MoCa: Modeling Object Consistency for 3D Camera Control in Video Generation

Anonymous authors
Paper under double-blind review

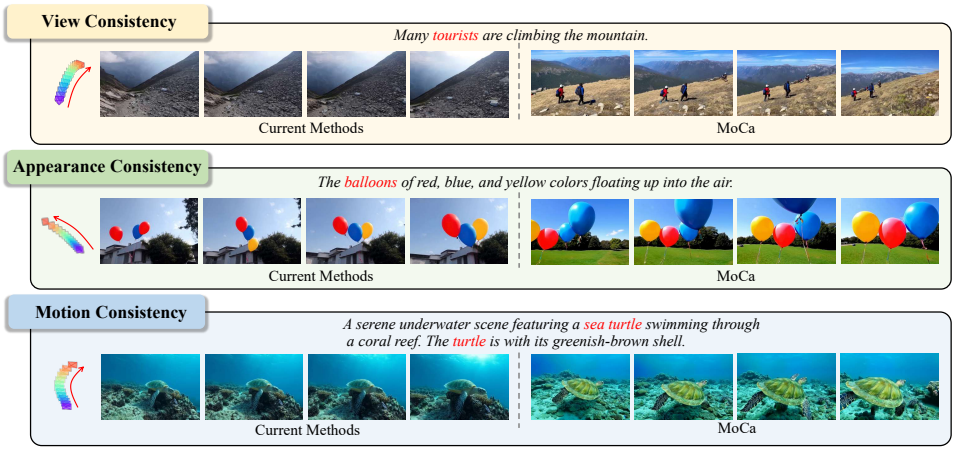

Figure 1: This figure outlines the core requirements for high-quality camera-controllable video generation: consistent object view, appearance, and motion. The foreground object should remain fully visible and retain its structure during camera movement, with stable texture and natural motion. Existing methods often fail to satisfy all three aspects simultaneously, whereas our approach demonstrates strong performance across all criteria.

Abstract

Camera control is important in text-to-video generation for achieving realistic scene navigation and view synthesis. This control is defined by parameters that describe movement through 3D space, thereby introducing a 3D consistency into the generation process. A core challenge for existing methods is achieving 3D consistency within the 2D pixel domain. Strategies that directly integrate camera conditions into text-to-video models often produce artifacts, while those relying on explicit 3D supervision face generalization issues. Both limitations originate from the gap between the 2D pixel space and the underlying 3D world. The key insight is that the projection of a smooth 3D camera movement produces consistency in object view, appearance, and motion across 2D frames. Inspired by this insight, we propose MoCa, a dual-branch framework that bridges this gap by modeling object consistency to implicitly learn 3D relationships between camera and scene. To ensure view consistency, we design a Spatial-Temporal Camera Encoder with Plücker embedding, which encodes camera trajectories into a geometrically grounded latent representation. For appearance consistency, we introduce a semantic guidance strategy that leverages persistent vision-language features to maintain object identity and texture across frames. To address motion consistency, we propose an object-aware motion disentanglement mechanism that separates object dynamics from global camera movement, ensuring precise camera control and natural object motion. Experiments show that MoCa achieves accurate camera control while preserving video quality, offering a practical and effective solution for camera-controllable video synthesis.

1 INTRODUCTION

Recent years have witnessed the significant success of video generation models supported by the foundational diffusion model in content creation and movie production owing to their excellent multi-modal understanding and powerful generation capabilities, especially in text-conditioned generation (Blattmann et al., 2023; Chen et al., 2023; He et al., 2022; Singer et al., 2023; Zhou et al., 2022). The growing demand for precise camera control to enhance video realism in applications like scene navigation and novel view synthesis remains unmet, as it requires models to understand the spatial relationship between camera trajectories and 3D scene.

While a standard text-to-video model learns a mapping $f(\mathbf{P}) \rightarrow \mathbf{V}^{X\times Y\times T}$, where X, Y denote pixel coordinates and T represents the temporal dimension, its objective is to ensure alignment between the text prompt $\mathbf{P}$ and the video volume $\mathbf{V}$. Camera-controlled generation introduces a specific trajectory condition $\mathbf{C}$. This requires the model to understand the spatial relationships of objects from changing viewpoints, learning a more complex mapping $f(\mathbf{P}, \mathbf{C}) \rightarrow \mathbf{V}^{X\times Y\times Z\times T}$. Here, the Z-dimension represents 3D spatial relation brought by the camera movement that must be consistently maintained. The challenge of learning this implicit 3D relation has led to different strategic approaches.

A common strategy is to treat camera control as an additional condition with standard 2D video generation frameworks. Methods such as MotionCtrl and CameraCtrl (Wang et al., 2024b; He et al., 2024) integrate camera parameters by temporal attention or element-wise addition in the Denoising U-Net. Without 3D spatial awareness, these approaches often struggle to keep view and motion consistency, leading to artifacts like unnatural object dynamics. Another line of work seeks to explicitly learn 3D relationship between the camera and objects with additional supervise. Techniques like VidCRAFT3 (Zheng et al., 2025) and ViewCrafter (Yu et al., 2024) convert video frames into 3D point clouds, while I2VControl-Camera (Feng et al., 2024) leverages RGB-D representations for better visual quality. These methods often rely on on accurate 3D estimation, which limits their practicality and generalizability. Therefore, the gap between the 3D scene and 2D video pixels limits current methods in addressing the challenges of camera-controlled video generation.

The key insight for building an implicit bridge between 2D pixels and the 3D scene is recognizing that a camera-controlled video, as a 2D projection of a 3D scene, will demonstrate consistent object view, appearance, and motion across frames as a result of smooth camera movement. Accordingly, our method focuses on modeling three types of consistency to produce natural object coherence under camera control: View, Appearance, and Motion. **View Consistency:** The camera's position and orientation determine which objects appear in the frame. For text-to-video generation, objects described in the text prompt must be visible throughout the majority of the clip. **Appearance Consistency:** As the camera moves, the structure and texture of objects in the 2D projection should remain continuous and stable over time. **Motion Consistency:** When both the camera and objects move, the task becomes more difficult because the generated video reflects a combination of both motions. Camera movements change the field of view, primarily causing translation and scaling of static objects. Meanwhile, dynamic objects must not only follow the motion described in the text prompt, but also exhibit translation and scaling consistent with the camera's movement. Therefore, the three types of consistency, view, appearance, and motion that observed in a 2D video imply a stable camera moving through a 3D scene. We believe if objects are consistent, the camera is controlled. We follow this insight to design our method.

In this paper, we proposed a dual-branch fusion framework named MoCa focusing on accurate **M**odeling **o**bject consistency to enhance **Ca**mera-controlled video generation. To maintain view consistency, we adopt Plücker embedding and Spatial-Temporal Camera Encoder (ST-Encoder) which provides a geometrically interpretable representation that encodes camera trajectories at the pixel level in a latent space. To enhance appearance consistency, we propose a semantic guidance strategy that incorporates vision-language features from a pre-trained foundational model. The vision-language features serve as persistent global scene information to guide the fusion of camera-conditioned visual features, mitigating issues such as object distortion and texture collapse. As for motion consistency, we decouple video motion into camera movements and object dynamics. For precise pixel-level camera control, the ST-Encoder with Plücker embedding enables precise camera controllability. Meanwhile, plausible object motion is also crucial for high-quality video generation. To achieve this, we propose an object-aware motion disentanglement that separates object dynamics

from global camera movements. Specifically, we extract the implicit structure and region information of foreground objects from the pre-trained foundational model, which serves as an object-aware mask to guide the motion disentanglement. This mechanism allows the model to maintain natural object dynamics while achieving precise camera control.

The main contributions of our work are as follows:

- We propose a method that learns view, appearance, and motion consistency to bridge 3D camera movement and its corresponding changes in 2D frames. This implicit learning of the scene-camera relationship results in enhanced camera-controlled video generation.
- We design a dual-branch framework comprising ReferenceNet and DenoisingNet, integrated with semantic guidance strategy that injects ReferenceNet visual-language features to improve the appearance consistency of objects.
- We introduce an object-aware disentangling mechanism to separate object motion from camera movement, ensuring object motion is both faithful to the text prompt and consistent with the camera's movement.

2 Related Work

2.1 Text-to-Video Generation

Text-to-video (T2V) generation is a challenging task that requires both high-fidelity visual realism and cross-modality consistency (Guo et al., 2024; Brooks et al., 2023; Wu et al., 2024; Ma et al., 2025; Menapace et al., 2024). In the early stage, research in video generation primarily relied on Generative Adversarial Networks (GANs) or Variational Autoencoders (VAEs) (Saito et al., 2017; Skorokhodov et al., 2022; Tulyakov et al., 2018; Vondrick et al., 2016) . Despite the progress they made, the performance of these methods was still far from expectations. Recent attempts at text-to-video (T2V) generation mainly leverage diffusion models for their impressive capability and well-established open-sourced communities (Blattmann et al., 2023; Chen et al., 2023; Guo et al., 2024; He et al., 2022; Ho et al., 2022; Singer et al., 2023; Wang et al., 2023). As a pioneer in this field, some methods (Ho et al., 2022; He et al., 2022; Hong et al., 2022; Karras et al., 2023; Ruan et al., 2023) commonly employ video diffusion models (VDMs) that incorporated temporal convolutional and attention layers into the pre-trained image diffusion models. Follow-up works, VideoCrafter (Chen et al., 2023) and SVD (Blattmann et al., 2023) expand the application of video diffusion models to larger datasets, while TF-T2V (Wang et al., 2024a) directly learn from extensive text-free videos. Nonetheless, these methods encounter limitations in generating long videos, owing to the inherent constraints on capacity and scalability within the U-Net design. To overcome these constraints, DiT-based models (Brooks et al., 2024; Peebles & Xie, 2023) have emerged as a promising alternative, enabling direct generation of videos extending up to tens of seconds. Notably, Sora (Zhang et al., 2023) adopts a unified visual representation, supporting large-scale training and synthesis of high-resolution videos exceeding one minute.

