MemCam: Memory-Augmented Camera Control for Consistent Video Generation

Xinhang Gao, Junlin Guan, Shuhan Luo, Wenzhuo Li, Guanghuan Tan, and Jiacheng Wang
Guilin University of Electronic Technology, China
Corresponding author: Junlin Guan (e-mail: guanjunlin@guet.edu.cn)

Fig. 1: **MemCam** generates videos with high scene consistency over long time horizons and under large camera rotations. Rows 1-2 show results on the test split of our dataset, including 360° round-trip and 360° single-direction rotation. Row 3 presents open-domain real-world scenes, demonstrating that MemCam maintains long-term scene consistency across varied scenes.

***Abstract*—Interactive video generation has significant potential for scene simulation and video creation. However, existing methods often struggle with maintaining scene consistency during long video generation under dynamic camera control due to limited contextual information. To address this challenge, we propose MemCam, a memory-augmented interactive video generation approach that treats previously generated frames as external memory and leverages them as contextual conditioning to achieve controllable camera viewpoints with high scene consistency. To enable longer and more relevant context, we design a context compression module that encodes memory frames into compact representations and employs co-visibility-based selection to dynamically retrieve the most relevant historical frames, thereby reducing computational overhead while enriching contextual information. Experiments on interactive video generation tasks show that MemCam significantly outperforms existing baseline methods as well as open-source state-of-the-art approaches in terms of scene consistency, particularly in long video scenarios with large camera rotations.**

***Index Terms*—interactive video generation, camera control, diffusion model**

I. Introduction

Recent advances in video generation models [1]–[4] have made substantial progress, enabling higher visual quality and realism in video synthesis. Within the broader landscape of video generation, the sub-field of interactive video generation has gained prominence as a key research focus. Its growing importance stems from promising applications in areas such as video or game scene generation [5]–[7] and world simulation [8]–[10]. Recent research on long video generation [11]–[14] has provided valuable methodologies to maintain coherence over extended sequences, thus significantly supporting and accelerating developments in interactive video generation.

Despite these advances, existing methods still face significant challenges when generating long interactive videos: they struggle to maintain the consistency of the scene content over extended temporal sequences [15], [16]. This limitation manifests itself as the model's tendency to forget previously generated content during continuous synthesis. When the camera viewpoint returns to a previously displayed region after complex motion, it often leads to content discrepancies within the same scene at different time points [5], [6], [15]. This is because these methods lack memory capability—they can only rely on camera information without explicit memory of past content, such as CameraCtrl [17], or depend on a fixed-length context window of limited scope, as in DFoT [16]. Some approaches attempt to improve this by incorporating 3D reconstruction, like GeometryForcing [18], but still face

issues such as error accumulation and inherent performance limitations of the reconstruction models themselves.

In this paper, we propose MemCam, a memory-augmented framework for scene-consistent interactive video generation. MemCam explicitly maintains historical frames as contextual memory and leverages them to condition future predictions, enabling long-term consistency under extended temporal horizons and large camera rotations. Without requiring additional 3D reconstruction modules, MemCam introduces a context compression module that effectively compresses historical frames selected via co-visibility, preserving scene structure across time. The overall framework is illustrated in Fig. 2. Our implementation is publicly available at https://github.com/newhorizon2005/MemCam.

Our main contributions can be summarized as follows:

- We propose MemCam, a memory-augmented framework that leverages historical frames and positional information as contextual cues to enable scene-consistent interactive video generation.
- We design a context compression module that, through a co-visibility-based context selection strategy, effectively filters and compresses historical frame information to provide rich and diverse contextual support for long-term video prediction.
- Experiments demonstrate that our method excels in long interactive video generation, significantly outperforming baseline approaches.

II. Related Work

A. Video Generation Models

Video generation models are developing rapidly, with mainstream model architectures largely based on diffusion models [19]–[21]. Early diffusion-based video generation methods primarily employed U-Net-based [22] architectures to extend image diffusion models, where temporal information was modeled using short-frame windows or temporal convolutions [23], [24]. Building on this progress, Transformer-based architectures, particularly Diffusion Transformers (DiT) [25], as adopted by recent models [1], [2], [26], replaced convolutional backbones with global self-attention mechanisms, allowing for more powerful spatiotemporal modeling and higher-quality video synthesis. Meanwhile, other architectural designs have also been explored, including hybrid attention mechanisms in video diffusion [27] and decomposed diffusion models [28]. These advances continue to improve the generative quality, temporal consistency, and controllability of video generation models.

