuations caused by noise while rejecting cross-region mismatches. The frame with the largest number of valid matches is selected as the anchor frame. Intuitively, this is the frame in which the model most strongly preserves the spatial organization of the reference image, and thus provides the most reliable basis for direct reference alignment. We then apply the retained correspondences to both the reference image I_{bg} and its depth map estimated by MoGe2 (Wang et al., 2025), rearranging the matched reference patches to construct the sparse anchor image $\hat{I}$ and anchor depth $\hat{D}$. Together with the camera pose $\hat{\Phi}$ associated with the anchor frame, $(\hat{I}, \hat{D}, \hat{\Phi})$ forms the sparse geometric anchor used for subsequent geometric guidance.

3.5 Anchor Distribution Alignment

SAS enables reliable sparse anchors to be recovered from endogenous correspondence at inference time. Directly applying the same procedure during training, however, would introduce a non-stationary conditioning distribution, since the extracted correspondence evolves together with the model parameters. We therefore propose **Anchor Distribution Alignment (ADA)**, which replaces model-dependent endogenous anchors during training with stable GT-derived sparse anchors that approximate the inference-time anchor distribution. TrajectoryCrafter (Yu et al., 2025) employs dual reprojection, which incurs additional computation and produces dense geometric guidance inconsistent with our sparse anchors. Instead, we randomly sample a frame from the GT video, remove its foreground region, and apply stochastic masking to construct a sparse training anchor. Using its corresponding depth and camera pose, we reproject this single sparse anchor along the camera trajectory. Because we sparsify before reprojection, the resulting missing regions remain temporally correlated and geometrically consistent across frames, closely approximating the sparse guidance SAS produces at inference. Notably, the GT frame used to construct the sparse anchor is sampled independently from the reference frame used for R2V conditioning. The model therefore cannot simply copy the reference appearance, but instead learns to synthesize a coherent background from partial geometric cues. In this way, ADA provides stable geometric conditioning during training while preserving the sparse and incomplete characteristics of inference-time anchors.

Table 1: Quantitative comparison with baselines. **Bold** means the best. "Syn." and "Real." refers to MultiCamVideo (Bai et al., 2025a) and DL3DV (Ling et al., 2024) dataset, respectively. FID, FVD, LPIPS, and SSIM are omitted for "Real." as ground-truth videos are unavailable.

		Camera Accuracy		Video Quality								Reference Consistency		Efficiency
	Method	TransErr↓	RotErr↓	FID↓	FVD↓	LPIPS↓	SSIM↑	SC↑	BC↑	TF↑	MS↑	DS-Max↑	DS-Avg↑	SF
Syn.	VACE+FT	0.2377	9.0644	210.27	1964.60	0.5721	0.3689	0.9165	0.9104	0.9537	0.9829	0.6810	0.6185	721M
	VACE+UCPE	0.1760	5.7711	210.79	1889.65	0.5683	0.3709	0.9293	0.9171	0.9599	0.9861	0.6694	0.6171	1.30G
	Ours	**0.1261**	**2.9499**	**169.48**	**1292.38**	**0.4884**	**0.4389**	**0.9372**	**0.9208**	**0.9706**	**0.9917**	**0.7303**	**0.6802**	784M
Real.	VACE+FT	0.2414	11.7765	-	-	-	-	0.9257	0.9030	0.9560	0.9856	0.6674	0.6225	721M
	VACE+UCPE	0.1558	5.5380	-	-	-	-	**0.9297**	0.9077	0.9594	0.9863	0.6622	0.6165	1.30G
	Ours	**0.1311**	**3.3483**	-	-	-	-	0.9290	**0.9111**	**0.9645**	**0.9890**	**0.7569**	**0.7038**	784M



Figure 5: Qualitative comparison with baselines on MultiCamVideo (Bai et al., 2025a) datasets.

3.6 Sparse Anchor Injection

Both SAS-derived inference anchors and ADA-generated training anchors are converted into trajectory-aware geometric guidance through **Sparse Anchor Injection (SAI)**. Given a sparse anchor $(\hat{I}, \hat{D}, \hat{\Phi})$, we propagate it along the target camera trajectory Φ using depth warpping (Yu et al., 2025), obtaining an anchor video $\hat{Z}$. This explicitly specifies how the recovered reference structures should evolve under the prescribed camera motion. The anchor video is encoded once as $\hat{z} = \mathcal{E}(\hat{Z})$, and cached throughout the denoising process. A lightweight linear layer projects $\hat{z}$ into the backbone feature space, after which it is added to the video hidden states at the input of each Post-anchor block. Through SAI, the recovered model-internal correspondence is converted into explicit, spatially aligned geometric guidance for camera-controlled background synthesis.

4 Experiments

4.1 Experiment Settings

Implementation details. Our framework is implemented based on the Wan2.1-VACE-1.3B model. We employ 1024-rank LoRA (Hu et al., 2022) exclusively on the self-attention layers of the backbone for efficient fine-tuning and a single linear layer for SAI. Optimization is performed using AdamW with a learning rate of 5×10^{-5} for 4,000 iterations. Training is conducted on two NVIDIA H200 GPUs with a total batch size of 32, utilizing 45-frame sequences at 480×832 resolution. End-to-end inference requires approximately 30 seconds per video using 20 denoising steps.