2.2 Camera-Controlled Video Generation

As a pioneering work, MotionCtrl (Wang et al., 2024b) learns camera control by conditioning pre-trained video models with extrinsic matrices. Follow-up works further improve the conditioning mechanisms. CameraCtrl (He et al., 2024) represents cameras as Plucker coordinates, which allows more stable ray-based rendering and view-dependent modeling. While I2VControl-Camera (Feng et al., 2024) introduces point trajectory guidance for precise object-centric control. Building upon this, CamCo (Xu et al., 2024) integrates epipolar constraints into attention layers, and CamTrol (Hou et al., 2024) leverages explicit 3D point cloud representations. Another line of work controls camera motion without training additional parameters. However, these methods often rely on additional guidance like depth or segmentation masks. Notably, all of these approaches adopt U-Net-based architectures as their backbone. More recently, (Bahmani et al., 2024b;a; He et al., 2025) incorporates camera control into a video diffusion transformer architecture. Despite these advances, existing methods fail to generate dynamic content (e.g., object motion) under camera control without constructing specialized dynamic video datasets. Our work enhances dynamic content generation without the need for curating additional dynamic videos.

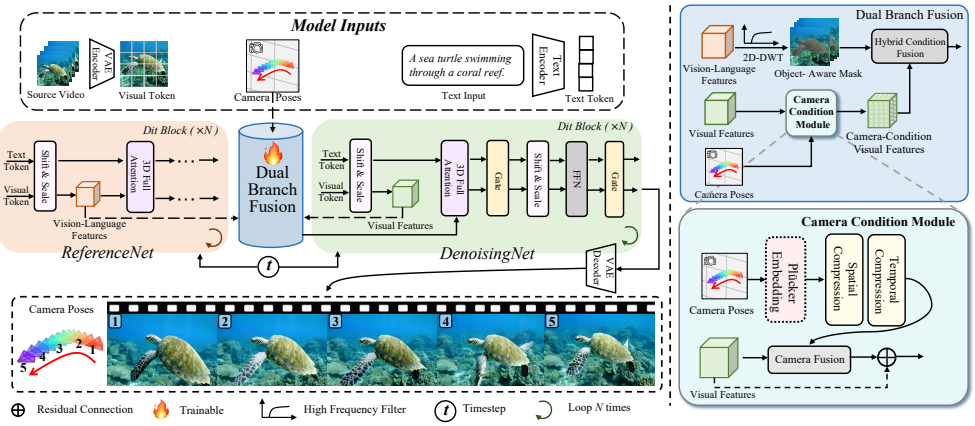

Figure 2: The overview of MoCa. To maintain view consistency, we utilize the Camera Condition Module with Plücker embedding to align camera rays with pixel-level visual representation. For appearance consistency, a semantic guidance strategy employs ReferenceNet's vision-language features to stabilize objects. Motion consistency is achieved by disentangling video motion into camera movement and object motion.

3 METHOD

Figure 2 illustrates the overall pipeline of MoCa. To model objects consistency of view, appearance and motion, we introduce a dual-branch fusion framework. For view consistency, we design a Spatial-Temporal Camera Encoder (ST-Encoder) with Plücker embedding as the primary form of camera parameters to align camera rays with pixel-level video representation. To enhance appearance consistency, we leverage vision-language alignment features from a foundational model and design a semantic guidance strategy to stabilize the appearance of objects. As for motion consistency, we propose an object-aware disentangling mechanism that guides the generation process to separate local object motion from camera movements.

3.1 CAMERA CONDITION MODULE

To ensure view consistency and prevent object misplacement, we introduce a camera condition module comprising a Camera Representation, Spatial-Temporal Camera Encoder and Camera Fusion Module. This system encodes camera trajectories into pixel-level representations and integrates them into the denoising process, guaranteeing that objects remain aligned with the camera's view.

Camera Representation. For the camera representation, we adopt Plücker embedding (Sitzmann et al., 2021) following recent works (He et al., 2024; Bahmani et al., 2024a), which provides strong geometric interpretation and fine-grained camera information. Specifically, given camera extrinsic matrix $\mathbf{E} = [\mathbf{R}; \mathbf{t}] \in \mathbb{R}^{3\times4}$ and intrinsic matrix $\mathbf{K} \in \mathbb{R}^{3\times3}$, we compute the Plücker embedding $\mathbf{p} = (\mathbf{o} \times \mathbf{d}', \mathbf{d}')$ for each pixel (u, v). Here, $\mathbf{o}$ represents the camera center in world coordinates, the ray direction from camera to pixel is defined as $\mathbf{d} = \mathbf{R}\mathbf{K}^{-1}[u, v, 1]^T + \mathbf{t}$, and $\mathbf{d}'$ is the normalized $\mathbf{d}$. The final Plücker embedding $\mathbf{P}_i \in \mathbb{R}^{6\times h\times w}$ is constructed for each frame, where h and w are the height and width for the frame.

Spatial-Temporal Camera Encoder. To integrate Plücker embedding into the generation process, we design a Spatial-Temporal Camera Encoder that transforms camera conditions into latent representations that are both spatially and temporally consistent with visual latents. In the spatial domain, a progressive convolutional architecture with downsampling, convolutional, and residual blocks extracts pixel-level spatial feature for camera motion. To incorporate temporal dynamics, we introduce dedicated 1D temporal convolutions across the frame sequence. The resulting spatial-temporal camera representations are fused with the visual features within the denoising process.

Camera Fusion Module. To effectively integrate camera representations with visual features, we adopted the fusion strategy from existing methods (Bahmani et al., 2024a; He et al., 2025) that injects control signals into each diffusion transformer (DiT) block. Specifically, we apply a cross-attention mechanism in each DiT block to fuse camera representations. This design allows the

model to dynamically modulate visual features based on spatial-temporal camera conditions, enabling fine-grained and precise controllability. By injecting geometry-aware camera embeddings into the generation process, the model maintains alignment between camera viewpoints and semantic objects, reducing cases where the described foreground fails to appear. Additionally, we explore alternative fusion strategies for camera signals which discussed in the ablation study.

3.2 Semantic Guidance Strategy

Appearance such as texture and scale of the object is critical in video generation, which directly affects the temporal consistency and visual integrity. Existing models suffer from object distortion or texture collapse in complex dynamic scenes, especially under intense camera movements. We trace this limitation to the additional conditioned camera signals, which weaken the base model's generative power. To address this, we propose a semantic guidance strategy that injects the visual-language features from a frozen foundational model into the generation process.

Specifically, we extract these vision-language features from the visual branch in each DiT block of the ReferenceNet, and inject them into each DiT block of the DenoisingNet. We consider that these features are aligned in both the visual and semantic spaces, serving as a stable appearance guidance for the whole scene. Therefore, our semantic guidance strategy reinforces vision-language alignment and enhancing the object appearance consistency of generated videos. Prior researches (Tian et al., 2024; Hu, 2024) have demonstrated the influence of utilizing analogous structures in maintaining the consistency of the object's identity. In our setup, the ReferenceNet shares the same structure with the DenoisingNet. Both the ReferenceNet and the DenoisingNet are initialized with weights inherited from the original pretrained DiT architecture.

3.3 Motion Disentanglement

In this section, we present an motion disentangling mechanism for separating object motion from camera movements, which is designed to enhance motion consistency. We decompose the overall video motion into camera movement and object motion. Given that camera control has been addressed in Section 3.1, this section focuses on our approach to modeling natural object motion. Specifically, we address the entanglement of object motion and camera motion by leveraging frequency-domain analysis. We utilize a 2D Discrete Wavelet Transform (2D-DWT) (Shahbahrami, 2012; Huang et al., 2005; Mushtaq et al., 2015) to extract high-frequency components from visual features, highlighting the structures and regions of objects. These high-frequency components guide the model to focus on natural object motion while improving precise camera controllability.

Current methods often fail to balance global camera movement and local object motion. When strong camera motion is applied, objects remain completely still, failing to show natural dynamics such as a fish swimming or a person walking. This limitation arises because video diffusion models entangle object and camera motion, making it difficult to maintain independent object dynamics.

High-Frequency Object-Aware Masking. To improve the realism of object motion, we propose an implicit object masking strategy that guides the model to separate local object motion from global camera movements. Specifically, we leverage the vision-language features from the foundational model to extract visual information that highlights foreground object structures and regions. Inspired by frequency-domain analysis in image processing, we apply a 2D Discrete Wavelet Transform (2D-DWT) (Shahbahrami, 2012; Huang et al., 2005; Mushtaq et al., 2015) to vision-language features across different orientations, capturing localized spatial-frequency information. This allows us to retain fine-grained structural cues that are critical for precise localization of object regions and effective motion disentanglement. The detail and visualization of 2D-DWT are discussed in appendix.

Hybrid Condition Fusion. We propose a hybrid conditioning fusion to fuse the object-aware mask with camera-conditioned visual features. It strategically employs cross-attention for spatial conditioning fusion with a temporal self-attention explicitly enforces inter-frame consistency. The cross-attention mechanism allows the model to dynamically modulate object-aware guidance and camera-conditioned features, enforcing accurate camera movements and preserving natural object motion. The temporal self-attention enables the model to maintain motion consistency across frames. Powered by the structure and localization information from the foundational model, our method enhances dual motion disentanglement and separates object motion from the camera movement.