B. Interactive Video Generation

The field of interactive video generation, where users provide control signals to guide the creation process, is also rapidly evolving. Existing models integrate various signals to enable fine-grained control over video content, such as camera trajectories [17], [29] and object motion [30], [31]. Recent work has explored the use of context guidance for interactive video generation. DFoT [16] introduces a training paradigm called "Diffusion Forcing," which independently perturbs the noise levels of different frames, allowing the model to extract conditional information from historical frames for flexible long-term video generation. GeometryForcing [18] takes a different approach by incorporating explicit 3D reconstruction as geometric constraints to enforce multi-view consistency during generation. Context-as-Memory [32] achieves scene-consistent interactive video generation without explicit 3D representations, but its code and model weights have not been publicly released.

III. Methodology

A. Context Compression Module

Directly conditioning on all historical frames is not only computationally inefficient, but may also introduce negative effects due to redundant or irrelevant information. To address this issue, we design a context compression module to aggregate contextual content. Specifically, we train a convolutional neural network to process the context portion, replacing the patchify layer of the base model. The weights of this network are initialized by extending the pre-trained model's patchify layer, ensuring meaningful outputs from the early stages of training. Given the lack of temporal correlation among historical frames, we keep the temporal dimension uncompressed, while the compression ratios for the spatial length and width dimensions are set to 2. The processed context tokens are only one-fourth the length of the unprocessed ones. For positional encoding, we use Rotary Positional Encoding (RoPE) [33], which allows variable input lengths. We keep the positional encoding for the prediction sequence consistent with the pre-training setup, assign new positional encodings to the context, and down-sample the spatial positional encodings of the context via pooling to align with the prediction sequence. Context compression enables the model to access more contextual frames simultaneously, thereby enriching the diversity of contextual information while reducing computational cost.

B. Camera Control

Camera control capability serves as the foundation for interactive video generation models. Following the work of RecamMaster [29], we add a single-layer MLP called Camera Encoder to each DiT Block. For any input sequence, we obtain camera information $\text{cam} \in \mathbb{R}^{F\times 12}$, which is derived by flattening a 3×4 matrix composed of $\mathbf{R}|\mathbf{t}$, where $\mathbf{R} \in \mathbb{R}^{3\times 3}$ is the rotation matrix, $\mathbf{t} \in \mathbb{R}^{3\times 1}$ is the translation vector, and $\mathbf{R}|\mathbf{t}$ denotes their concatenation. This camera information is mapped through the encoder to the channel dimension of the model's main feature, expanded via repetition, and then added element-wise to the main feature. The operation is formulated as:

$$\mathbf{F}_{\text{out}} = \mathbf{F}_{\text{in}} + \text{CameraEncoder}(\text{cam}), \tag{1}$$

where $\mathbf{F}_{\text{in}}$ denotes the input to the DiT Block, and $\mathbf{F}_{\text{out}}$ represents the features input to the self-attention layer after camera conditioning is incorporated.