Dataset. We utilize a combination of synthetic and real-world datasets to train and evaluate our framework. Training is conducted on the MultiCamVideo (Bai et al., 2025a) synthetic training set, comprising 21,150 sequences. These sequences are preprocessed with Grounded-SAM (Ren et al.,

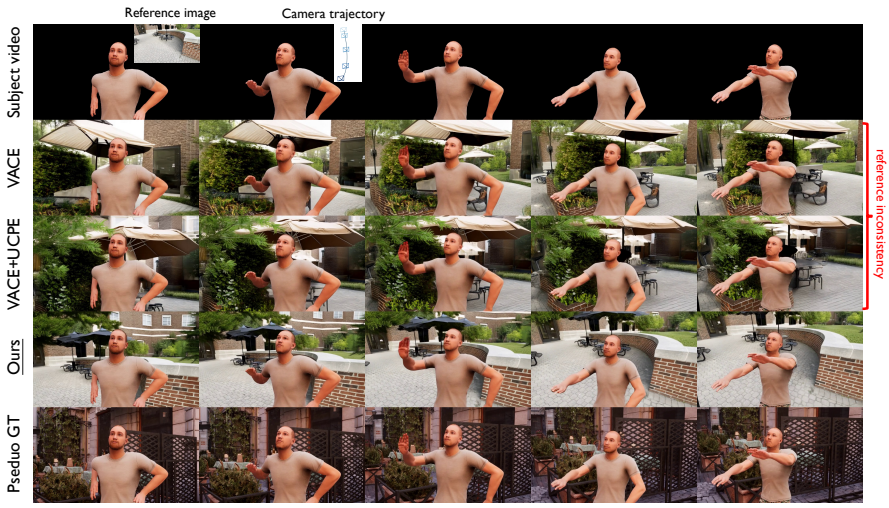

Figure 6: Qualitative comparison with baselines on DL3DV (Ling et al., 2024) datasets. The "Pseudo GT" serves exclusively as a visual reference for evaluating camera trajectory adherence.

2024) for subject segmentation and Qwen2.5-VL (Team, 2025) for captioning. For evaluation, we employ a two-pronged approach: first, we use 1,000 synthetic test sequences from MultiCamVideo, where ObjectClear (Zhao et al., 2025) is applied to synthesize clean background references from random frames. Second, to assess real-world generalization, we test on the DL3DV (Ling et al., 2024) dataset, rendering 10 camera trajectories for each of the 100 selected real-world images.

Evaluation metrics. We evaluate our framework across four dimensions: camera accuracy, video quality, reference consistency, and efficiency. To assess camera followability, we estimate trajectories via ViPE (Huang et al., 2025) and compute translation and rotation errors (TransErr and RotErr) (He et al., 2024; Duan et al., 2025). Video quality is measured using FID, FVD, LPIPS, and SSIM, and VBench (Huang et al., 2024) metrics, including Subject/Background Consistency (SC/BC), Temporal Flickering (TF), and Motion Smoothness (MS). Reference consistency is quantified via temporal DINO feature similarities (DS-Max and DS-Avg). Finally, we report the storage footprint (SF) of learnable parameters to demonstrate our lightweight fine-tuning paradigm.

4.2 Comparison with baselines

Given the absence of open-source implementations for ActAnywhere (Pan et al., 2024) and DynaScene (Yao et al., 2025), and the shift toward Diffusion Transformers (DiT), we establish a rigorous benchmark using DiT-based baselines: (1) VACE (Jiang et al., 2025) + Fine-tuning (FT), replicating ActAnywhere's implicitly decoupling paradigm, and (2) VACE + UCPE (Zhang et al., 2025a), mirroring DynaScene's use of implicit camera embeddings. This setup ensures a fair comparison between our explicit geometric guidance and prevailing implicit and appearance-driven strategies within a unified architecture. As shown in Figures 5 and 6 and Table 1, our method consistently outperforms baselines across both synthetic and real-world datasets. We achieve significant gains in camera followability while delivering superior video quality and reference consistency. Notably, our injection module uses only lightweight linear layers integrated into selected DiT blocks. This results in a 63MB footprint, which is far more efficient than methods requiring full-backbone injection.

4.3 Ablation studies

Impact of NRM. We further investigate the Noise-Resilient Matching (NRM) by comparing performance with and without its integration. As illustrated in Table 2 and Figure 7, omitting NRM results in chaotic anchor videos heavily corrupted by noise, which subsequently leads to distorted generations. While NRM provides marginal improvements in Temporal Flickering (TF) and Motion Smoothness (MS), these gains come at the high cost of severe loss in fine-grained details.

Table 2: Quantitative results of ablation studies. **Bold**: best results; underline: the second best results.

Setting			Camera Accuracy		Video Quality								Reference Consistency		Efficiency
NRM	ADA	$\hat{l}$	TransErr↓	RotErr↓	FID↓	FVD↓	LPIPS↓	SSIM↑	SC↑	BC↑	TF↑	MS↑	DS-Max↑	DS-Avg↑	SF
	✓	15	0.1370	3.3218	236.70	1926.10	0.5788	0.3841	0.9274	0.9180	**0.9741**	**0.9921**	0.6292	0.5838	784M
✓		15	0.1961	5.5981	198.61	1673.05	0.5597	0.3874	0.9354	0.9182	0.9637	0.9892	0.6785	0.6322	784M
✓	✓	10	0.1598	4.2475	196.01	1604.06	0.5456	0.3870	0.9295	0.9138	0.9636	0.9895	0.6829	0.6365	806M
✓	✓	18	0.1552	4.3125	190.96	1553.14	0.5177	0.4128	0.9330	0.9189	0.9691	0.9911	0.7011	0.6507	770M
✓	✓	15	**0.1261**	**2.9499**	**169.48**	**1292.38**	0.4884	0.4389	**0.9372**	**0.9208**	0.9706	0.9917	**0.7303**	**0.6802**	784M

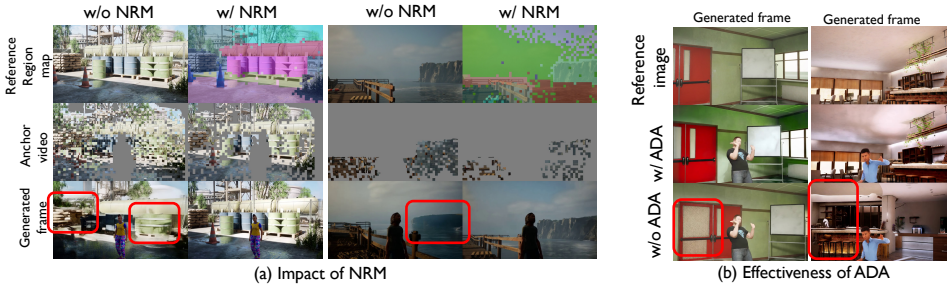

Figure 7: Ablation studies on NRM and ADA.