Methods	FID ↓	FVD ↓	CLIPSIM ↑	TransErr ↓	RotErr ↓	OC ↑	BC ↑	MS ↑
RealEstate10K								
MotionCtrl(Animatediff-Based) (Wang et al., 2024b)	246.6	870.8	0.309	0.716	0.213	94.6%	95.8%	97.8%
CameraCtrl (He et al., 2024)	255.8	931.5	0.305	0.708	0.204	94.3%	94.7%	97.7%
AC3D (Bahmani et al., 2024a)	225.2	683.4	0.309	0.695	0.196	95.1%	95.3%	98.5%
Ours	207.4	667.9	0.312	0.703	0.208	94.9%	96.4%	98.5%
VidGen								
MotionCtrl(Animatediff-Based) (Wang et al., 2024b)	274.0	1858.2	0.333	0.722	0.107	92.6%	93.2%	97.1%
CameraCtrl (He et al., 2024)	266.3	1905.1	0.339	0.731	0.089	92.9%	93.1%	96.9%
AC3D (Bahmani et al., 2024a)	228.4	1712.0	0.345	0.727	0.084	93.5%	94.7%	97.7%
Ours	232.2	1643.7	0.349	0.724	0.081	94.7%	95.1%	98.3%

Table 1: Quantitative comparison on RealEstate10K and VidGen datasets. Lower is better (↓), higher is better (↑). Red indicates top-1 and blue indicates top-2 performance.

4 Experiments

4.1 Implementation Details

Datasets. During training, our model is fine-tuned from CogVideoX (Yang et al., 2024) on RealEstate10K (Zhou et al., 2018), which has around 65K video clips with per-frame camera parameters (extrinsics and intrinsics). This setup aligns with those used in prior works. For evaluation, we assess performance on both RealEstate10K and the VidGen dataset. VidGen (Tan et al., 2024) consists of a large collection of text-video pairs, primarily featuring dynamic scenes. In the contrast, RealEstate10K focus on static scenes, showcasing furnishings in indoor settings and natural landscapes in outdoor environments. We leverage this dataset to validate the effectiveness of our approach in object stability and motion consistency, thereby demonstrating its generalization to complex, dynamic scenes.

Metrics. We evaluate performance using a comprehensive set of quantitative metrics. For common evaluation, we report FID, FVD, and CLIPSIM scores. To evaluate camera accuracy, we follow CameraCtrl (He et al., 2024), using rotation and normalized translation errors from Mega-SAM (Li et al., 2024) reconstructed trajectories. To further assess foreground object consistency and background consistency, we adopt Object Consistency (OC), Background Consistency (BC) and Motion Smoothness(MS) scores from VBench (Huang et al., 2024), respectively. We consider OC, BC, and MS scores from VBench as standard and widely used evaluation metrics for text-to-video generation. Moreover, these metrics collectively capture the different aspects of consistency that our method aims to improve.

4.2 Quantitative Comparison

To evaluate the effectiveness of our method, we compare it with existing methods on Realestate10K and VidGen. For a fair comparison, all videos are uniformly downsampled to 16 frames. It demonstrates that our method achieves superior visual quality on RealEstate10K compared to existing methods, while maintaining competitive performance in camera controllability. These results validate our method's effectiveness for static scene camera controllability, particularly in achieving enhanced realism and stability beyond existing methods. Moreover, we conduct extensive experiments on the dynamic scene dataset VidGen. Our method outperforms previous methods across key metrics, such as RotErr, CLIPSIM, OC and MS scores. In addition to achieving superior performance, our method achieves suboptimal performance on metrics such as FID, FVD, and TransErr. Owing to the disentanglement of the object motion from the camera movement, our method achieves stable performance in motion smoothness(MS), even in the dynamic scene, unlike other methods that exhibit significant degradation on this metrics. Experiments show that our method achieves high quality generation with view, appearance and motion consistency.

4.3 Qualitative Comparison

We present qualitative comparisons in Figure 3, where the sequence from left to right represents the start to the end of the video. For the upper example, MotionCtrl fails to maintain view consistency, with no dog appearing in all frames. CameraCtrl effectively preserves view consistency but falls short in maintaining appearance consistency of the dog, showing obvious texture errors. AC3D is

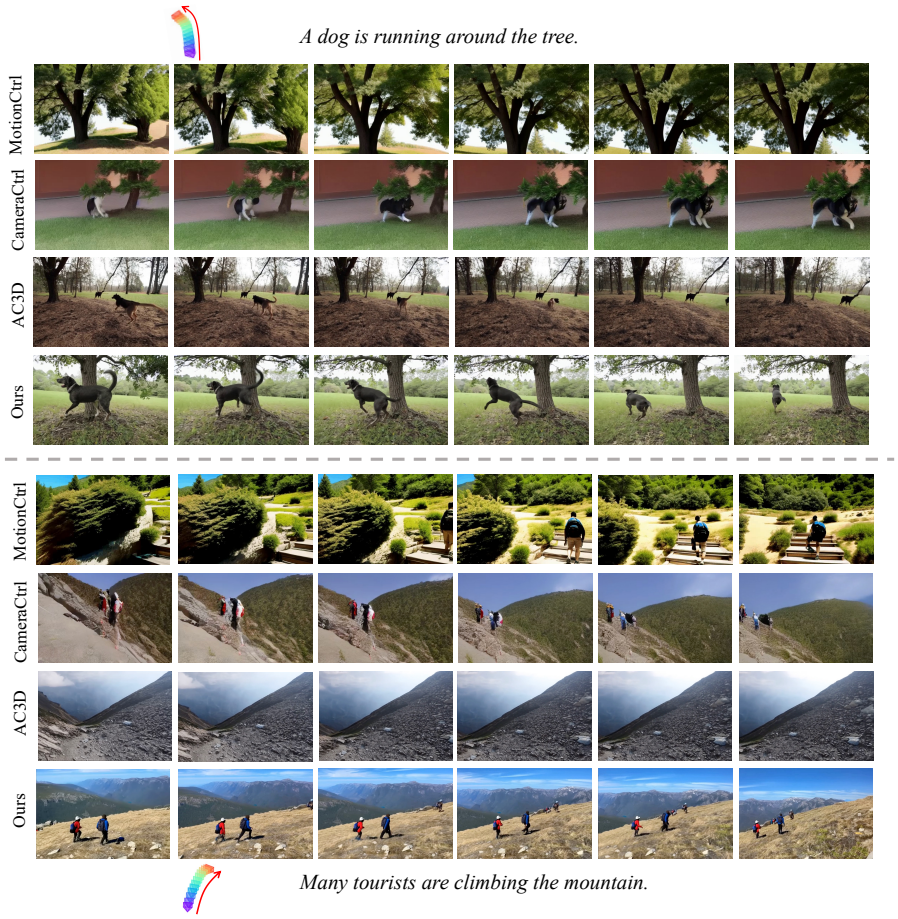

Figure 3: This figure presents a qualitative comparison between our method and existing approaches.

the best among the comparison methods overall but still exhibits unnatural motion. In contrast, our method achieves excellent consistency in view, appearance, and motion.

For the lower example, no tourists appear in all frames of AC3D, which clearly violates view consistency. CameraCtrl depicts the tourists' appearance very blurrily. The tourists' outlines in MotionCtrl are relatively accurate, but the motion continuity is poor, and they don't appear in the first few frames, violating view consistency. Our method performs much better overall. From the qualitative results, existing methods struggle to effectively maintain view, appearance, and motion consistency in the frames. The fundamental reason is still the lack of understanding of 3D space. Our method achieves this through constraints on objects.

4.4 Qualitative Results of MoCa Under Conflicting Motion

To evaluate the effectiveness of motion disentanglement, we further test MoCa in scenarios where the object motion direction described in the text conflicts with the camera input. For example, given the prompt "a bird flying from right to left" while the camera pans to the right, MoCa generates a bird moving against the camera motion correctly. As shown in Figure 5, MoCa ensures that the object motion (guided by text) are not overridden or distorted by the camera movements (guided by pose). This result indicates that our motion disentanglement mechanism effectively decouples object motion from camera movements.

To further validate this, we compare our model with AC3D under the same prompt and camera inputs. We observe that they often fail to realize the text-specified object motion direction, as their object motion remains entangled with the camera motion. This contrast further confirms the necessity and effectiveness of our motion disentanglement strategy.