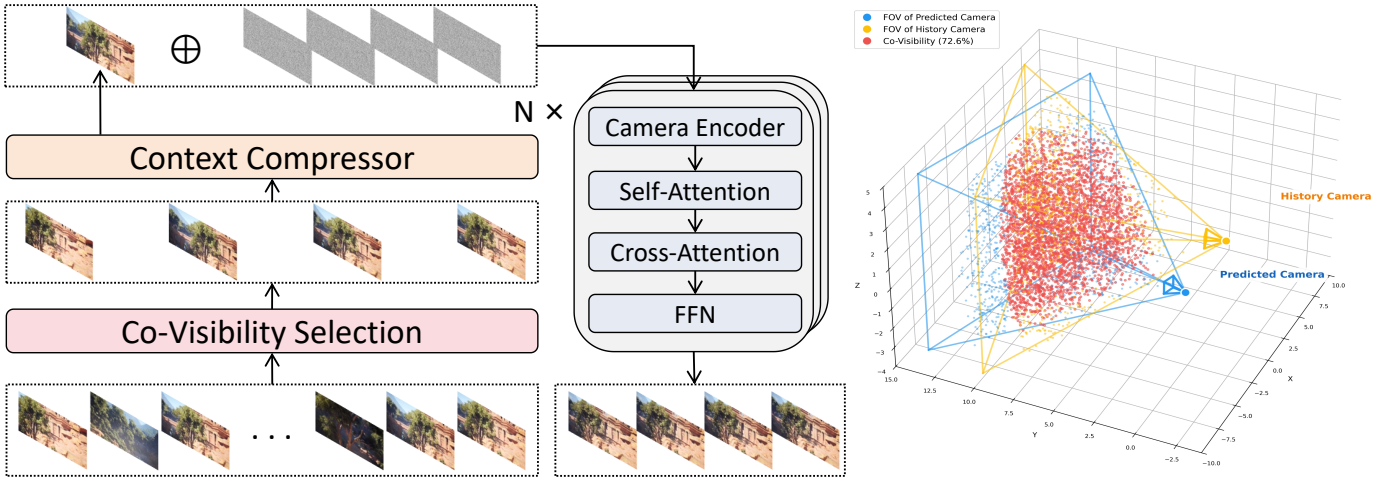

Fig. 2: **Methodology.** (Left) Overview of MemCam: the Context Compressor encodes historical frames selected via co-visibility into compact representations, which are concatenated with the noisy prediction sequence and fed into the DiT Block. (Right) Illustration of co-visibility computation between predicted and historical camera FOVs.

C. Memory Module with Context Selection

We introduce a memory module that explicitly incorporates historical frames as contextual guidance into the generation process. Specifically, we maintain a sequence of historical frames along with their corresponding camera information, called the historical sequence. For each frame in the predicted sequence, we leverage its camera information to compute the co-visibility score with every frame in the historical sequence, measuring the overlap between their respective fields of view. Based on this score, we select a subset of historical frames for each predicted frame and concatenate them as contextual input for the model.

We estimate the co-visibility between two camera poses via Monte Carlo sampling of 3D points to approximate the field-of-view (FOV) overlap. Given camera positions and orientations, the co-visibility is defined as the Intersection over Union (IoU) of the visible point sets:

$$\text{IoU}(\mathcal{C}_1, \mathcal{C}_2) = \frac{\sum_{i=1}^{N} \mathcal{V}_1(\mathbf{x}_i) \wedge \mathcal{V}_2(\mathbf{x}_i)}{\sum_{i=1}^{N} \mathcal{V}_1(\mathbf{x}_i) \vee \mathcal{V}_2(\mathbf{x}_i)} \in [0, 1], \tag{2}$$

where $\mathcal{C}_1$ and $\mathcal{C}_2$ denote the two camera configurations, N is the number of sampled points $\mathbf{x}_i$, and $\mathcal{V}_j(\mathbf{x}_i)$ indicates the visibility of point $\mathbf{x}_i$ from camera j. We set $N = 10^4$ to achieve a robust approximation.

D. Training and Inference

During training, we randomly sample a clip from a scene as the target sequence. The first frame of the target sequence serves as a fixed context frame to ensure smooth transitions between consecutive video segments. This fixed context frame is concatenated directly with the prediction sequence and jointly encoded by the 3D VAE [34], without passing through the compression module. For each frame in the prediction sequence, we assign one context frame selected from the remaining frames in the scene based on co-visibility: any frame with non-zero overlap is eligible for selection. These selected context frames are individually encoded by the 3D VAE and then processed through our context compression module. To support image-to-video generation, with 10% probability, we zero-pad all the selected context to simulate scenarios where no historical frames are available.

During inference, we adopt a segment-wise generation approach. Unlike training, inference selects the context frame with the highest co-visibility score for each predicted frame from historical frames, ensuring the most relevant content is retrieved. The memory is updated after each segment is generated, and this process iterates until the full video is completed.