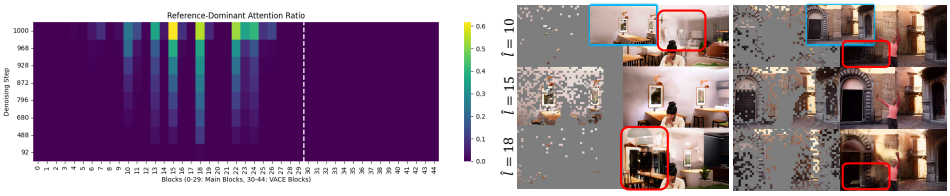

Figure 8: Ablation study on Anchor Block selection. Left: Ratio of reference-dominant video tokens across blocks, averaged over 10 samples. Right: Qualitative comparison with different anchor blocks.

Anchor Block selection. We evaluate the impact of Anchor Block selection in Table 2 and Figure 8, indicating that selecting a block with a relatively low response ratio (e.g., Block 10) fails to extract sufficient geometric anchors. Conversely, selecting a later block (e.g., Block 18) reduces the number of downstream blocks available for guidance injection, limiting control efficacy. Consequently, we select Block 15 as the optimal Anchor Block to balance geometric richness with injection capacity.

Effectiveness of ADA. To verify the efficacy of the Anchor Distribution Alignment (ADA), we conduct an ablation study by removing it and directly using the endogenous anchor videos extracted from the DiT as the injection condition during training. As shown in Table 2 and Figure 7, all results exhibit a noticeable decline. This suggests that our ADA module effectively simulates the anchor distribution encountered during inference, thereby stabilizing the fine-tuning process.

5 Conclusion

In this paper, we presented **ReCoG**, a framework for subject-aware camera-controllable video background generation that brings explicit geometric control to flexible R2V generation. Our key finding is that meaningful reference-to-video correspondence emerges endogenously during diffusion, despite being absent from the R2V formulation. Building on this insight, ReCoG identifies informative correspondence within the DiT backbone, distills noisy token-level matches into reliable sparse anchors, and stabilizes training through anchor distribution alignment; the recovered anchors are further propagated along the target camera trajectory as explicit geometric guidance. Extensive experiments on synthetic and real-world data demonstrate consistent improvements in camera control accuracy and reference consistency while preserving flexible adaptation to the subject. Our results highlight endogenous correspondence as a promising source of geometric control for reference-conditioned video generation.

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic datasets and assist with translation. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, clean and reformat datasets, support qualitative and thematic data analysis, and interpreting results, and providing critical ingredients for proving mathematical claims are not applicable to this work. Additionally, we used generative AI tools to edit a research paper to improve readability. We have reviewed all AI-assisted work. All AI-assisted content was carefully reviewed and revised by the authors to ensure correctness, consistency, and fidelity to the intended meaning. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, et al. World simulation with video foundation models for physical ai. *arXiv preprint arXiv:2511.00062*, 2025.

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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14834–14844, 2025a.

Jianhong Bai, Menghan Xia, Xintao Wang, Ziyang Yuan, Zuozhu Liu, Haoji Hu, Pengfei Wan, and Di Zhang. Syncammaster: Synchronizing multi-camera video generation from diverse viewpoints. In *International Conference on Learning Representations*, volume 2025, pp. 58038–58060, 2025b.

Wei Cao, Hao Zhang, Fengrui Tian, Yulun Wu, Yingying Li, Shenlong Wang, Ning Yu, and Yaoyao Liu. Freeorbit4d: Training-free arbitrary camera redirection for monocular videos via geometry-complete 4d reconstruction. *arXiv preprint arXiv:2601.18993*, 2026.

Yipeng Chen, Zhichao Ye, Zhenzhou Fang, Xinyu Chen, Xiaoyu Zhang, Jialing Liu, Nan Wang, Haomin Liu, and Guofeng Zhang. Postcam: Camera-controllable novel-view video generation with query-shared cross-attention. *arXiv preprint arXiv:2511.17185*, 2025.

Yufan Deng, Xun Guo, Yuanyang Yin, Jacob Zhiyuan Fang, Yiding Yang, Yizhi Wang, Shenghai Yuan, Angtian Wang, Bo Liu, Haibin Huang, et al. Magref: Masked guidance for any-reference video generation. *arXiv e-prints*, pp. arXiv–2505, 2025.

Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. Worldscore: A unified evaluation benchmark for world generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 27713–27724, 2025.

Zhengcong Fei, Debang Li, Di Qiu, Jiahua Wang, Yikun Dou, Rui Wang, Jingtao Xu, Mingyuan Fan, Guibin Chen, Yang Li, et al. Skyreels-a2: Compose anything in video diffusion transformers. *arXiv preprint arXiv:2504.02436*, 2025.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Teng Hu, Zhentao Yu, Zhengguang Zhou, Sen Liang, Yuan Zhou, Qin Lin, and Qinglin Lu. Hunyuancustom: A multimodal-driven architecture for customized video generation. *arXiv preprint arXiv:2505.04512*, 2025a.

Teng Hu, Zhentao Yu, Zhengguang Zhou, Jiangning Zhang, Yuan Zhou, Qinglin Lu, and Ran Yi. Polyvivid: Vivid multi-subject video generation with cross-modal interaction and enhancement. *arXiv preprint arXiv:2506.07848*, 2025b.

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.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17191–17202, 2025.

Min-Jung Kim, Jeongho Kim, Hoiyeong Jin, Junha Hyung, and Jaegul Choo. Infinite-homography as robust conditioning for camera-controlled video generation. In *European Conference on Computer Vision*, pp. 201–216. Springer, 2026.

Chunyang Li, Yuanbo Yang, Jiahao Shao, Hongyu Zhou, Katja Schwarz, and Yiyi Liao. Rerope: Repurposing rope for relative camera control. *arXiv preprint arXiv:2602.08068*, 2026a.

