CINESCENE: Implicit 3D as Effective Scene Representation for Cinematic Video Generation

Supplementary Material

A. Dataset Construction

We follow [2] to use Unreal Engine 5 to construct Scene-Decoupled Video Dataset as described in Section 3.

3D Environments. We collect 35 different 3D environment assets from here. We mitigate the domain shift from rendered data to real-world distributions by prioritizing high-fidelity 3D assets while incorporating surreal environments as an auxiliary. Environmental variety is achieved through a diverse selection of internal and external venues, ranging from metropolitan streets and commercial interiors to pastoral settings.

Subjects. We collect 70 different human 3D models as subjects from source A and source B, covering a wide range of styles such as realistic, anime, game-style.

Animations. We collect around 100 different animations from source A and source B, including diverse dynamic animations (*e.g.*, waving, dancing, cheering.

Camera Trajectories In order to achieve large view changes, we add constraints to the camera movements. The speed of camera movements is set to be linear.

1. **Arc Movements (Left/Right)**: For "arc" movements, the camera traverses an arc spanning 75 degrees over 77 frames, continuously tracking the dynamic subject to ensure it remains in view.
2. **Pan Movements (Left/Right)**: For "pan" movements, we require a 75-degree rotation over 77 frames. This movement involves only camera rotation, with no change in its spatial position.
3. **Arc and Tilt Movements (Up/Down)**: The vertical camera motion is constrained between 10 and 45 degrees relative to its initial orientation. For "arc" movements, the camera tracks the dynamic subject. For "tilt" movements, the camera only rotates with no change in translation.
4. **Dolly, Truck, and Pedestal Movements (Left/Right/Up/Down/Forward/Backward)**: For dolly, truck, and pedestal movements, we randomly select the distance thresholds relative to the dynamic subject. The specific ranges for these movements are defined as follows:
   - Left: $[1/4, 2]$
   - Right: $[1/4, 2]$
   - Forward: $[1/4, 5/4]$
   - Backward: $[1/4, 2]$
   - Up: $[1/4, 2/3]$
   - Down: $[1/4, 2/3]$

   These values represent the relative distance changes with respect to the subject.

Initial Viewpoint: The starting viewpoint for all camera movements and the panoramic images is randomly selected within a range of -45 to 45 degrees. This angle is measured relative to the front-facing direction of the dynamic subject, introducing variability in initial perspectives.

B. Implementation Details

Our model's trainable components include: 1) the 3D attention, projector, and feedforward layers within the DiT Blocks; 2) a convolutional layer and a Layer Normalization module for projecting the implicit 3D features; and 3) a learnable camera encoder for projecting the camera condition [2]. For the baselines compared [2, 45, 63], we utilize the default parameters (*e.g.*, number of generated frames) provided in their official GitHub repositories to ensure a fair evaluation of their standard performance.

C. Experimental Results

C.1. Out-of-Domain Test

To evaluate the model's generalization capabilities, we conduct an out-of-domain (OOD) test using the DiT360 dataset [7, 16]. This section first explains the construction of our test set and then presents the quantitative and qualitative results.

OOD Test Set Construction. The DiT360 dataset [7, 16] comprises static panoramic views stitched from RGB-D images of building-scale scenes. From this dataset, we selected 50 panoramic images as our test samples. Since each panorama originates from a single viewpoint, we are limited to "pan" camera trajectories, which involve only rotational movements without any translation. To construct the test set, we process each panoramic image in two ways: 1) We project the panorama into a series of context images by sampling viewpoints along the horizontal plane at 18° increments; 2) We define three distinct camera trajectories to generate the ground truth sequences. It is important to note that this test set exclusively contains static scenes. Consequently, this evaluation focuses on comparing the methods' abilities to generate coherent and consistent static environments.

Quantitative Results on OOD Test Set.

The quantitative results are presented in Table 4. Our proposed method demonstrates superior performance in scene consistency, as evidenced by its leading scores across the Mat. Pix., PSNR, SSIM, and LPIPS metrics. Furthermore, our model achieves the second-highest CLIP-V score.