IV. Experiment

A. Experiment Settings

Implementation Details. Our method is built on the Wan2.1 1.3B Text-to-Video Diffusion Transformer [1]. We train on the Context-as-Memory dataset [32], containing 100 scenes with 7901 frames each. Each training instance consists of 77 frames at 640×352 resolution: the first frame serves as fixed context, and the remaining 76 frames form the prediction sequence. For each predicted frame, one context frame is selected from the other frames in the scene based on co-visibility. The fixed context and prediction sequence are jointly encoded by the 3D VAE into 20 latent frames, while the 76 selected context frames are individually encoded and processed through our compression module. All model parameters are trainable. The model was trained for 20,000 iterations using AdamW [35] with learning rate 1×10^{-5}, batch size 16, on 2 NVIDIA H20 GPUs. During inference, we use 50 denoising steps. We evaluate on 5% of Context-as-Memory scenes as test set, and additionally on RealEstate10K [36] for zero-shot generalization.

Evaluation Metrics. To effectively evaluate the memory capacity of the video generation model, we adopt the following

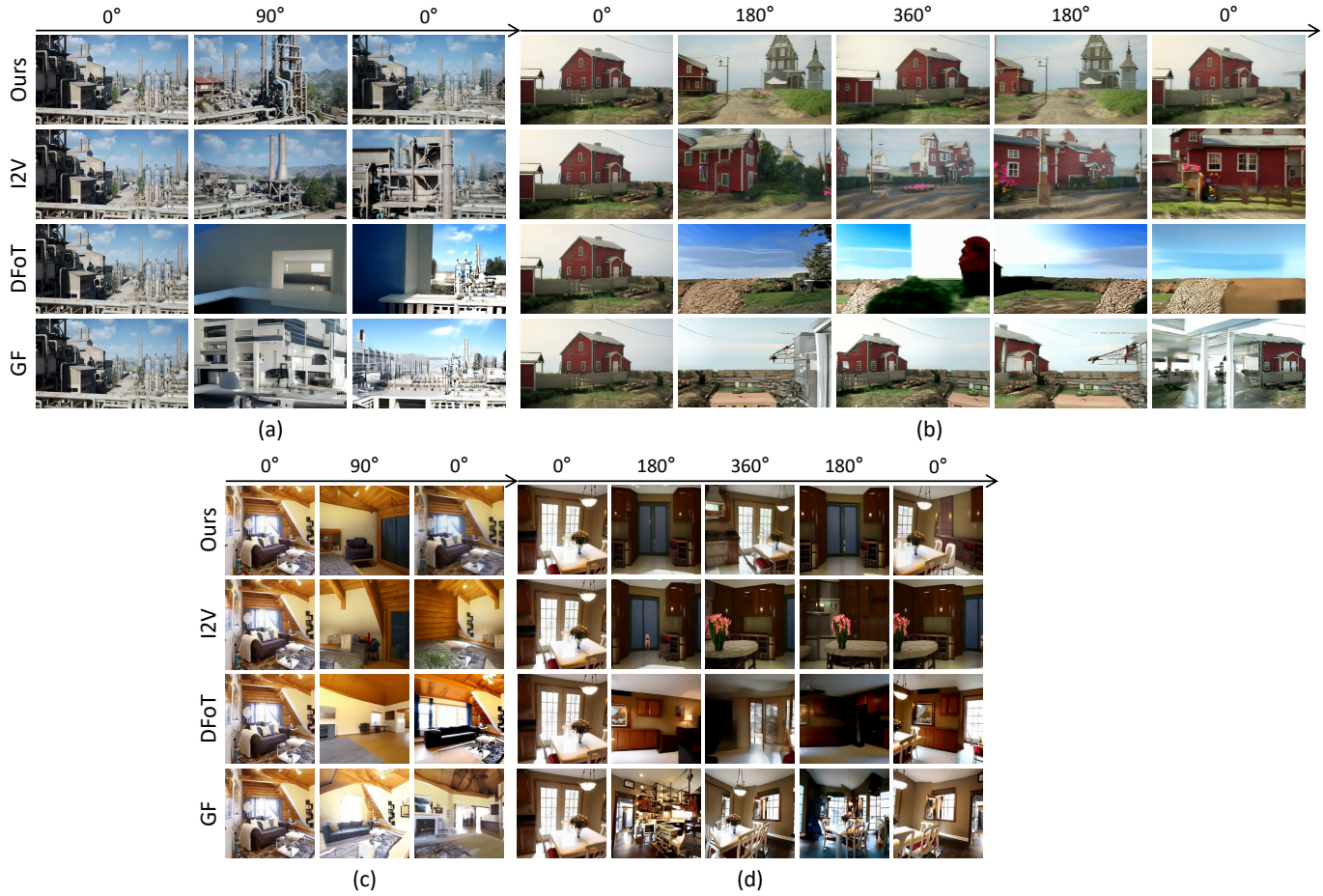

Fig. 3: **Qualitative Comparison Results.** (a) and (b) are evaluated on the Context-as-Memory dataset, and (c) and (d) on RealEstate10K. MemCam achieves superior performance in scene memory retention and overall generation quality. In contrast, other methods exhibit varying degrees of scene inconsistency due to insufficient utilization of contextual information.

two assessment methods: (1) 90-degree round-trip generation: A video of $1 + 76 \times 2 = 153$ frames is generated, where the camera trajectory first rotates 90 degrees to the right and then rotates 90 degrees to the left back to the origin. (2) 360-degree round-trip generation: A video of $1 + 76 \times 8 = 609$ frames is generated, where the camera first completes a full 360-degree clockwise rotation and then a full 360-degree counterclockwise rotation, both returning to the starting viewpoint. These two methods effectively test the memory capacity of the model. Each generated video can be viewed as two sequences. Ideally, if one sequence is temporally reversed, it should be identical to the other sequence. We quantify this difference by using the PSNR, SSIM [37], and LPIPS [38] metrics and employ FVD [39] to assess overall video quality.