Yixuan Li, Yanhong Zeng, Ka Leong Cheng, Jiayi Zhu, Hanlin Wang, Wen Wang, Yihao Meng, Hao Ouyang, Qiuyu Wang, Yue Yu, et al. Cameraanything: Refilming videos with arbitrary camera control. *arXiv preprint arXiv:2607.24591*, 2026b.

Zhaoyang Li, Dongjun Qian, Kai Su, Qishuai Diao, Xiangyang Xia, Chang Liu, Wenfei Yang, Tianzhu Zhang, and Zehuan Yuan. Bindweave: Subject-consistent video generation via cross-modal integration. *arXiv preprint arXiv:2510.00438*, 2025.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22160–22169, 2024.

Haofeng Liu, Yang Zhou, Ziheng Wang, Zhengbo Xu, Zhan Peng, Jie Ma, Jun Liang, Shengfeng He, and Jing Li. Mocam: Unified novel view synthesis via structured denoising dynamics. In *European Conference on Computer Vision*, pp. 354–372. Springer, 2026.

Lijie Liu, Tianxiang Ma, Bingchuan Li, Zhuowei Chen, Jiawei Liu, Gen Li, Siyu Zhou, Qian He, and Xinglong Wu. Phantom: Subject-consistent video generation via cross-modal alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14951–14961, 2025.

Yawen Luo, Xiaoyu Shi, Jianhong Bai, Menghan Xia, Tianfan Xue, Xintao Wang, Pengfei Wan, Di Zhang, and Kun Gai. Camclonemaster: Enabling reference-based camera control for video generation. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–10, 2025.

Boxiao Pan, Zhan Xu, Chun-Hao P Huang, Krishna K Singh, Yang Zhou, Leonidas J Guibas, and Jimei Yang. Actanywhere: Subject-aware video background generation. *Advances in Neural Information Processing Systems*, 37:29754–29776, 2024.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6121–6132, 2025.

Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, et al. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026.

Vincent Sitzmann, Semon Rezchikov, Bill Freeman, Josh Tenenbaum, and Fredo Durand. Light field networks: Neural scene representations with single-evaluation rendering. *Advances in Neural Information Processing Systems*, 34:19313–19325, 2021.

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

Qwen Team. Qwen2.5-vl, January 2025. URL `https://qwenlm.github.io/blog/qwen2.5-vl/`.

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.

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.

Mingyang Xie, Numair Khan, Tianfu Wang, Naina Dhingra, Seonghyeon Nam, Haitao Yang, Zhuo Hui, Christopher Metzler, Andrea Vedaldi, Hamed Pirsiavash, et al. Lavr: Scene latent conditioned generative video trajectory re-rendering using large 4d reconstruction models. *arXiv preprint arXiv:2601.14674*, 2026.

Zhufeng Xu, Xuan Gao, Bailin Deng, Yikang Ding, Xiaoqiang Liu, Haoxian Zhang, Pengfei Wan, Hongbo Fu, and Lin Gao. Camerasquad: Achieving content consistency in parallel multi-trajectory camera-controlled video generation. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–12, 2026.

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.

Mingshuai Yao, Mengting Chen, Qinye Zhou, Yabo Zhang, Ming Liu, Xiaoming Li, Shaohui Liu, Chen Ju, Shuai Xiao, Qingwen Liu, et al. Beyond static scenes: Camera-controllable background generation for human motion. *arXiv preprint arXiv:2504.02004*, 2025.

Mark Yu, Wenbo Hu, Jinbo Xing, and Ying Shan. 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.

Cheng Zhang, Boying Li, Meng Wei, Yan-Pei Cao, Camilo Cruz Gambardella, Dinh Phung, and Jianfei Cai. Unified camera positional encoding for controlled video generation. *arXiv preprint arXiv:2512.07237*, 2025a.

Zhenxing Zhang, Jiayan Teng, Zhuoyi Yang, Tiankun Cao, Cheng Wang, Xiaotao Gu, Jie Tang, Dan Guo, and Meng Wang. Kaleido: Open-sourced multi-subject reference video generation model. *arXiv preprint arXiv:2510.18573*, 2025b.

Jixin Zhao, Shangchen Zhou, Zhouxia Wang, Peiqing Yang, and Chen Change Loy. Objectclear: Complete object removal via object-effect attention. *arXiv preprint arXiv:2505.22636*, 2025.

Zijian Zhou, Shikun Liu, Haozhe Liu, Haonan Qiu, Zhaochong An, Weiming Ren, Zhiheng Liu, Xiaoke Huang, Kam Woh Ng, Tian Xie, et al. Scaling zero-shot reference-to-video generation. *arXiv preprint arXiv:2512.06905*, 2025.

A TECHNICAL APPENDICES OVERVIEW

The appendices provide additional evidence and analyses for ReCoG. Section B presents results on real-world scenes, and Section C provides additional comparison with an I2V-based method. Section D studies the effect of regional granularity and robustness to depth estimation errors. Section E discusses limitations and future directions, while Section F provides additional qualitative results on synthetic and real-world datasets.

B RESULTS ON REAL-WORLD SCENES

Figure 9 demonstrates the generalization of ReCoG to real-world scenes. ReCoG is built upon a pretrained video foundation model trained on large-scale real-world videos and is adapted using lightweight LoRA modules, rather than trained from scratch on synthetic data. In our training, synthetic videos provide accurate geometry and camera supervision, while the pretrained backbone supplies the visual prior required for realistic video synthesis. Similar synthetic-supervision strategies have been widely adopted in camera-controllable video generation (Bai et al., 2025a; Luo et al., 2025; Li et al., 2026a; Liu et al., 2026; Kim et al., 2026; Li et al., 2026b; Bai et al., 2025b; Xu et al., 2026). Additional results are provided in Section F.