Table 4. **Quantitative comparision with previous methods on OOD test set.** We compare CINESCENE with FramePack [71] on scene consistency, Context-as-Memory [68] and Gen3C [45] on both scene consistency and camera accuracy, Traj-Attn [63] and RecamMaster [2] on camera accuracy. We follow [64] to evaluate video quality on VBench.

Model	Scene Consistency					Camera Accuracy			Text Alignment	Video Quality
	Mat. Pix.(K)↑	CLIP-V↑	PSNR↑	SSIM↑	LPIPS↓	RotErr↓	TransErr↓	CamMC↓	CLIP-T↑	VBench↑
Context-based Method										
FramePack [71]	4025.98	0.8784	9.9902	0.3529	0.6625	-	-	-	**0.3223**	0.7813
Context-as-Memory [68]	4338.75	0.8976	10.1676	0.364	0.6698	5.5482	11.2567	14.0722	0.3112	**0.8093**
Explicit 3D Guidance Method										
Gen3C [45]	4246.33	**0.9049**	11.9062	0.4412	0.6309	4.2975	10.8731	13.1951	0.3004	0.738
Camera-Controlled Method										
Traj-Attn [63]	-	-	-	-	-	4.3117	13.2458	14.899	0.3072	0.7635
RecamMaster [2]	-	-	-	-	-	3.7125	11.3909	13.0711	0.3052	0.7705
Ours	**4726.57**	0.9038	**12.0215**	**0.4470**	**0.5121**	**3.6981**	**10.8435**	**12.3307**	0.3025	0.7965

Table 5. **Ablation study on number of scene context images.**

# Num	Scene Consistency					Camera Accuracy		
	Mat. Pix. (K)↑	CLIP-V↑	PSNR↑	SSIM↑	LPIPS↓	RotErr↓	TransErr↓	CamMC↓
1	3614.14	0.7876	10.3426	0.1682	0.6478	3.9551	8.7780	10.8669
4	4200.82	0.8167	10.7466	0.2215	0.6069	**2.3948**	7.7557	8.9438
10	4519.10	0.8389	11.9169	0.2890	0.5447	3.0033	7.9496	9.5312
20	**4617.51**	**0.8633**	**14.5094**	**0.4133**	**0.4241**	2.6825	**5.1460**	**6.8819**

Table 6. **Ablation on camera control condition**.

	RotErr↓	TransErr↓	CamMC↓
W/o Implicit	2.7362	5.4411	7.1805
W/o Camera	11.3678	11.1976	19.5482
Ours	**2.6825**	**5.146**	**6.8819**

In terms of camera accuracy, the RotErr, TransErr, and CamMC metrics confirm the effectiveness of our method.

Qualitative Results on OOD Test Set. Please see the qualitative results in the supplementary video.

C.2. Ablation Study

Ablation on Number of Scene Context Images.

We conduct four experiments on different numbers of scene context images: 1, 4, 10, and 20. The results are shown in Table 5. The results show that increasing the number of scene context images leads to a consistent improvement in performance across both scene consistency and camera accuracy. For scene consistency, all metrics show a clear positive trend. As the number of views increases from 1 to 20, the Mat. Pix. and CLIP-V scores steadily rise, while the LPIPS error consistently decreases. This demonstrates that more contextual information helps the model generate scenes with higher geometric fidelity and semantic coherence. PSNR and SSIM, show a marked increase, signifying a substantial enhancement in the clarity and structural integrity of the generated videos. For camera accuracy, the overall trend also confirms the benefit of using more views, with the 20-view configuration achieving the lowest error rates across RotErr, TransErr, and CamMC. We observe a slight, non-monotonic degradation in performance when moving from 4 to 10 views. This might suggest that a moderate number of views can introduce temporary ambiguities for camera pose estimation, which are effectively resolved when a richer set of 20 views provides more robust constraints.

Impact of Camera Control. We conduct three experiments on camera accuracy: (1) without implicit 3D information, with camera condition; (2) with implicit 3D information, without camera condition; (3) with both (ours). Results in Table 6 show that the input camera instruction provides fine-grained and accurate control, while implicit 3D information further boosts the camera control.