Comparison Methods. We compare MemCam with the following methods: (1) Using only the first frame as context (I2V): implemented on our base model with the same training setup as MemCam; (2) Diffusion Forcing Transformer (DFoT) [16]: based on a fixed-length context window; (3) Geometry-Forcing (GF) [18]: based on 3D representation reconstruction.

B. Main Results

Quantitative Results. As shown in Table I, MemCam achieves competitive or superior performance across most metrics, with particularly significant gains in the 360° scenario where longer duration and larger camera rotation pose greater challenges. The I2V baseline, lacking historical memory, generally performs poorly. DFoT, constrained by its limited context window, fails to retain early content in long sequences. GF achieves competitive PSNR/SSIM on 90° tasks, but its performance drops notably in 360° scenarios as reconstruction errors accumulate. MemCam achieves the best FVD across all settings, indicating superior temporal consistency and visual quality. The results show that effective utilization of historical frames through compression and retrieval is crucial for memory retention and generation quality.

Qualitative Results. As shown in Fig. 3, MemCam faithfully preserves the scene structure even after large camera rotations, while other methods exhibit noticeable drift and content hallucination. When the camera returns to the starting viewpoint, MemCam reconstructs the original scene appearance, whereas baselines produce inconsistent textures or altered layouts.

TABLE I: **Quantitative Comparison Results.** We evaluate on two datasets with 90° and 360° round-trip benchmarks. ↑ means higher is better, ↓ means lower is better. Best results are in **bold**, second best are underlined.

Context-as-Memory Dataset								
Methods	90° Round-trip				360° Round-trip			
	PSNR↑	SSIM↑	LPIPS↓	FVD↓	PSNR↑	SSIM↑	LPIPS↓	FVD↓
I2V	15.81	0.452	0.470	528.51	9.75	0.332	0.603	988.82
DFoT	16.76	0.474	0.393	683.59	8.94	0.252	0.613	1188.34
GF	16.57	0.486	**0.348**	557.66	10.07	0.402	0.565	852.05
MemCam (Ours)	**17.83**	**0.506**	0.357	**215.71**	**14.81**	**0.423**	**0.504**	**167.87**
RealEstate10K Dataset								
Methods	90° Round-trip				360° Round-trip			
	PSNR↑	SSIM↑	LPIPS↓	FVD↓	PSNR↑	SSIM↑	LPIPS↓	FVD↓
I2V	16.26	0.488	0.362	552.01	10.16	0.284	0.611	789.62
DFoT	17.17	0.505	0.399	539.89	10.43	0.281	0.564	1002.39
GF	**17.70**	**0.597**	0.316	519.78	11.19	0.379	0.405	419.60
MemCam (Ours)	17.61	0.544	**0.314**	**269.82**	**16.52**	**0.550**	**0.400**	**131.96**

Sufficient context conditioning not only strengthens memory retention, but also suppresses error accumulation in long-range generation.