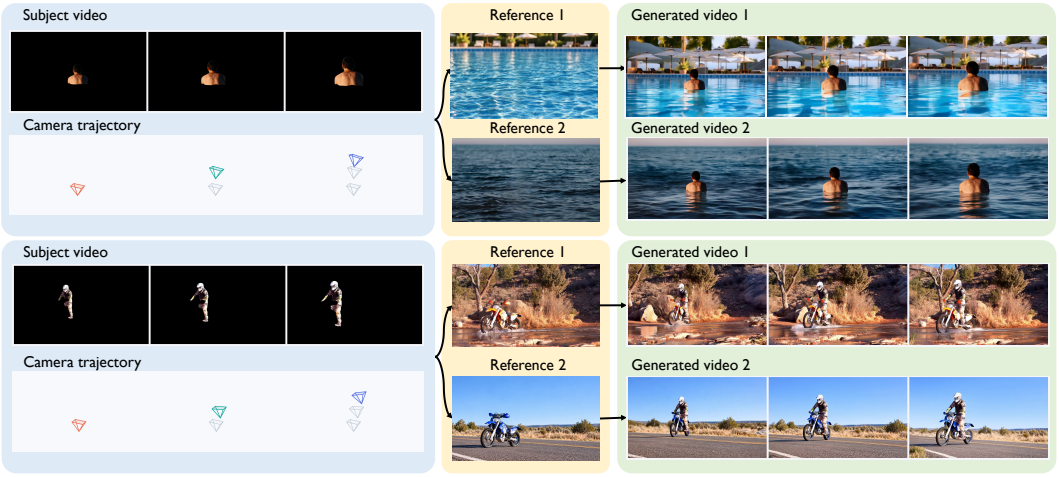

Figure 9: Results on real-world scenes.



Figure 10: Comparison with the I2V-based GEN3C (Ren et al., 2025).

C ADDITIONAL COMPARISON WITH AN I2V-BASED METHOD

We further compare ReCoG with GEN3C (Ren et al., 2025), a representative camera-controlled Image-to-Video (I2V) method, to examine the suitability of I2V conditioning for subject-aware background generation.

Given a reference image, GEN3C estimates its geometry and warps the image along the target camera trajectory to construct a condition video. This warped video provides explicit geometric guidance to a DiT that generates the final video from noise. To adapt GEN3C to our setting, we fuse its condition video with the prescribed foreground video so that the generated result preserves the target subject and motion.

As shown in Figure 10, this formulation has clear limitations for our task. Because reference content is directly propagated through geometric warping, structures that are incompatible with the prescribed

foreground may still be imposed on the generated video. This becomes particularly problematic when the reference viewpoint or layout is misaligned with the foreground, or when the reference image itself contains a subject. Moreover, subject-induced effects such as cast shadows are absent from the reference and therefore from the warped condition, making natural subject-background interaction difficult to synthesize.

In contrast, ReCoG retains the flexible conditioning of R2V generation. Rather than treating the reference as a fixed geometric template, it selectively inherits and adapts reference content to the foreground context, while using sparse correspondence-derived anchors for camera control. This enables more natural scene adaptation and subject–background interaction.

D Additional analysis

Table 3: Quantitative results. Darker blue denotes superior results.

Setting		Camera Accuracy		Video Quality								Reference Consistency	
noised $\hat{D}$	L	TransErr↓	RotErr↓	FID↓	FVD↓	LPIPS↓	SSIM↑	SC↑	BC↑	TF↑	MS↑	DS-Max↑	DS-Avg↑
	5	0.1222	3.0767	169.77	1365.64	0.4921	0.4329	0.9308	0.9197	0.9718	0.9922	0.7260	0.6729
	10	0.1261	2.9499	169.48	1292.38	0.4884	0.4389	0.9372	0.9208	0.9706	0.9917	0.7303	0.6802
	15	0.1245	2.7180	162.44	1275.55	0.4801	0.4430	0.9343	0.9199	0.9705	0.9919	0.7403	0.6886
	20	0.1289	3.7537	158.39	1222.89	0.4775	0.4471	0.9321	0.9196	0.9704	0.9919	0.7439	0.6914
✓	10	0.1242	2.8677	167.74	1297.29	0.4884	0.4412	0.9347	0.9203	0.9708	0.9918	0.7311	0.6793

Impact of regional granularity We further analyze the impact of varying clustering granularities L during inference in Table 3. Generally, a larger L leads to a more fine-grained regional partitioning, which inherently reduces the number of successfully matched anchors. Since these anchors serve as explicit guidance for camera control, a reduction in their density inevitably weakens the model's ability to strictly follow the prescribed camera trajectory. However, we observe an interesting trade-off: fewer anchors allow the model to better leverage its capability to hallucinate the global background from sparse cues, thereby enhancing reference consistency.

Robust to inaccurate depth estimation Explicit depth-warping methods are often susceptible to inaccuracies in depth estimation. To evaluate the resilience of our sparse anchors, we conduct a sensitivity analysis by introducing random perturbations to the estimated anchor depth $\hat{D}$, formulated as $\alpha\hat{D} + \beta$ (where $\alpha \in [0.5, 1.5]$ and $\beta \in [0, 1]$). As shown in Table 3, our camera accuracy remains remarkably stable despite these significant depth fluctuations. This suggests that our sparse representation effectively mitigates depth estimation errors, highlighting the model's robust capability to reconstruct geometrically consistent videos from sparse and potentially noisy geometric cues.

E Limitation and future work

We show failure cases of our method in Figure 11. We observe that extreme scale inconsistencies between the reference image and the subject video can lead to suboptimal outcomes. Although the model attempts to synthesize reasonable videos, the significant geometric discrepancy often results in perceptually inconsistent videos. Consequently, providing a reference image with a congruent scale and perspective relative to the subject video is essential for achieving high-fidelity background synthesis.

While our proposed method effectively enables camera control in the R2V setting, its performance inherently inherits the generative boundaries of the underlying base model. The fidelity and temporal stability of the synthesized backgrounds are, to some extent, bounded by the pre-trained model's capacity. To further push these limits, we plan to scale up our framework by integrating it with more powerful large-scale architectures, such as Wan2.1-VACE-14B. Additionally, expanding the scale and diversity of the training data will be a key focus in our future work to further enhance the model's robustness and generative quality in complex real-world scenarios.