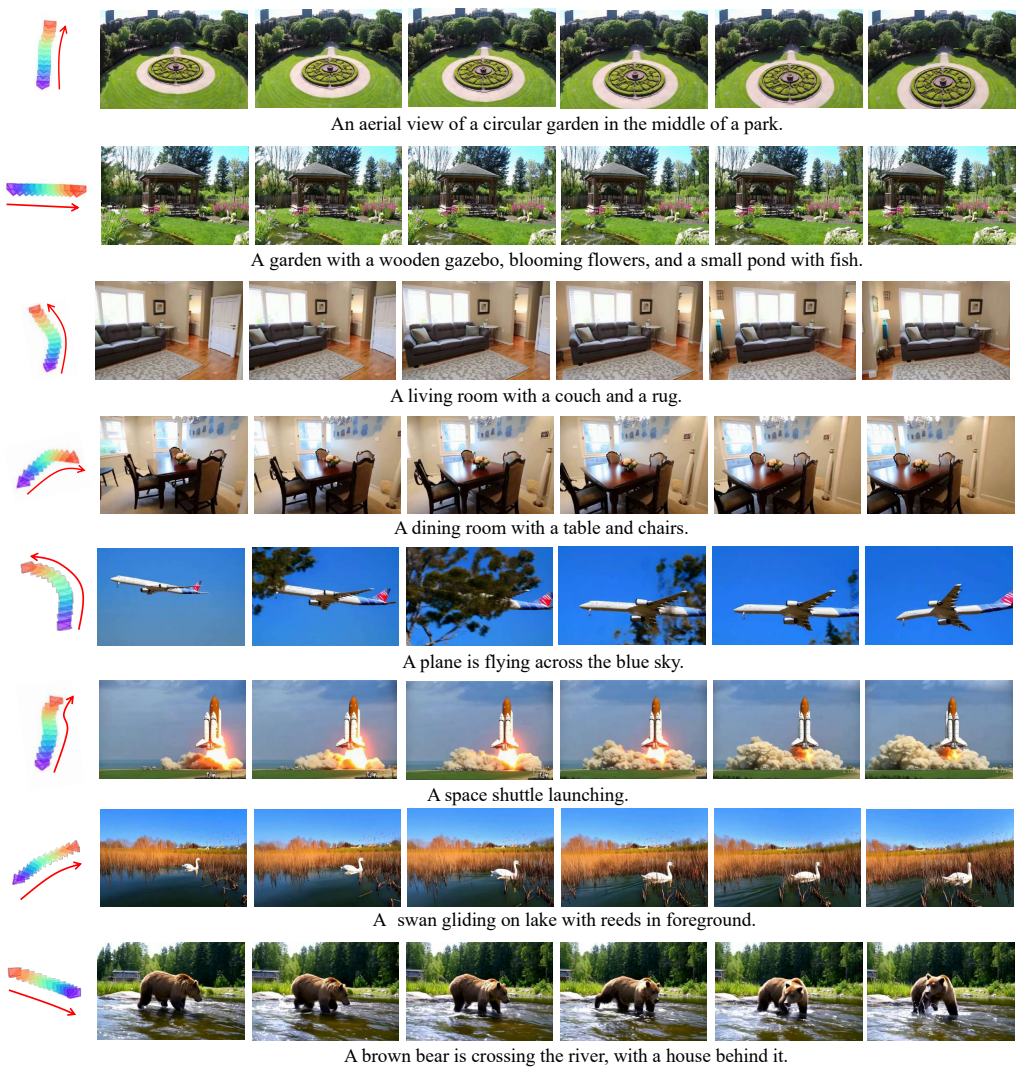

Figure 4: Qualitative results of our method in both static and dynamic scenes.

4.5 Qualitative Results of MoCa Across Diverse Scenarios

In this section, we present more results of our MoCa, especially on tasks of varying difficulty. There are eight examples from top to bottom, divided into 4 different types with increasing difficulty. The first group is outdoor scenes where most objects are in the distance and there are no complex spatial relationships. The second group is indoor scenes, where objects have clearer 3D relationships, such as a table in the center of the picture and a wall behind the table. Both groups of examples require the model to maintain good view and appearance consistency for a good visual effect. The following two groups of examples are more complex, describing the results of a dynamic object under camera movement. In the third group, airplanes and rockets move over a large area in the scene, and there are even complex occlusions. For instance, trees block the airplane, and smoke blocks the tail flame. Thanks to the understanding of 3D space, our method can handle these scenes well and achieve good view, appearance and motion consistency. For the fourth group of examples, the object are animals that have self-initiated movements, which requires the model to not only understand camera movement but also decouple the animals' own dynamics. Thanks to our motion decoupling mechanism, we can handle these situations well. In summary, our method achieve good performance in camera-controlled video generation especially in view, appearance and motion consistency.

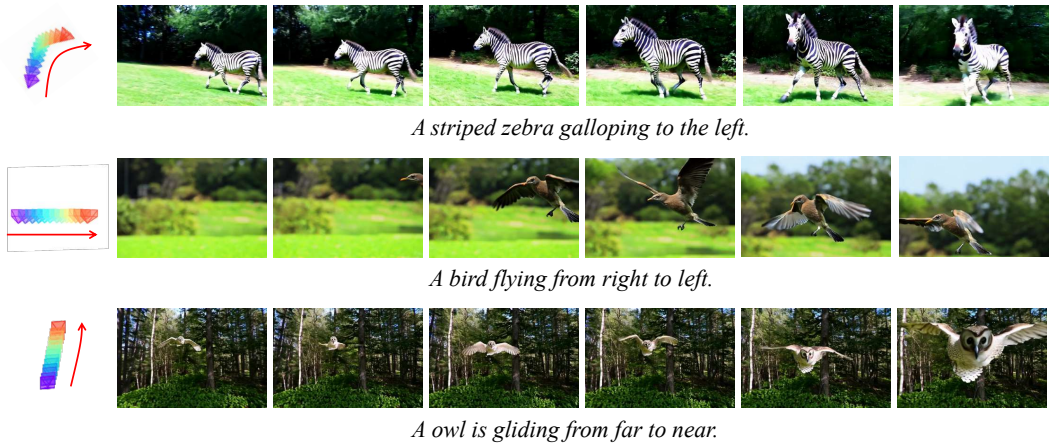

Figure 5: Qualitative results of our method under the conflicting motion. It shows that our motion disentanglement strategy seperates object motion from camera movements effectively. The foreground object motion are not overriden or distorted by the camera input.

4.6 Ablation Studies

Plücker Embedding. We directly use the numerical values of camera parameters (extrinsics and intrinsics) to evaluate the contribution of Plücker embedding. We use the linear projection to downsample the camera parameters along the spatial-temporal dimension of the visual features in the generation process. The experimental results are illustrated in Table 2. We find that using the Plücker embedding as the camera representation yields more precise camera control. Meanwhile, due to the strong geometric interpretation, the powerful geometric interpretation directly results in superior consistency for both objects and the background. Incorporating original numerical values directly could compromise geometric relationships.

Semantic Guidance Strategy. To evaluate the contribution of vision-language features from the foundational model in maintaining appearance consistency, we conduct both qualitative and quantitative studies to verify this. As in Table 2 and Figure 6, the introduction of the semantic guidance strategy leads to notable enhancements in object appearance consistency. Under strong camera movements, objects maintain their appearance without distortion. For example, without the semantic guidance strategy, the sea turtle in Figure 6 exhibits significant geometric distortion. Furthermore, the improved preservation of object appearance enables more effective high-frequency decomposition in subsequent processing stages.

High-Frequency Object Masking. As discussed in Sec. 3.3, we extract an object-aware mask through high-frequency decomposition for achieving motion disentanglement. We recongnize the mask as the fined-grained cues of the region and localization of the foreground object.To validate its importance, we perform an ablation by removing this decomposition and directly fusing camera-conditioned visual features with vision-language features. Results in Table 2 show that without high-frequency masking, both object and background consistency scores drop considerably. The mask helps the model better identify objects in the video frame by emphasizing foreground structures and localization, achieving motion disentanglement. Meanwhile, our motion disentanglement leads to a marked increase in the motion smoothness(MS) score, particularly in dynamic scenes.

Camera Fusion Strategies. Regarding camera fusion strategies, we evaluated two distinct approaches: element-wise addition fusion and cross-attention fusion. Existing approaches, such as CameraCtrl (He et al., 2024) and AC3D (Bahmani et al., 2024a), typically fuse camera conditions via element-wise addition. Specifically, they directly fuse image latent features and camera pose features through pixel-wise addition. However, as shown in Table 2, this fusion strategy falls short in achieving accurate camera control due to its limited capacity for pixel-wise understanding between camera parameters and visual features. In contrast, we adopt cross attention approach to achieve better performance on TransErr and RotErr. Benefiting from its stronger pixel-level understanding, the attention fusion mechanism also outperforms addition fusion on metrics such as FID, FVD, and ClipSim, leading to more realistic and semantically consistent video generation.

Methods	FID ↓	FVD ↓	CLIPSIM ↑	TransErr ↓	RotErr ↓	OC ↑	BC ↑	MS ↑
RealEstate10K								
W/O PLÜCKER EMBEDDING	225.8	694.7	0.309	0.758	0.210	93.5%	95.1%	98.4%
W/O SEMANTIC GUIDANCE	243.1	705.6	0.308	0.722	0.198	94.1%	95.8%	97.9%
W/O HIGH-FREQUENCY MODELING	235.4	649.8	0.309	0.744	0.209	94.5%	94.9%	98.0%
OURS (FULL, ADDITION FUSION)	236.2	771.8	0.310	0.738	0.211	94.6%	95.1%	98.2%
OURS (FULL, ATTENTION FUSION)	207.4	667.9	0.312	0.703	0.208	94.9%	96.4%	98.5%
VidGen								
W/O PLÜCKER EMBEDDING	258.5	1716.4	0.340	0.747	0.109	93.2%	94.5%	97.6%
W/O SEMANTIC GUIDANCE	231.7	1739.8	0.336	0.723	0.096	93.4%	94.0%	95.2%
W/O HIGH-FREQUENCY MODELING	233.4	1735.4	0.339	0.733	0.092	94.3%	94.6%	97.4%
OURS (ADDITION FUSION)	248.1	1738.6	0.345	0.732	0.084	94.9%	94.7%	97.4%
OURS (ATTENTION FUSION)	232.2	1643.7	0.349	0.724	0.081	94.7%	95.1%	98.3%

Table 2: Ablation studies on RealEstate10K and VidGen datasets. Lower is better (↓), higher is better (↑). Red indicates top-1 and blue indicates top-2 performance.

