FreeMem: Enhancing Consistency in Long Video Generation via Tuning-Free Memory

Jibin Peng[1*], Di Lin[1*], Zhecheng Xu[1], Haoran Lu[1], Ruonan Liu[2†], Wuyuan Xie[3], Miaohui Wang[3], Lingyu Liang[4], Yi Wang[3], Qing Guo[5]

[1]Tianjin University,
[2]Shanghai Jiao Tong University,
[3]Shenzhen University,
[4]South China University of Technology,
[5]Nankai University,
jibinpeng@tju.edu.cn, ande.lin1988@gmail.com

Abstract

Text-to-Video (T2V) generation has advanced greatly, yet maintaining consistency remains challenging, especially for tuning-free long video generation. We attribute the consistency problem to cumulative deviations for long video generation at three levels: the random noise lacking correlation results initial deviation between frames; discrepancy in semantic feature tokens between denoising network blocks gradually accumulates as the frame count grows, leading to greater deviations; attention mechanisms struggle to capture global relationships across distant frames in long videos. To address these, we propose FreeMem, a tuning-free framework leveraging hierarchical memory update and injection: the noise memory stabilizes consistency by manipulating low and high frequency components in the initial noise space; the token memory combats inconsistency through adaptive fusion of historical and current semantic feature tokens between denoising network blocks; and the attention memory establishes persistent cache to model long-range relationships within self attention layers. Evaluated on VBench, FreeMem improves subject and background consistency matrics across various methods, offering a practical solution for low-cost, high-consistency long video generation.

Introduction

With the development of text-to-video (T2V) diffusion models (Chen et al. 2024; Guo et al. 2024; Wang et al. 2023c; Zheng et al. 2024b; Polyak et al. 2024; Yang et al. 2024b; Deng et al. 2024; HaCohen et al. 2024; Yin et al. 2025; Lin et al. 2025; Li et al. 2024d), T2V has gained significant interest, showing extensive potential applications in film production, advertising, and education.

The videos from T2V should have three main characteristics: controllability, diversity and consistency. Controllability denotes the video's capacity to faithfully align with user-defined attributes, enabled by conditional inputs such as text prompts, control signals, or reference frames. Diversity refers to the videos' diverse styles and contents, typically achieved through randomness and style transfer techniques. Consistency represents the video's coherence across video frames, including appearance consistency, typically achieved by temporal modules in diffusion.

While diversity and controllability have seen significant advancements, the problem of consistency has become prominent. High consistency requires strong appearance consistency, meaning that frames must maintain uniformity in style and object features to avoid visual abruptness or disharmony, especially in long videos. The generation of long videos is crucial in long-form content creation and storytelling. But training-based long video generation works require substantial resources, heavily relying on computational power and large-scale data. Thus many works have shifted to tuning-free approaches, leveraging existing short video generation models to generate long videos through autoregressive strategies or another, which are cost-effective and easy to implement. However, in autoregressive tuning-free approaches, consistency deteriorates as video length increases, as illustrated in Figure 1. The key point of the consistency problem of tuning-free long videos lies in the cumulative deviation between frames. For long videos, subtle visual discrepancies of adjacent frames will accumulate, resulting in significant reduction in consistency between the early and later frames. This phenomenon can be seen from Figure 1 where the similarity curve compared to the first frame decreases as the video length increases.

The cumulative deviation in long videos has multifaceted nature. Specifically: (1) The multiple random initial noise independently sampled from standard Gaussian distribution are inherently different and uncorrelated, resulting the initial deviation between frames. (2) Between the blocks of the denoising network, discrepancy in the representation of identical semantic feature tokens across frames leads to challenges in maintaining consistent semantic alignment. As the number of generated frames increase, the cumulative discrepancy in cross-frame semantic features accumu-

*These authors contributed equally.
†Corresponding author.

Figure 1: The visualization of the cumulative deviation. As the number of frame increases, the consistency gradually decreases as the color of the parrot's wings and body does not match. The similarity curve computed by DINO (Caron et al. 2021) shows that consistency between frames declines as video length increases. Therefore, we propose FreeMem to reduce the cumulative deviation and improve consistency between frames in long video generation.

lates progressively, thereby increasing deviation. (3) Attention mechanisms predominantly focus on local interactions within limited window sizes, and temporal attention struggle to capture long-range relationships across distant frames in long videos, resulting deviations between distant frames with weak relationships. The above three factors contribute to the cumulative deviation as video length increases, reducing the consistency.

To address the problem, we propose FreeMem, employing multi-level memory modules to store and update features from previous frames, and inject them into subsequent frames to reduce the deviation in long video generation. As illustrated in Figure 2(a), FreeMem is a tuning-free framework that can be seamlessly integrated into various long video generation works, offering a convenient and effective solution to enhance consistency.