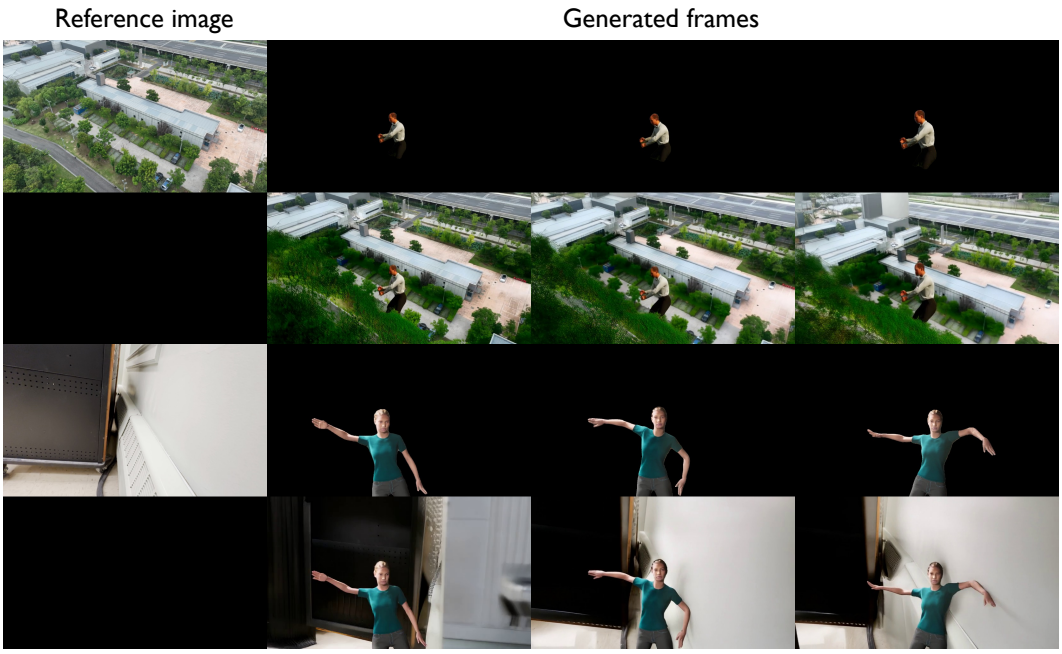

Figure 11: Failure cases.

F ADDITIONAL RESULTS

We provide additional comparisons on the MultiCamVideo and DL3DV datasets in Figures 12 to 14, together with further real-world results in Figure 15.

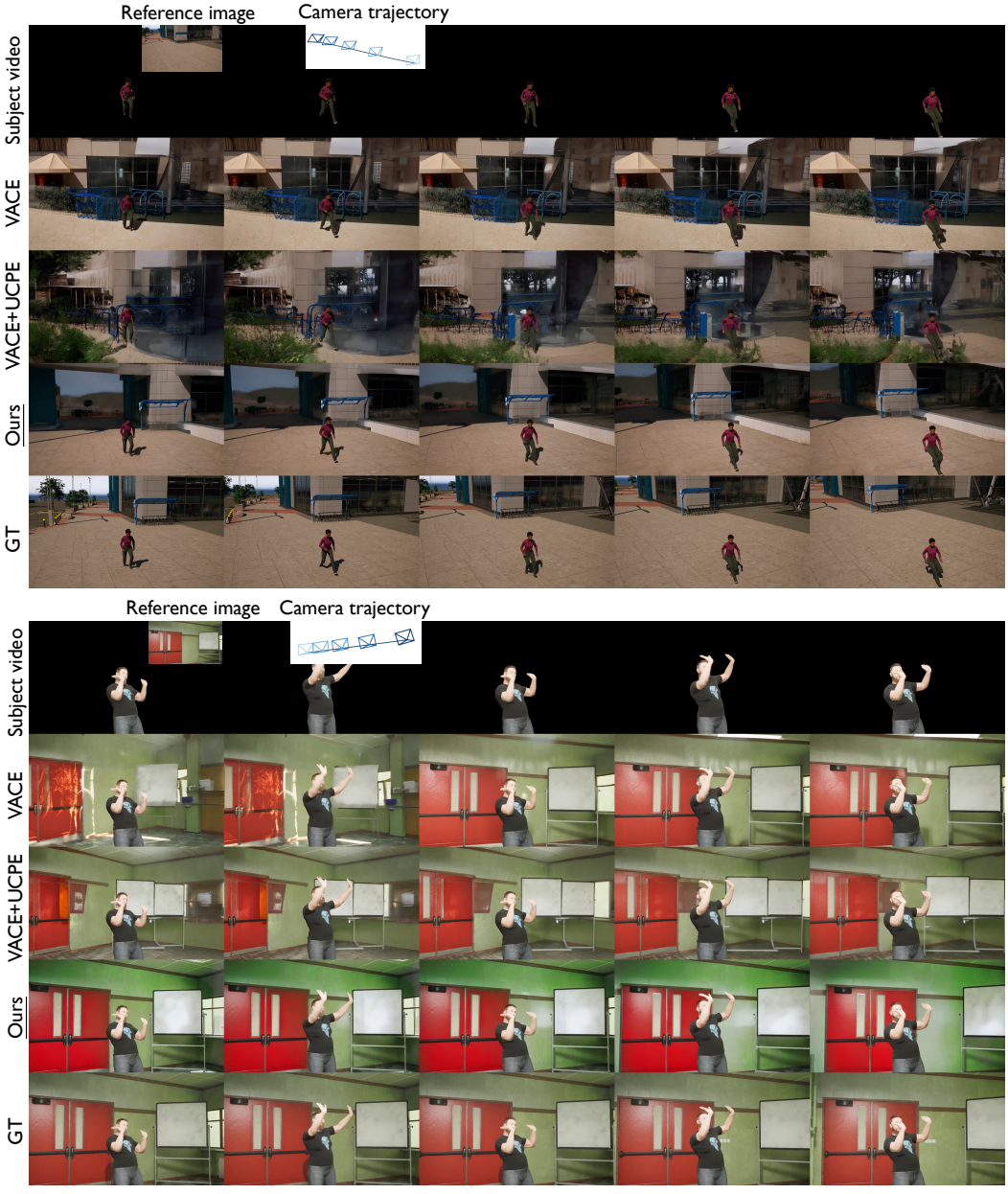

Figure 12: More comparisons with baselines on MultiCamVideo Dataset.