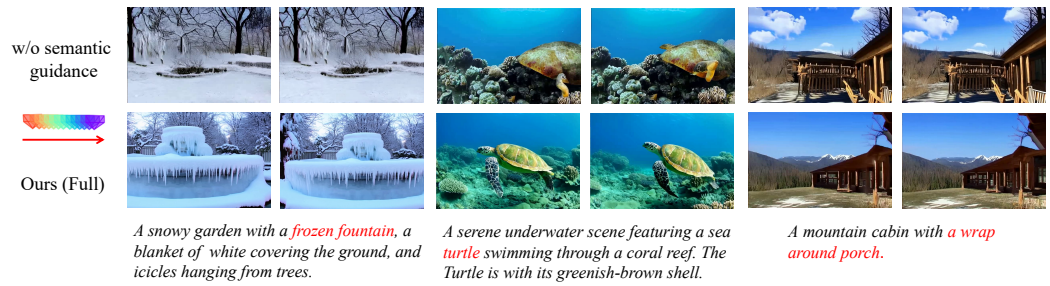

Figure 6: It shows our ablation study on the semantic guidance strategy. Without it, we observe that the generated videos suffer from object distortions.

4.7 ANALYSIS OF MODEL COMPLEXITY AND RUNTIME OVERHEAD.

In this section, we report both training cost and inference-time overhead of MoCa. During training, MoCa is fine-tuned from CogVideoX on 16×H200 GPUs with a batch size of 64 for 20,000 iterations. For inference, we measure per-sample latency on a single H200 GPU using BF16 precision and 50 denoising steps. Our MoCa has an inference latency of 291.2s, compared to 104.9s for the baseline model. The dual-branch fusion introduces 74.6s of overhead, providing persistent semantic guidance and improving appearance consistency. The 2D-DWT and hybrid condition fusion add only 9.73s (≈3.3% of the total time). Despite this small cost, the motion disentanglement module effectively enhances object motion and appearance consistency. Although MoCa has a higher inference latency than the baseline, it produces better visual-quality and camera-controllable videos. In practical video generation, users are typically far more concerned with visual quality and controllability than the inference speed. For latest diffusion-based video generation models, generating just a few seconds of a satisfying video requires several minutes of computation. This inherent mismatch between video length and inference time highlights that the quality of the generated video is more important than gains in the inference latency.

5 CONCLUSION

This work introduces MoCa, a framework for camera-controllable video generation that addresses the challenge of 3D consistency in 2D pixel space. By modeling object consistency across view, appearance, and motion, MoCa bridges the gap between explicit camera parameters and realistic video synthesis. Our approach incorporates a Spatial-Temporal Camera Encoder with Plücker embedding to maintain view consistency, a semantic guidance strategy using vision-language features to preserve object appearance, and an object-aware motion disentanglement mechanism to separate local object motion from global camera movements. Experimental results demonstrate that MoCa achieves accurate camera control while maintaining high video quality, offering a practical solution for consistent and controllable video generation.

6 Reproducibility Statement

To ensure the reproducibility of the results presented in this paper, we have made comprehensive efforts to provide necessary information and resources across the main content and appendix. Details regarding our project homepage and the anonymous code repository are available in Appendix A. The Sec. 3 in main content offers a detailed description of our model architecture, including key components, layer designs, and functional mechanisms. We also explicitly provide specific training parameters and detailed configurations of the datasets in Appendix E, ensuring clarity on the experimental setup. Beyond implementation details, the performance of our proposed method is fully validated through extensive experiments. The Sec. 4.1, 4.2, 4.3 in main content presents core results, comparative analyses and ablation studies. The Appendix F provides more results. With the integrated resources and detailed descriptions outlined above, other researchers should be able to reproduce our experimental results and verify the claims made in this paper.

7 Ethics Statement

All data used in this study are publicly available open-source datasets, with clear and legitimate sources that comply with relevant data usage licenses. This research does not involve any human subjects, human experiments, or collection of private or sensitive information from individuals. No potential ethical risks (such as bias, discrimination, harm to individuals or groups, or misuse of research outcomes) are associated with our study. Throughout the entire research process, we have strictly adhered to the principles outlined in the ICLR Code of Ethics, ensuring full compliance with academic ethics and social responsibility.

References

Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B Lindell, and Sergey Tulyakov. AC3D: Analyzing and Improving 3D Camera Control in Video Diffusion Transformers. *arXiv preprint arXiv:2411.18673*, 2024a.

Sherwin Bahmani, Ivan Skorokhodov, Aliaksandr Siarohin, Willi Menapace, Guocheng Qian, Michael Vasilkovsky, Hsin-Ying Lee, Chaoyang Wang, Jiaxu Zou, Andrea Tagliasacchi, et al. VD3D: Taming Large Video Diffusion Transformers for 3D Camera Control. *arXiv preprint arXiv:2407.12781*, 2024b.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. InstructPix2Pix: Learning to Follow Image Editing Instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18392–18402, 2023.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video Generation Models as World Simulators. *OpenAI Blog*, 1:8, 2024.

Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. VideoCrafter1: Open Diffusion Models for High-Quality Video Generation. *arXiv preprint arXiv:2310.19512*, 2023.

Wanquan Feng, Jiawei Liu, Pengqi Tu, Tianhao Qi, Mingzhen Sun, Tianxiang Ma, Songtao Zhao, Siyu Zhou, and Qian He. I2VControl-Camera: Precise Video Camera Control with Adjustable Motion Strength. *arXiv preprint arXiv:2411.06525*, 2024.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning. *International Conference on Learning Representations*, 2024.

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.

Hao He, Ceyuan Yang, Shanchuan Lin, Yinghao Xu, Meng Wei, Liangke Gui, Qi Zhao, Gordon Wetzstein, Lu Jiang, and Hongsheng Li. CameraCtrl II: Dynamic Scene Exploration via Camera-controlled Video Diffusion Models. *arXiv preprint arXiv:2503.10592*, 2025.

Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent Video Diffusion Models for High-Fidelity Long Video Generation. *arXiv preprint arXiv:2211.13221*, 2022.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video Diffusion Models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers. *arXiv preprint arXiv:2205.15868*, 2022.

Chen Hou, Guoqiang Wei, Yan Zeng, and Zhibo Chen. Training-free Camera Control for Video Generation. *arXiv preprint arXiv:2406.10126*, 2024.

Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8153–8163, 2024.

Chao-Tsung Huang, Po-Chih Tseng, and Liang-Gee Chen. Analysis and VLSI Architecture for 1-D and 2-D Discrete Wavelet Transform. *IEEE Transactions on signal processing*, 53(4):1575–1586, 2005.

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.

Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. DreamPose: Fashion Video Synthesis with Stable Diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22680–22690, 2023.

Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. MegaSaM: Accurate, Fast, and Robust Structure and Motion from Casual Dynamic Videos. *arXiv preprint arXiv:2412.04463*, 2024.

Xin Ma, Yaohui Wang, Xinyuan Chen, Gengyun Jia, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent Diffusion Transformer for Video Generation. *Transactions on Machine Learning Research*, 2025.

Willi Menapace, Aliaksandr Siarohin, Ivan Skorokhodov, Ekaterina Deyneka, Tsai-Shien Chen, Anil Kag, Yuwei Fang, Aleksei Stoliar, Elisa Ricci, Jian Ren, et al. Snap Video: Scaled Spatiotemporal Transformers For Text-to-Video Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7038–7048, 2024.

Al-Shuraifi Mushtaq, Al-Anssari Ali Ihsan, and Nameer Qasim. 2D-DWT vs. FFT OFDM Systems in fading AWGN channels. *Radioelectronics and Communications Systems*, 58(5):228–233, 2015.

William Peebles and Saining Xie. Scalable Diffusion Models with Transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. MM-Diffusion: Learning Multi-Modal Diffusion Models for Joint Audio and Video Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10219–10228, 2023.

Masaki Saito, Eiichi Matsumoto, and Shunta Saito. Temporal Generative Adversarial Nets with Singular Value Clipping. In *Proceedings of the IEEE international conference on computer vision*, pp. 2830–2839, 2017.

Asadollah Shahbahrami. Algorithms and architectures for 2D discrete wavelet transform. *The Journal of Supercomputing*, 62:1045–1064, 2012.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-A-Video: Text-to-Video Generation without Text-Video Data. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=nJfylDvgzlq.

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.

Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. StyleGAN-V: A Continuous Video Generator with the Price, Image Quality and Perks of StyleGAN2. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3626–3636, 2022.

Zhiyu Tan, Xiaomeng Yang, Luozheng Qin, and Hao Li. Vidgen-1M: A Large-Scale Dataset for Text-to-Video Generation. *arXiv preprint arXiv:2408.02629*, 2024.

Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive generating expressive portrait videos with audio2video diffusion model under weak conditions. In *European Conference on Computer Vision*, pp. 244–260. Springer, 2024.

Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. MoCoGan: Decomposing Motion and Content for Video Generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1526–1535, 2018.

Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating Videos with Scene Dynamics. *Advances in neural information processing systems*, 29, 2016.

Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. VideoComposer: Compositional Video Synthesis with Motion Controllability. *Advances in Neural Information Processing Systems*, 36:7594–7611, 2023.

Xiang Wang, Shiwei Zhang, Hangjie Yuan, Zhiwu Qing, Biao Gong, Yingya Zhang, Yujun Shen, Changxin Gao, and Nong Sang. A Recipe for Scaling up Text-to-Video generation with Text-free Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6572–6582, 2024a.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. MotionCtrl: A Unified and Flexible Motion Controller for Video Generation. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024b.

Ruiqi Wu, Liangyu Chen, Tong Yang, Chunle Guo, Chongyi Li, and Xiangyu Zhang. LAMP: Learn A Motion Pattern for Few-Shot Video Generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Dejia Xu, Weili Nie, Chao Liu, Sifei Liu, Jan Kautz, Zhangyang Wang, and Arash Vahdat. Camco: Camera-Controllable 3D-Consistent Image-to-Video Generation. *arXiv preprint arXiv:2406.02509*, 2024.

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.

Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024.

Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2VGen-XL: High-Quality Image-to-Video Synthesis via Cascaded Diffusion Models. *arXiv preprint arXiv:2311.04145*, 2023.

Sixiao Zheng, Zimian Peng, Yanpeng Zhou, Yi Zhu, Hang Xu, Xiangru Huang, and Yanwei Fu. Vidcraft3: Camera, object, and lighting control for image-to-video generation. *arXiv preprint arXiv:2502.07531*, 2025.

Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. MagicVideo: Efficient Video Generation With Latent Diffusion Models. *arXiv preprint arXiv:2211.11018*, 2022.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo Magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018.

A Appendix / Supplemental Material

For a more comprehensive evaluation, additional visualizations and generated examples are provided on our project page at: https://anonymous.4open.science/w/MoCa-31E5/. The implementation code is also publicly available at: https://anonymous.4open.science/r/MoCa-31E5/ to facilitate reproducibility and further research. We encourage readers to explore these resources for a deeper understanding of our method.

This supplementary document offers additional details, extended analyses, and further experimental results that support the main content of the paper. It includes implementation specifics, extra qualitative comparisons, and more visualizations of the generated videos. We structure the appendix as follows: Appendix B claim the usage of large language models. Appendix C provides a brief review of ReferenceNet-based camera-controlled video generation. Appendix D explains the use of 2D Discrete Wavelet Transform for high-frequency object-aware masking. Appendix E presents additional experiment details, including settings and dataset information. Appendix F includes more qualitative results, especially in dynamic scenes. Appendix G discusses the limitations.

B The Usage of Large Language Models

We acknowledge the use of a large language model for limited assistance in the writing of this paper. The tool was used exclusively for proofreading and improving the linguistic fluency of the text. All scientific content, including the research ideas, methodology, experiment and analysis, was conducted solely by the authors.

C ReferenceNet Based Camera-Controlled Video Generation

C.1 Text-to-Video Diffusion Models.

Modern text-to-video (T2V) models typically builds upon a pre-trained text-to-image (T2I) model, such as Stable Diffusion (SD), a pioneering framework that follows the design of the Latent Diffusion Model (LDM). In this work, we follow the T2V model definition from CameraCtrl. As an essential part of SD, Variational AutoEncoder compresses the feature distribution of the original image, denoted as x_0 into a latent space representation z_0. The encoding operation extracts the image essence as $z_0 = E(x_0)$ whereas the decoding counterpart reconstruct the image from the latent via $x_0 = D(z_0)$. The diffusion process is then conducted in the latent space, which significantly reduces computational overhead without compromising generation performance.

During the diffusion phase, these models usually corrupt the latent z_0 by adding Gaussian noise ϵ according to a predefined schedule inherited from Denoising Diffusion Probabilistic Model (DDPM) or its deterministic variant DDIM. Then the denoising network is optimized to reverse this process by progressively eliminating the introduced noise directed by some conditional embeddings c, to yield videos that adhere to the prescribed text prompts. The training objective can be formulated as follows:

$$\mathcal{L} = E_{t,\epsilon,\mathbf{z}_0^{1:N},\mathbf{c}}\left[\left\|\epsilon - \epsilon_\theta(\mathbf{z}_t^{1:N}, t, \mathbf{c})\right\|^2\right]$$

where N denotes the number of video frames, c is the text embeddings transformed from the input prompts utilizing the CLIP ViT-L/14 text encoder, ϵ_θ is a Unet with learnable weights θ. The Unet is composed of pairs of down/up blocks as well as a middle block. Each block consist of ResNets, spatial and temporal self-attention layers, together with cross-attention layers that interact with text conditions, thereby promoting the model's capability to generate videos that are semantically consistent with text.

C.2 Camera-controlled text-to-video generation.

In the field of text-to-video generation, task about adding camera trajectory control s_t has already been extensively explored. By incorporating structural camera signals, the video generation results can be more controllable. Specifically, the camera trajectories are first processed by a special encoder $\phi(\cdot)$ and then fed into the video generator for further operations. Consequently, the objective

of the generator with guidance from camre signals can be formulated as follows:

$$\mathcal{L} = E_{t,\epsilon,\mathbf{z}_0^{1:N},\mathbf{c},s_t}\left[\left\|\epsilon - \epsilon_\theta(\mathbf{z}_t^{1:N}, t, \mathbf{c}, \phi(s_t))\right\|^2\right]$$

where N is the number of video frames, c is the text embeddings, and s_t means camera trajectories at different timesteps.

C.3 REFERENCENET.

Prior research has pointed out that utilizing analogous structures is crucial in maintaining the identity consistency of the target object. Therefore, ReferenceNet that mirrors the architectural design of SD and operates in parallel with the Denoising Unet, is widely adopted to assist in modelling complex image or video features. EMO and EchoMimic facilitate the self-attention mechanism in the ReferenceNet to extract reference image features into the attention layers of corresponding block in the Denoising U-Net, making the facial identity more consistent throughout the video. Hallo integrates features from the same spatial resolution layers into the Denoising Unet to enhance the visual texture information of both portraits and backgrounds in the generated videos. Meanwhile, as ReferenceNet shares identical network structure and initialization weights, the Denoising Unet can selectively learn some correlated features from it in the same feature space. AnimateAnyone leverages the learned reference image features from ReferenceNet to produce a well-initialized latent, thus accelerating the entire network training process of the Denoising Unet.

D 2D-DWT IN HIGH-FREQUENCY OBJECT-AWARE MASKING

As mentioned in Sec.3.3, we use high-frequency object masking strategy to extract object-aware mask for better separating objects and background. In this section, we give more detail about this strategy. Formally, given the vision-language features from the foundational model $\mathbf{X} \in \mathbb{R}^{B\times C\times H\times W}$, we decompose them into four frequency sub-bands using 2D-DWT:

$$\text{DWT}(\mathbf{X}) \to \{\mathbf{LL},\ \mathbf{LH},\ \mathbf{HL},\ \mathbf{HH}\}\ , \tag{1}$$

where $\mathbf{LL}$ denotes the low-frequency approximation coefficients, and $\mathbf{LH}$, $\mathbf{HL}$, $\mathbf{HH}$ represent high-frequency details in horizontal, vertical, and diagonal directions, respectively. To obtain these components, we adopt separable 1D Haar wavelet filters and perform sequential convolutions:

$$h_L = \frac{1}{\sqrt{2}}[1,1], \quad h_H = \frac{1}{\sqrt{2}}[-1,1], \tag{2}$$

$$\begin{aligned} \mathbf{LL} &= (\mathbf{X} * h_L^\top) * h_L, \quad \mathbf{LH} = (\mathbf{X} * h_L^\top) * h_H, \\ \mathbf{HL} &= (\mathbf{X} * h_H^\top) * h_L, \quad \mathbf{HH} = (\mathbf{X} * h_H^\top) * h_H. \end{aligned} \tag{3}$$

Unlike conventional 1D temporal wavelet transforms or global 2D Fourier transforms, our 2D-DWT formulation preserves localized spatial-frequency characteristics, which are essential for accurately identifying object contours in dynamic scenes. To emphasize fine-grained structural cues, we discard the low-frequency component $\mathbf{LL}$ and retain only the high-frequency sub-bands. These are then used to reconstruct a high-frequency-enhanced representation via inverse DWT (iDWT):

$$\mathbf{X}_{\text{high}} = \text{iDWT}(0, \mathbf{LH}, \mathbf{HL}, \mathbf{HH}). \tag{4}$$

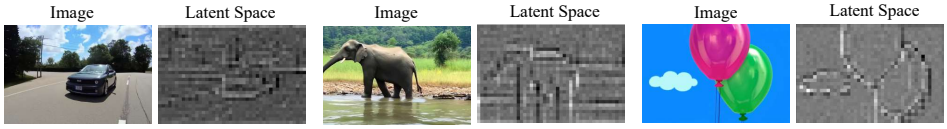

Figure 7: This figure presents a visualization of applying high-frequency decomposition to visual features in a latent space. Our strategy yields an object-aware mask that effectively captures the structure and localization of the foreground object. As shown in the right case, our high-frequency decomposition can accurately extract the structure of all objects, even in scenes containing multiple objects of different classes.

As visualized in Fig. 7, this operation highlights the structure and localization of foreground objects in the latent space. We view the high-frequency representation as an object-aware mask. From the ablation study in the main content, the mask enhances motion consistency performance by improving the model's object identification capability. It achieves this by emphasizing foreground structures in the latent space, which promotes a clearer separation between foreground and background elements.

E More Experiment Details

E.1 Training and Inference Details

Our model is built upon CogVideoX (Yang et al., 2024), a transformer-based text-to-video diffusion model with approximately 5B parameters, which demonstrates leading performance on both automated metrics and human evaluations. Similar to ReferenceNet, the weights of our DenoisingNet are inherited from the original CogVideoX and remain frozen during training. Only the dual-branch fusion module is optimized. We employ the AdamW optimizer with a batch size of 64, an initial learning of 1×10^{-4}, epsilon set to 1×10^{-8} weight decay of 1×10^{-4} and beta values of 0.9 and 0.95. Training is conducted using 16 NVIDIA H200 GPUs.

During inference, to increase the magnitude of camera motion and enhance the challenge of camera-controllable video generation, we sample the input camera trajectory by selecting 49 frames at equal intervals from the original 98-frame sequence.

E.2 Dataset

During training, our model is fine-tuned from CogVideoX on the RealEstate10K dataset, which comprises approximately 80,000 video clips extracted from around 10,000 YouTube videos, totaling about 10 million frames. For each clip, camera parameters (extrinsics and intrinsics) are provided for every frame, forming a continuous trajectory estimated by running SLAM and bundle adjustment algorithms on the original videos. This dataset supports view synthesis and 3D computer vision tasks, with scenes often focusing on static environments such as furnished indoor spaces and natural outdoor landscapes.