C. Ablation Study

All ablation experiments are conducted on the Context-as-Memory dataset.

Context Selection Strategy. We compare different strategies for selecting context frames. "Recent" selects the most recent historical frames; "Random" randomly samples from historical frames; "TopK" aggregates all frames that overlap with any frame in the prediction sequence, ranks them by total occurrence count, and selects the top-ranked ones. As shown in Table II, our selection strategy consistently outperforms all alternatives. On the 90° benchmark, Recent performs comparably to Ours since the limited camera rotation ensures that recent frames remain spatially relevant. On the 360° benchmark, Recent degrades dramatically as it cannot retrieve content from distant viewpoints. Random, despite lacking targeted selection, outperforms Recent because it can access frames from earlier parts of the video that cover distant viewpoints. TopK tends to favor frames from the middle of the sequence, resulting in uneven coverage. Our method achieves the best results by providing both relevance and uniform coverage through per-frame dynamic selection.

Context Compression Module. We evaluate the context compression module on the 360° round-trip benchmark by varying total context length (76, 38, or 19 frames, i.e., one context frame per 1, 2, or 4 predicted frames). "None" denotes direct concatenation without compression, while "Ours" applies our compression module. We report inference time as seconds per frame (s/frame). As shown in Table III, without compression, increasing context length yields marginal quality gains but incurs substantial computational overhead. With compression, longer context provides richer information that better tolerates the compact encoding, leading to improved performance. Notably, Ours-76 attains similar quality to None-76 while being nearly 5× faster, and even outperforms None-19 at comparable time cost. These results demonstrate that context compression effectively enables richer context without prohibitive computational burden.

TABLE II: **Ablation on Context Selection Strategy.**

90° Round-trip				
Methods	PSNR↑	SSIM↑	LPIPS↓	FVD↓
Recent	17.80	0.498	0.375	231.75
Random	17.36	0.502	0.400	237.87
TopK	17.39	0.497	0.394	256.88
Ours	**17.83**	**0.506**	**0.357**	**215.71**
360° Round-trip				
Methods	PSNR↑	SSIM↑	LPIPS↓	FVD↓
Recent	10.26	0.340	0.581	915.41
Random	12.13	0.417	0.511	238.98
TopK	14.67	0.414	0.506	262.52
Ours	**14.81**	**0.423**	**0.504**	**167.87**

TABLE III: **Ablation on Context Compression Module.**

Methods	PSNR↑	SSIM↑	LPIPS↓	FVD↓	s/frame↓
None-19	14.69	0.417	0.513	184.42	4.45
None-38	14.76	0.424	0.494	179.29	8.82
None-76	14.88	0.430	0.489	164.37	22.15
Ours-19	13.88	0.398	0.534	203.30	2.14
Ours-38	14.61	0.407	0.527	182.99	2.81
Ours-76	14.81	0.423	0.504	167.87	4.47

V. Conclusion

In this work, we propose MemCam, a memory-augmented framework for scene-consistent interactive video generation that treats previously generated frames as external memory. We design a context compression module that encodes historical frames into compact representations, enabling the model to incorporate richer contextual information while significantly reducing computational cost. Combined with a co-visibility-based context selection strategy, our framework effectively retrieves the most relevant historical frames for each predicted

frame, enabling faithful scene reconstruction even under large camera rotations. Experiments on two datasets demonstrate that MemCam significantly outperforms existing methods, particularly in long video scenarios with large viewpoint changes.

Limitations and Future Work. The current inference speed is relatively slow due to the computational overhead of bidirectional attention. In future work, we plan to explore diffusion distillation for acceleration and scale up training with larger datasets to further improve generation quality.