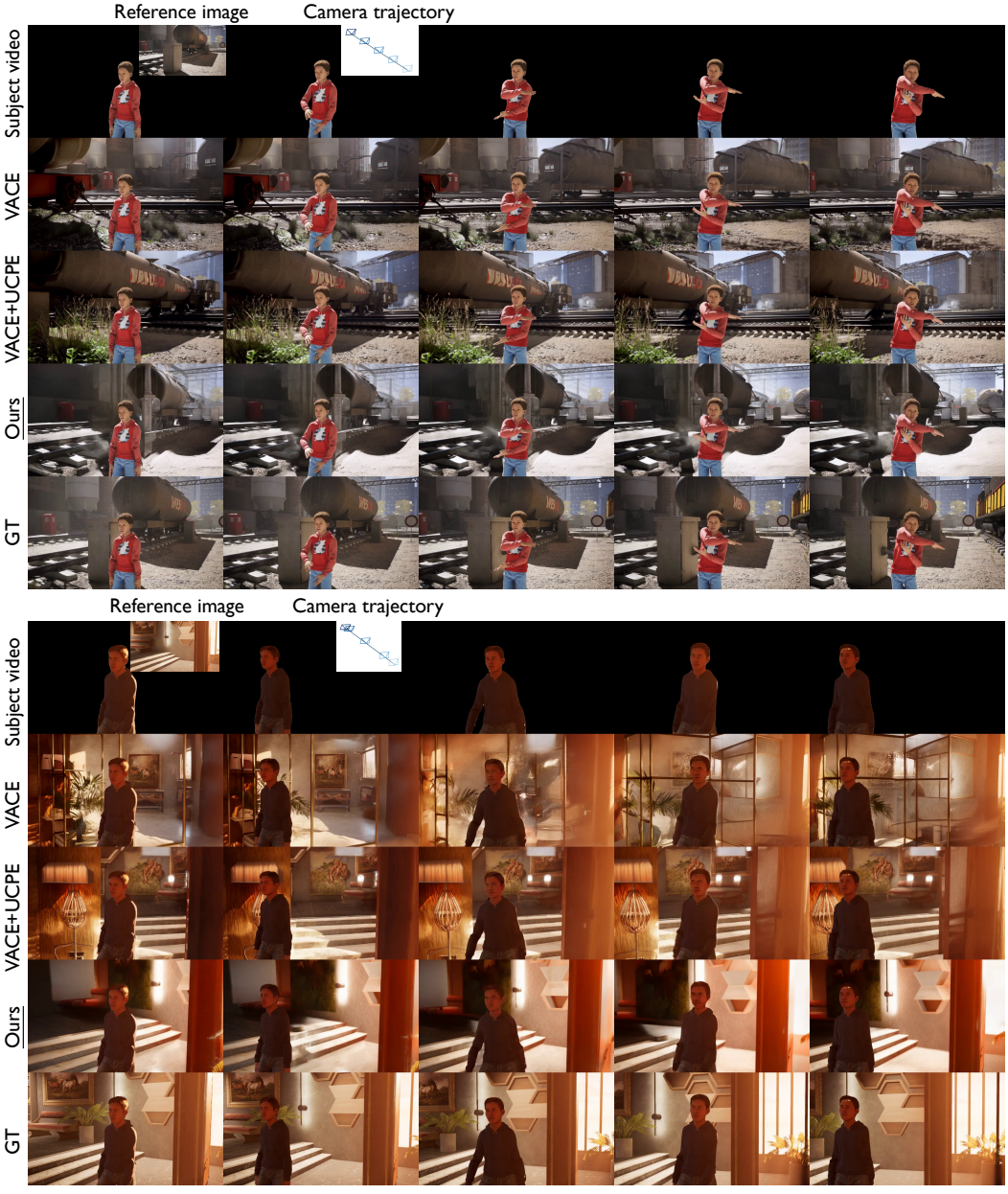

Figure 13: More comparisons with baselines on MultiCamVideo Dataset.