Different Condition Mechanisms. We show the results in Table 8. (1) Different concatenations. We ablate injecting conditions on channel/view-dimension [2]. We found that frame-dimension (ours) provides a more robust way for the spatio-temporal interactions among conditions, which is also observed in [2]. (2) Different Fusion Strategies. We compare ours with concatenating F_i and F_c directly into tokens without element-wise addition. Our fusion strategy better integrates scene content and viewpoints, enhancing consistency. (3) The camera accuracy are on par among the different mechanisms, as it is mainly impacted by injecting camera trajectory (also validated in Table 6 and Table 9).

Inconsistent Modalities. We evaluate robustness by removing scene descriptions or using inconsistent prompts

Table 7. **Ablation study on scene descriptions.**

Method	Scene Consistency					Camera Accuracy			Text Alignment	Video Quality
	Mat. Pix.(K)↑	CLIP-V↑	PSNR↑	SSIM↑	LPIPS↓	RotErr↓	TransErr↓	CamMC↓	CLIP-T↑	VBench ↑
remove scene description	4589.18	0.8576	14.3361	**0.4159**	0.4293	2.4974	5.3	6.8412	0.3057	0.8022
add inconsistent description	4577.54	0.8547	14.2669	0.4103	0.4308	**2.474**	5.2749	**6.7745**	0.3048	0.8006
Ours	**4617.51**	**0.8633**	**14.5094**	0.4133	**0.4241**	2.6825	**5.146**	6.8819	**0.3212**	**0.8053**

Table 8. **Ablation study on different feature fusion strategies.**

Method	Scene Consistency					Camera Accuracy			Text Alignment	Video Quality
	Mat. Pix.(K)↑	CLIP-V↑	PSNR↑	SSIM↑	LPIPS↓	RotErr↓	TransErr↓	CamMC↓	CLIP-T↑	VBench ↑
channel-dim	4564.53	0.8301	14.0434	0.4024	0.4372	2.4813	5.0500	6.6230	0.3170	0.8041
view-dim	4407.87	0.8265	12.8336	0.3757	0.4974	2.4915	5.9913	7.4144	0.3213	0.7982
separate F_i, F_c	4546.81	0.8533	14.1516	0.4037	0.4449	2.7345	5.4331	7.1764	**0.3229**	0.8022
Ours	**4617.51**	**0.8633**	**14.5094**	**0.4133**	**0.4241**	2.6825	5.1460	6.8819	0.3212	**0.8053**

Table 9. Ablation on camera control condition. Supplement to Table 2

	RotErr↓	TransErr↓	CamMC↓
w/o Implicit	2.7362	5.4411	7.1805
w/ Image feature	2.4967	5.2081	6.7754
w/ Camera feature	2.7631	5.1620	6.9542
w Implicit (Ours)	2.6825	5.146	6.8819

Table 10. **Ablation on OOD trajectories.**

Method	Camera Accuracy			Text Alignment	Video Quality
	RotErr↓	TransErr↓	CamMC↓	CLIP-T↑	VBench↑
ReCamMaster [2]	**2.2485**	7.4892	8.5504	0.3051	0.8015
Traj-Attn [63]	3.4825	8.6681	10.5595	0.2669	0.7728
Ours	2.4147	**6.7794**	**8.0997**	**0.3163**	**0.8022**

(e.g., "in the garden"). Despite a slight performance drop, our method remains robust to missing or mismatched descriptions, shown in Table 7.

Out-of-range trajectories. Following [39], we evaluate 300 unseen trajectories from RealEstate10K [77]. As groud truth videos are lack for consistency evaluation, we report all other available metrics in Table 10. Our model shows generalization to out-of-range trajectories, benefiting from the 46K diverse trajectories in our dataset.

C.3. More Qualitative Results

Please refer to our project page.

D. Limitations

Our current work presents several limitations that motivate future research directions:

1) We explore the generation of short video clips (77 frames) with a maximum view change of 75 degrees. Extending scene consistency to longer videos with larger view changes remains a challenging but important area for future investigation.

2) To simplify the problem, we provide the first scene context image with the same viewpoint as the first frame of the generated video. Future work will address the generation of videos from random camera positions.

3) CINESCENE inherits limitations from the pre-trained T2V models, such as distortion in humans' large motion movements.