Acknowledgment

This work was supported by the National Innovation Training Program for College Students of Guilin University of Electronic Technology (Grant No. 202510595051).

References

[1] T. Wan, A. Wang, B. Ai, B. Wen, C. Mao, C.-W. Xie, D. Chen, F. Yu, H. Zhao, J. Yang *et al.*, "Wan: Open and advanced large-scale video generative models," *arXiv preprint arXiv:2503.20314*, 2025.

[2] W. Kong, Q. Tian, Z. Zhang, R. Min, Z. Dai, J. Zhou, J. Xiong, X. Li, B. Wu, J. Zhang *et al.*, "Hunyuanvideo: A systematic framework for large video generative models," *arXiv preprint arXiv:2412.03603*, 2024.

[3] X. Yang and X. Wang, "Compositional video generation as flow equalization," *arXiv preprint arXiv:2407.06182*, 2024.

[4] F. Bao, C. Xiang, G. Yue, G. He, H. Zhu, K. Zheng, M. Zhao, S. Liu, Y. Wang, and J. Zhu, "Vidu: a highly consistent, dynamic and skilled text-to-video generator with diffusion models," *arXiv preprint arXiv:2405.04233*, 2024.

[5] J. Yu, Y. Qin, X. Wang, P. Wan, D. Zhang, and X. Liu, "Gamefactory: Creating new games with generative interactive videos," *arXiv preprint arXiv:2501.08325*, 2025.

[6] D. Valevski, Y. Leviathan, M. Arar, and S. Fruchter, "Diffusion models are real-time game engines," *arXiv preprint arXiv:2408.14837*, 2024.

[7] R. Li, P. Torr, A. Vedaldi, and T. Jakab, "Vmem: Consistent interactive video scene generation with surfel-indexed view memory," *arXiv preprint arXiv:2506.18903*, 2025.

[8] A. Hu, L. Russell, H. Yeo, Z. Murez, G. Fedoseev, A. Kendall, J. Shotton, and G. Corrado, "Gaia-1: A generative world model for autonomous driving," *arXiv preprint arXiv:2309.17080*, 2023.

[9] L. Russell, A. Hu, L. Bertoni, G. Fedoseev, J. Shotton, E. Arani, and G. Corrado, "Gaia-2: A controllable multi-view generative world model for autonomous driving," *arXiv preprint arXiv:2503.20523*, 2025.

[10] Z. Xiao, Y. Lan, Y. Zhou, W. Ouyang, S. Yang, Y. Zeng, and X. Pan, "Worldmem: Long-term consistent world simulation with memory," *arXiv preprint arXiv:2504.12369*, 2025.

[11] J. Gao, Z. Chen, X. Liu, J. Feng, C. Si, Y. Fu, Y. Qiao, and Z. Liu, "Longvie: Multimodal-guided controllable ultra-long video generation," *arXiv preprint arXiv:2508.03694*, 2025.

[12] J. Gao, Z. Chen, X. Liu, J. Zhuang, C. Xu, J. Feng, Y. Qiao, Y. Fu, C. Si, and Z. Liu, "Longvie 2: Multimodal controllable ultra-long video world model," *arXiv preprint arXiv:2512.13604*, 2025.

[13] S. Yang, W. Huang, R. Chu, Y. Xiao, Y. Zhao, X. Wang, M. Li, E. Xie, Y. Chen, Y. Lu *et al.*, "Longlive: Real-time interactive long video generation," *arXiv preprint arXiv:2509.22622*, 2025.

[14] L. Zhang and M. Agrawala, "Packing input frame context in next-frame prediction models for video generation," *arXiv preprint arXiv:2504.12626*, 2025.

[15] E. Decart, Q. McIntyre, S. Campbell, X. Chen, and R. Wachen, "Oasis: A universe in a transformer," *URL: https://oasis-model.github.io*, 2024.

[16] K. Song, B. Chen, M. Simchowitz, Y. Du, R. Tedrake, and V. Sitzmann, "History-guided video diffusion," *arXiv preprint arXiv:2502.06764*, 2025.