FreeMem is designed at three distinct levels: initial input noise, feature tokens between blocks of the denoising network and attention tensors in attention layers. This multi-level design directly corresponds to the hierarchical structure of diffusion models shown in Figure 2(b): (1) Noise level operates at the input space of the denoising process. Noise memory with Noise Memory Process specifically manipulates and rearranges the low and high frequency parts of this latent space to improve the correlation between frames and stabilize consistency. (2) Token level intervenes at the intermediate feature space of the denoising network blocks. Token memory with Token Memory Block dynamically merges redundant similar tokens across frames while injecting previous semantic features, reducing feature discrepancy between distant frames. (3) Attention level acts on the relational space within self attention layers. Attention memory with Memory Self Attention maintains the persistent cache of historical attention features, enabling explicit modeling of long-range relationships.

Experiments across multiple methods and frame lengths demonstrate significant consistency improvements in long videos under the VBench (Huang et al. 2024) benchmark.

Related Work

Long Video Generation With the rapid development of video diffusion models, numerous models (Guo et al. 2024; Wang et al. 2023c,b; Chen et al. 2024) have emerged. Recent advances further extend video length generation (Zheng et al. 2024b; Polyak et al. 2024; Yang et al. 2024b; Deng et al. 2024; HaCohen et al. 2024; Lin et al. 2025; Li et al. 2024d), yet their training remains computationally expensive.

Some works have attempted to extend video length using current models through tuning-free strategies. ExVideo (Duan et al. 2024) extends Stable Video Diffusion (Blattmann et al. 2023) to generate up to 128 frames. Gen-L-Video (Wang et al. 2023a), CoNo (Wang, Li, and Chen 2024) and Confiner2025 (Li et al. 2024b) employ post-processing methods to combine multiple short video segments into long videos. Video-Infinity (Tan et al. 2024) can generate hundreds of frames through parallel processing but it relies heavily on computational resources. FreeLong (Lu et al. 2024) and GLC-Diffuison (Ma et al. 2025) extend the length by modeling the global context and local context, respectively. FreeNoise (Qiu et al. 2024) fuses cross-frame features by sliding window mechanism to gererate longer videos. FIFO-Diffusion (Kim et al. 2024) and StreamingT2V (Henschel et al. 2025) adopt autoregressive approaches, generating frames sequentially in a time-ordered or block-by-block manner, which inevitably suffer from cumulative deviations affecting consistency.

Video Consistency For consistency related works, they can be divided into training-based works and tuning-free works. Training-based works (Si et al. 2025; Li et al. 2024a; Yang et al. 2023; Dong et al. 2025; Yang et al. 2024a; Ge et al. 2023) enhance consistency in diffusion by designing noise process, training strategies, cache modules, which inevitably require high-quality training and large computation.

For tuning-free works, they can be divided into attention-based and noise-based methods. Attention-based methods (Fan et al. 2024; Zhou et al. 2024) like UniCtrl (Chen, Xia, and Xu 2024) and VideoTetris (Tian et al. 2024), fuse reference-frame or cross-frame features through attention mechanisms. Ouroboros-Diffusion (Chen et al. 2025) further enhances consistency via Subject-Aware Cross-Frame Attention, while TiARA (Li et al. 2024c) reweights temporal attention scores through time-frequency analysis. Noise-based methods leverage noise manipulation to improve consistency. ConsistI2V (Ren et al. 2024) introduces low-frequency noise initialization in the first frame alongside spatio-temporal attention, while FreeInit (Wu et al. 2024) unifies low-frequency information in the initial noise space.