For evaluation, we test performance on both RealEstate10K and the VidGen dataset. VidGen is a large-scale collection of text-video pairs designed for text-to-video generation. The dataset is created through a meticulous curation process involving rough and fine-grained filtering to ensure high-quality videos with detailed, temporally consistent captions. In contrast to the static scenes in RealEstate10K, VidGen features predominantly dynamic scenes. Figure 8 shows samples of scenes from the two datasets.

We leverage the VidGen dataset to validate the effectiveness of our approach in terms of object stability and motion consistency. The diverse and complex dynamics within VidGen's videos enable a rigorous test of our model's ability to generalize beyond static environments, thereby demonstrating its robustness in handling complex, dynamic scenes, as visualized in the qualitative results shown in Figure 11, 12.

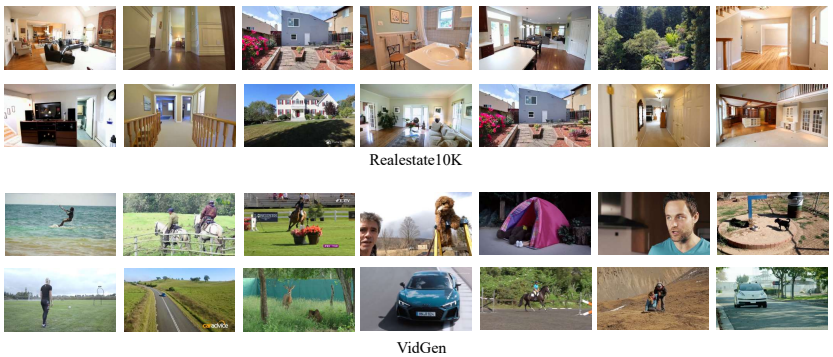

Figure 8: The overview of the dataset Realestate10K and VidGen.

F More Experiment Results

F.1 Quantitative Comparison on Dynamic Scenes

In this section, we present additional qualitative comparison results. The sequence from left to right in the figure represents the progression from the start to the end of the video. As shown in Figure 9, we provide three new challenging cases, where each case involves animals or objects with self-driven motion. Our method demonstrates excellent performance in terms of view, appearance, and motion consistency.

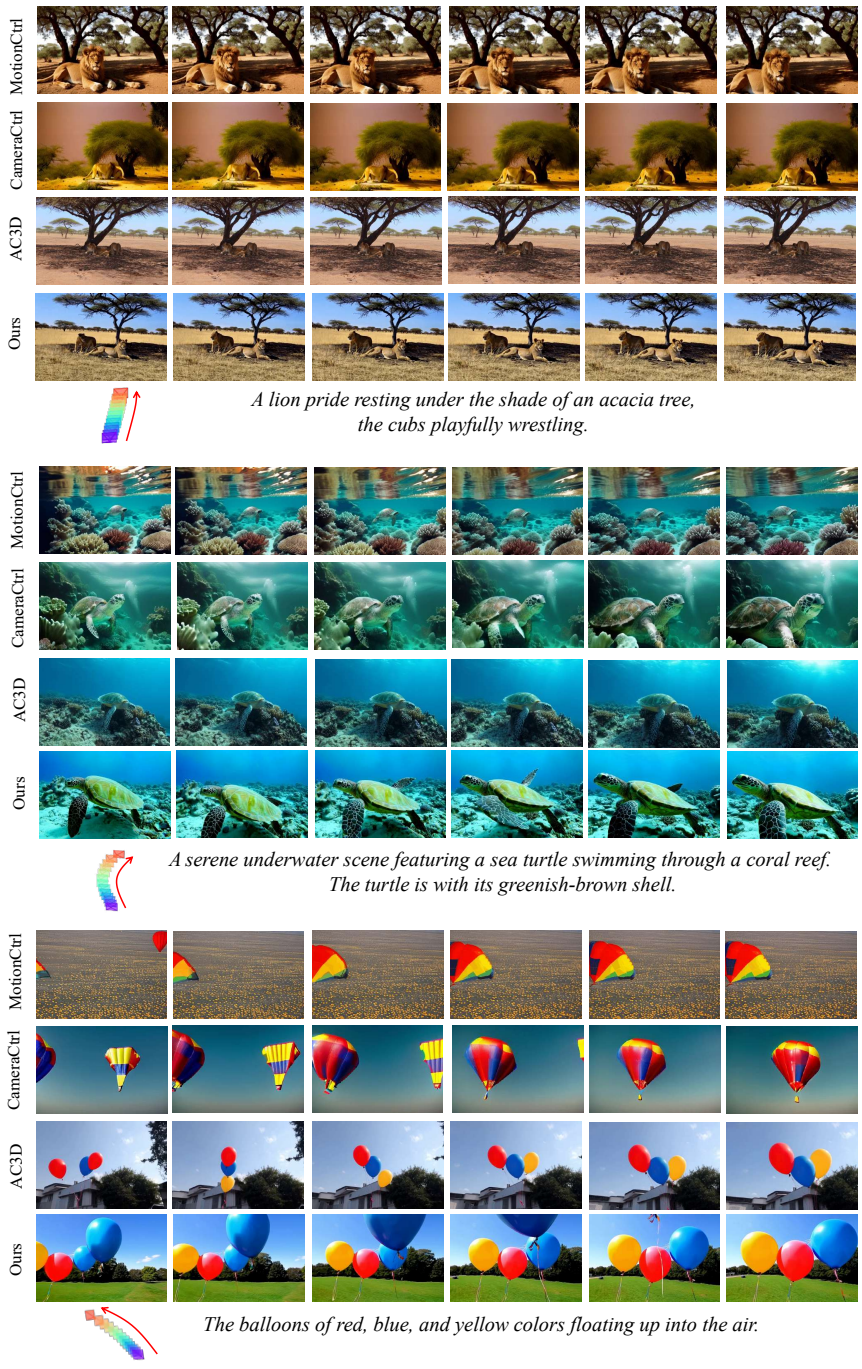

Figure 9: More qualitative comparison between our method and existing approaches.

F.2 Quantitative Comparison on Static Scenes

In this section, we present additional qualitative comparison results. The sequence from left to right in the figure represents the progression from the start to the end of the video. As shown in Figure 10, we provide three new cases of static scenes. Our method demonstrates excellent performance in terms of view, appearance, and motion consistency.

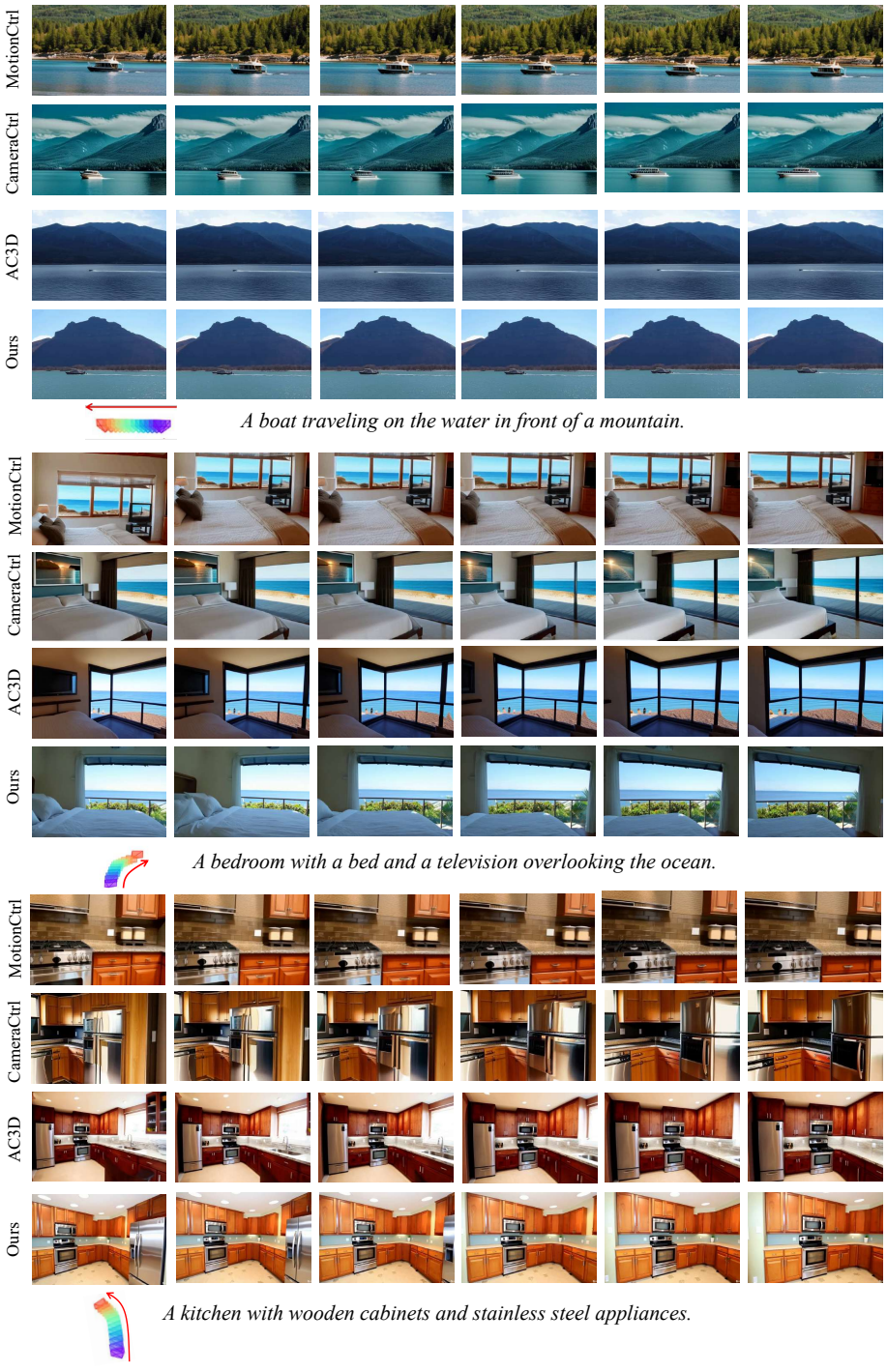

Figure 10: More qualitative comparison between our method and existing approaches.