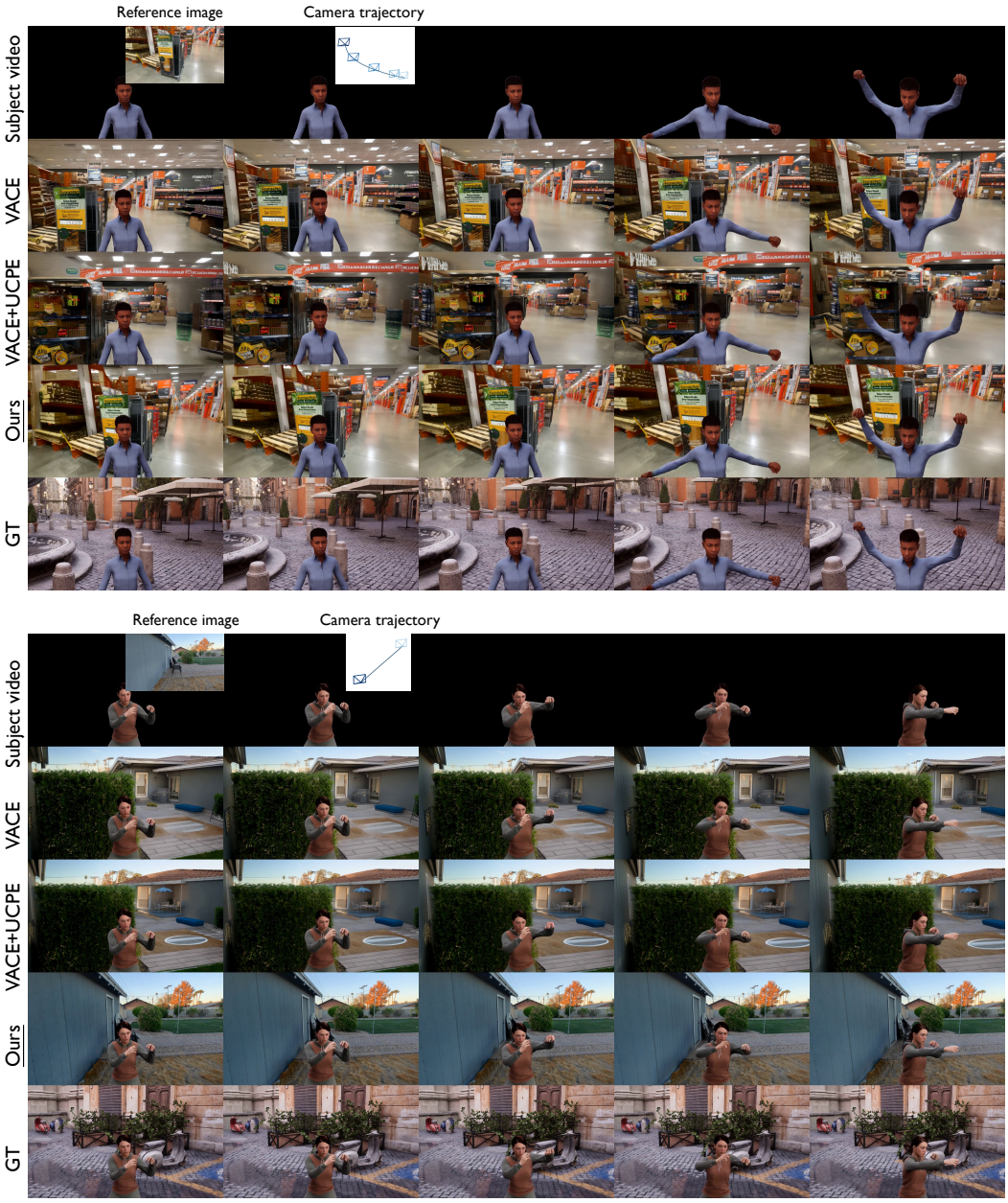

Figure 14: More comparisons with baselines on DL3DV Dataset.

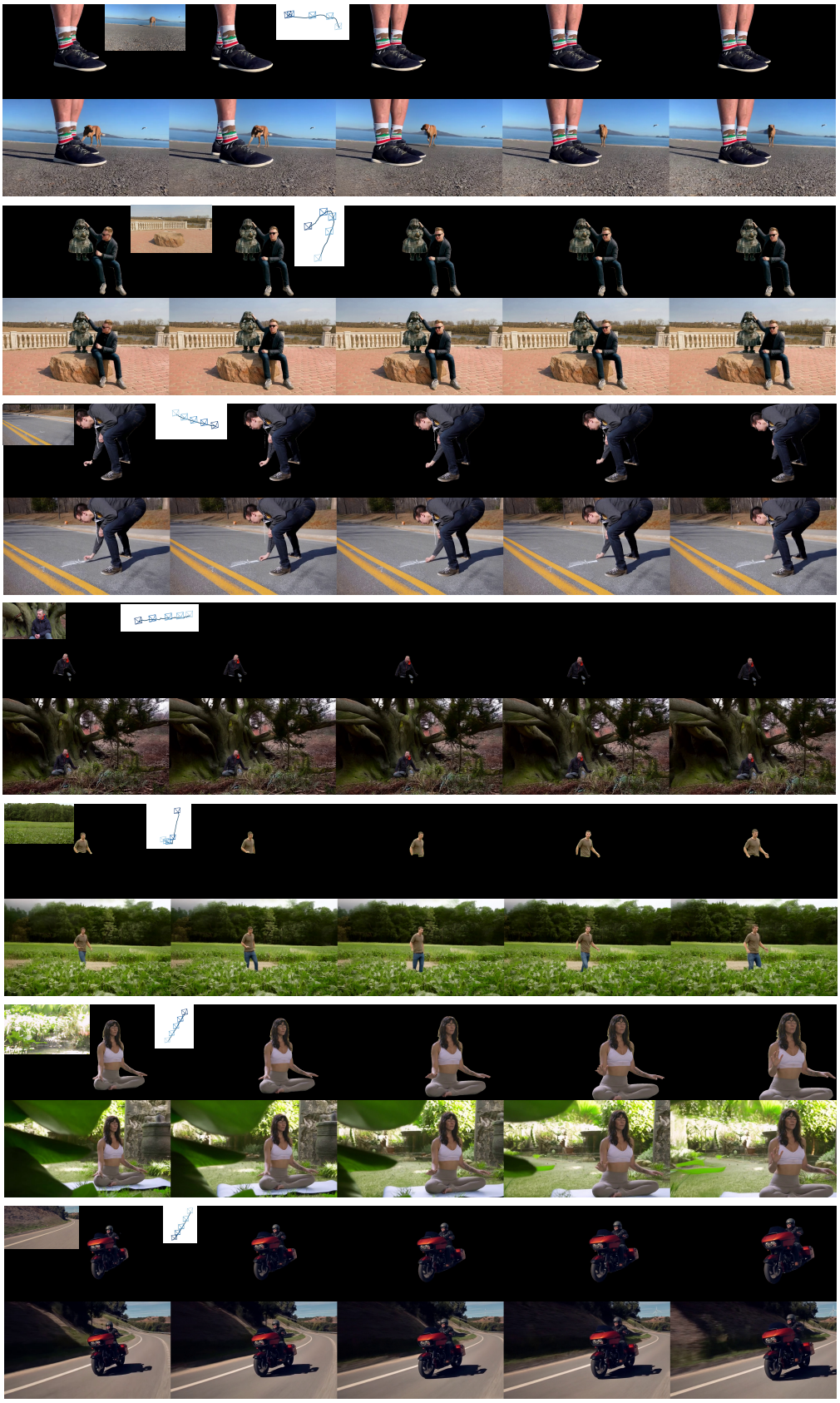

Figure 15: More results for real-world scenes.