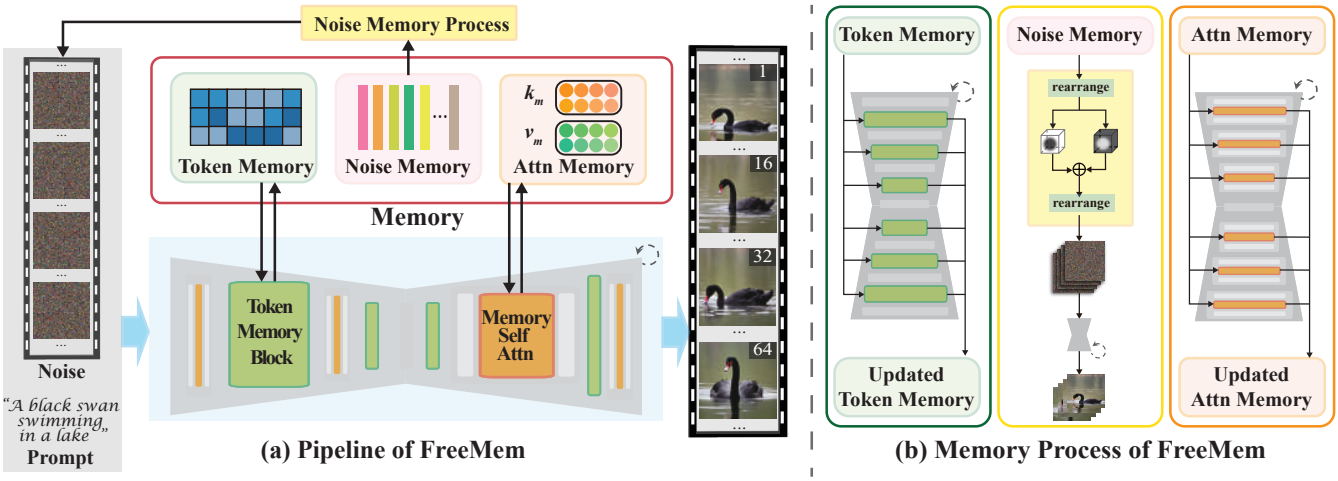

Figure 2: (a) The pipeline of video generation with FreeMem. It takes a text prompt and initialized noise as input to produce video frames while interacting with memory. Specifically, the token memory interacts through Token Memory Block between the adjacent blocks of denoising network, the attention memory interacts through Memory Self Attention in attention layers, and the noise memory interacts through Noise Memory Process. (b) The memory process of FreeMem. The token and attention memory are integrated into the generation process to improve consistency in long video generation, while also updating the corresponding memory. The noise memory is used to generate the initial noise for each frame.

Similarly, RAVE (Kara et al. 2024) improves consistency by applying randomized noise shuffling. But they are mostly for short videos or rely on reference images, reducing diversity.

Memory Mechanisms Memory has been widely used in tasks related to time sequences. In image synthesis, DM-GAN (Zhu et al. 2019) employs dynamic memory modules to iteratively refine image details by adaptively fusing the features from memory and images. For video processing, MANA (Yu et al. 2022) designs a memory-augmented attention module to memorize general video details for video super-resolution. Recent efforts extend memory to long-term scenarios: SLOWFAST-VGEN (Hong et al. 2024) encodes episodic memory in LoRA parameters for action-driven video synthesis, while MEMO (Zheng et al. 2024a) ensures identity consistency in talking head generation through memory-guided temporal module. MA-LMM (He et al. 2024) proposes a memory bank to autoregressively store and accumulate past features, enhancing the capability of large multi-modal models in long video understanding tasks.

With memory mechanisms' inherent capability to preserve long-term features, we construct memory modules to reduce cumulative deviations in long video generation.

Details of FreeMem

Overview

The overall pipeline is illustrated in Figure 2(a). Given a text prompt and initialized noise, the model generates the long video through a diffusion-based process. FreeMem systematically preserves and updates historical features in the generation process. Shown in Figure 2(b), the memory process of FreeMem consists of noise memory with Noise Memory Process, token memory with Token Memory Block and attention memory with Memory Self Attention.

The overall memory M is defined as the combination of M_{noise} , M_{token} , M_{attn}:

$$\begin{aligned} M_{\text{noise}} &= \{\eta^i \mid i = 1, 2, \ldots, p\}, \\ M_{\text{token}} &= \{\varphi_t^b | t \in [0, T], b \in [0, B]\}, \\ M_{\text{attn}} &= \{\text{cache}_t^l(k_m, v_m) | t \in [0, T], l \in [0, L]\}. \end{aligned} \tag{1}$$

M_{noise} is the noise memory storing the fixed noise whose low-frequency part is used to initialize the noise for each frame to improve the correlation between frames. M_{token} is the token memory storing the feature tokens between blocks of the denoising network to reduce feature discrepancy in the long run. M_{attn} is the attention memory storing the key-value pairs in the attention layers to model long-range relationships. p is the number of noise in the noise memory, T is the total number of time steps, L is the total number of layers related to attention, and B is the total number of blocks of the denoising network. $\text{cache}_t^l(k_m, v_m)$ is the key-value pairs stored in the cache for attention memory at the t step in the l layer. Similarly, φ_t^b is the feature tokens in token memory at the t step in the b block. FreeMem effectively reduces cumulative deviations with multi-level memory modules.

Noise Memory

Noise serves as the initial input of denoising networks. FreeInit (Wu et al. 2024) has shown that the low-frequency part of noise is crucial in video quality, which retain spatio-temporal correlations during the denoising process. Noise memory leverages the low-frequency components from noise memory M_{noise} to construct temporally correlated noise frames, while integrating high-frequency information from random noise to preserve several variations.