F.3 More Quantitative Results on Dynamic Scenes

In this section, we provide additional qualitative results under dynamic scene settings. Dynamic scenes are characterized by the presence of distinct objects with self-driven motion and a clear separation between foreground and background elements, presenting a particularly challenging scenario for video generation models.

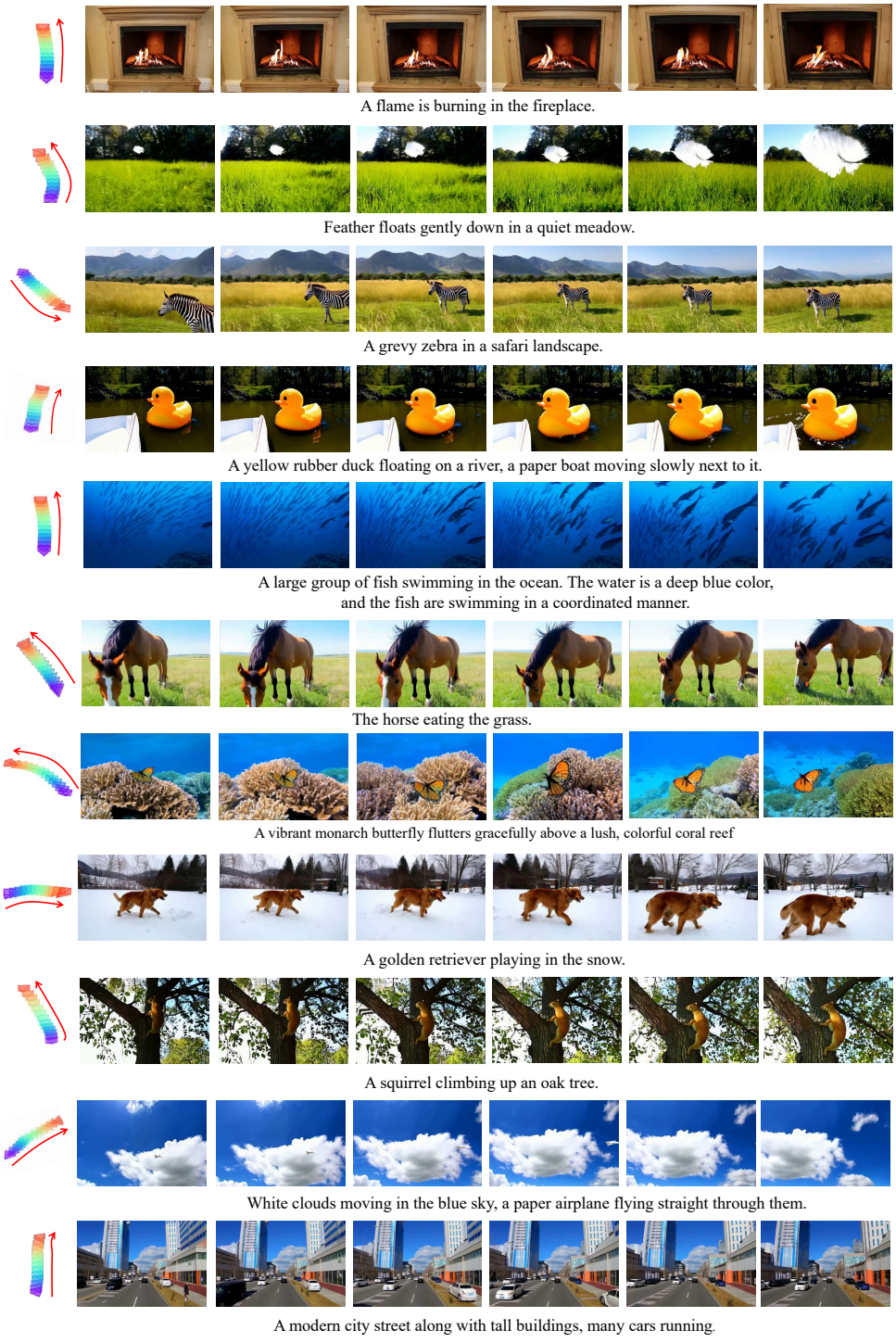

Figure 11: More quantitative results of MoCa on dynamic scenes.

F.4 More Quantitative Results on Static Scenes

In this section, we provide additional qualitative results under static scene settings. Static scenes typically encompass both indoor and outdoor environments featuring complex textures and geometric structures. These scenarios place strong emphasis on a model's ability to comprehend spatial layout and accurately model camera motion through the sequence.

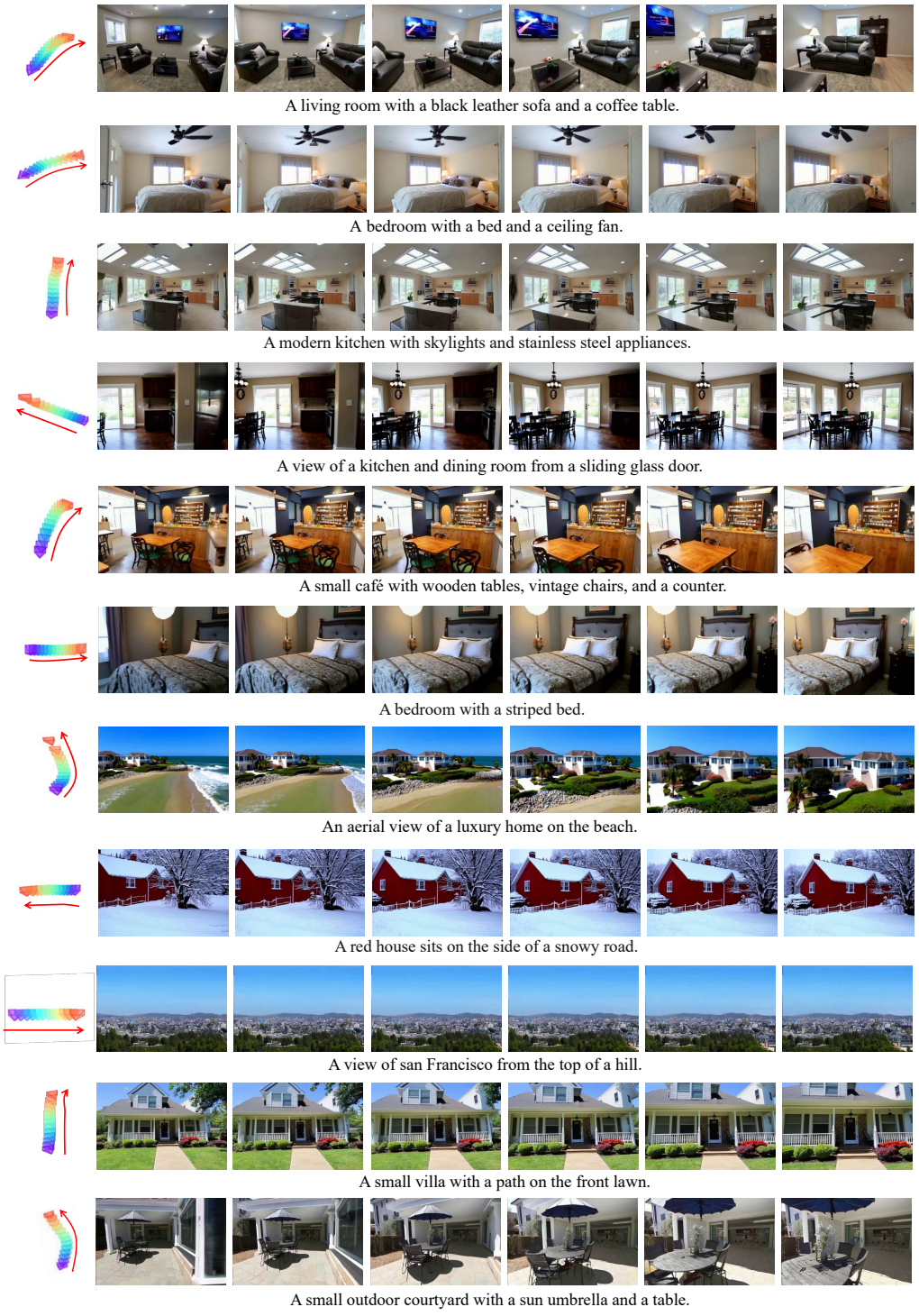

Figure 12: More quantitative results of MoCa on static scenes

G LIMITATIONS

The proposed MoCa, while demonstrating capability in generating high-quality camera-controlled videos, exhibits two certain limitations. Specifically, the current method primarily focuses on integrating camera control into text-to-video generation frameworks, without extending this capability to other multimodal data inputs. Future research should prioritize the idea of multimodal video generation capable of processing diverse input modalities, such as secondary editing of object regions and video style transfer applications. Secondly, although our approach successfully maintains object stability with camera movements, the current framework cannot precisely control where moving objects are positioned in the frame. For example, an object might unintentionally appear near the edges, resulting in less visual effect. Future work should address these challenges through enhanced object control generation and multimodal fusion techniques, ultimately aiming to achieve superior video generation quality with expanded creative possibilities.