[17] H. He, Y. Xu, Y. Guo, G. Wetzstein, B. Dai, H. Li, and C. Yang, "Cameractrl: Enabling camera control for text-to-video generation," *arXiv preprint arXiv:2404.02101*, 2024.

[18] H. Wu, D. Wu, T. He, J. Guo, Y. Ye, Y. Duan, and J. Bian, "Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling," *arXiv preprint arXiv:2507.07982*, 2025.

[19] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[20] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," *arXiv preprint arXiv:2210.02747*, 2022.

[21] X. Liu, C. Gong, and Q. Liu, "Flow straight and fast: Learning to generate and transfer data with rectified flow," *arXiv preprint arXiv:2209.03003*, 2022.

[22] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015. [Online]. Available: https://arxiv.org/abs/1505.04597

[23] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, "Video diffusion models," 2022. [Online]. Available: https://arxiv.org/abs/2204.03458

[24] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni, D. Parikh, S. Gupta, and Y. Taigman, "Make-a-video: Text-to-video generation without text-video data," 2022. [Online]. Available: https://arxiv.org/abs/2209.14792

[25] W. Peebles and S. Xie, "Scalable diffusion models with transformers," 2023. [Online]. Available: https://arxiv.org/abs/2212.09748

[26] Z. Yang, J. Teng, W. Zheng, M. Ding, S. Huang, J. Xu, Y. Yang, W. Hong, X. Zhang, G. Feng, D. Yin, Y. Zhang, W. Wang, Y. Cheng, B. Xu, X. Gu, Y. Dong, and J. Tang, "Cogvideox: Text-to-video diffusion models with an expert transformer," 2025. [Online]. Available: https://arxiv.org/abs/2408.06072

[27] W. Wang, H. Yang, Z. Tuo, H. He, J. Zhu, J. Fu, and J. Liu, "Swap attention in spatiotemporal diffusions for text-to-video generation," 2024. [Online]. Available: https://arxiv.org/abs/2305.10874

[28] Z. Luo, D. Chen, Y. Zhang, Y. Huang, L. Wang, Y. Shen, D. Zhao, J. Zhou, and T. Tan, "Videofusion: Decomposed diffusion models for high-quality video generation," 2023. [Online]. Available: https://arxiv.org/abs/2303.08320

[29] J. Bai, M. Xia, X. Fu, X. Wang, L. Mu, J. Cao, Z. Liu, H. Hu, X. Bai, P. Wan *et al.*, "Recammaster: Camera-controlled generative rendering from a single video," *arXiv preprint arXiv:2503.11647*, 2025.

[30] Z. Wang, Z. Yuan, X. Wang, T. Chen, M. Xia, P. Luo, and Y. Shan, "Motionctrl: A unified and flexible motion controller for video generation," 2024. [Online]. Available: https://arxiv.org/abs/2312.03641

[31] R. Feng, H. Zhang, Z. Yang, J. Xiao, Z. Shu, Z. Liu, A. Zheng, Y. Huang, Y. Liu, and H. Zhang, "The matrix: Infinite-horizon world generation with real-time moving control," 2024. [Online]. Available: https://arxiv.org/abs/2412.03568

[32] J. Yu, J. Bai, Y. Qin, Q. Liu, X. Wang, P. Wan, D. Zhang, and X. Liu, "Context as memory: Scene-consistent interactive long video generation with memory retrieval," *arXiv preprint arXiv:2506.03141*, 2025.

[33] J. Su, Y. Lu, S. Pan, A. Murtadha, B. Wen, and Y. Liu, "Roformer: Enhanced transformer with rotary position embedding," 2023. [Online]. Available: https://arxiv.org/abs/2104.09864

[34] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[35] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2019. [Online]. Available: https://arxiv.org/abs/1711.05101

[36] T. Zhou, R. Tucker, J. Flynn, G. Fyffe, and N. Snavely, "Stereo magnification: Learning view synthesis using multiplane images," 2018. [Online]. Available: https://arxiv.org/abs/1805.09817

[37] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[38] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," 2018. [Online]. Available: https://arxiv.org/abs/1801.03924

[39] T. Unterthiner, S. van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, "Towards accurate generative models of video: A new metric & challenges," 2019. [Online]. Available: https://arxiv.org/abs/1812.01717