Rearrange Rearrange step shuffles noise block by block to extend length, which maintains randomness and diver-

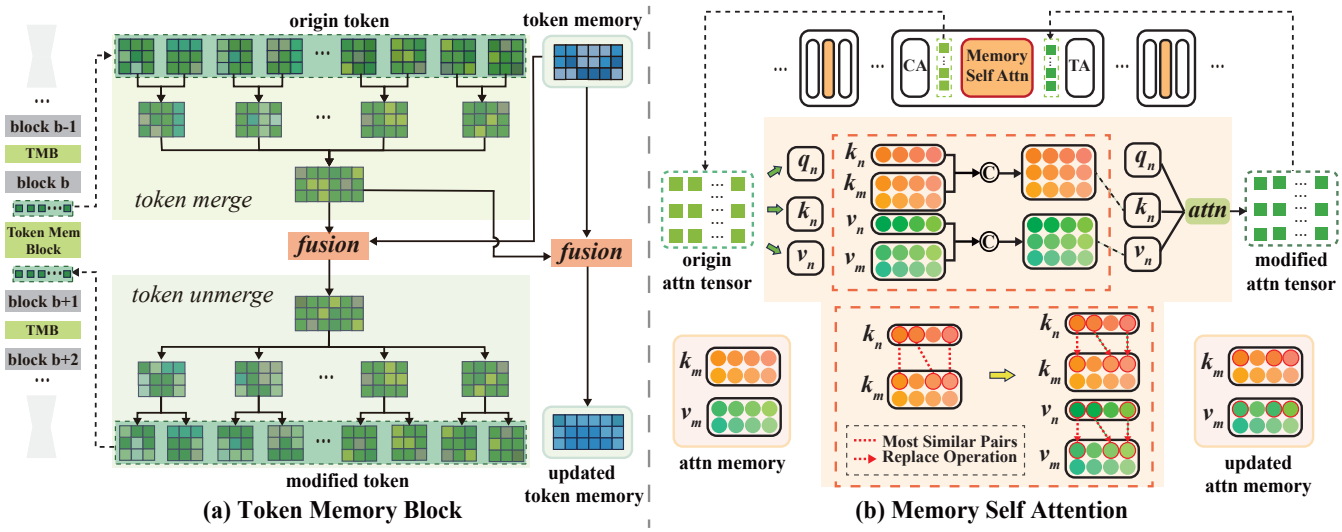

Figure 3: (a) The Token Memory Block is responsible for updating token memory and refining original token. The origin token from the block is merged into local token via token merge, fused with the token memory to update both, and then unmerged via token unmerge to get the modified token as input for the next block. (b) Memory Self Attention is an adaptation of the original self attention mechanism, incorporating attention memory. The key-value pairs from the origin tensor are concatenated with the attention memory to compute the modified attn tensor via the attention mechanism. Meanwhile, the most similar key-value pairs are selected to update the attention memory.

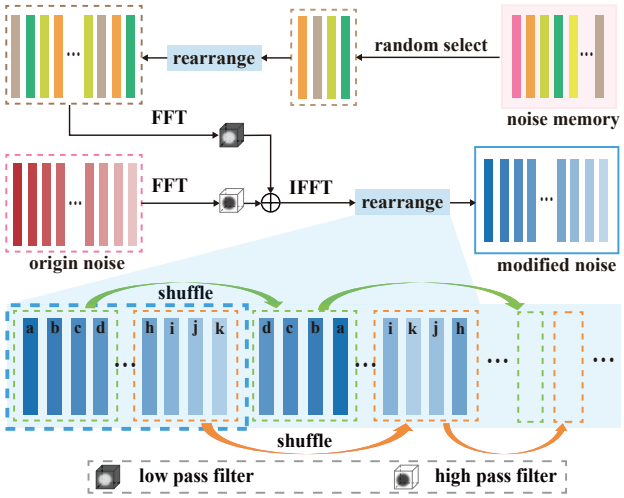

Figure 4: The illustration of the Noise Memory Process. Small rectangles represent individual noise. The rearrange step extends the length by shuffling noise block by block.

sity while preserving overall characteristics. As shown in Figure 4, it segments the initial noise sequence into small blocks, randomly shuffles every noise within each block, and appends them to the end of the noise sequence. It can be formulated as follows:

$$N^b = \mathcal{R}(N^a, a, b), \quad \text{s.t.} \quad a \mid b. \tag{2}$$

It means that the noise $N^a = \{\eta_1, \eta_2, \ldots, \eta_a\}$ is rearranged to the noise $N^b = \{\hat{\eta_1}, \hat{\eta_2}, \ldots, \hat{\eta_b}\}$. Rearrange step is implemented twice to extend length for long videos, ensuring the long-term correlation.

Noise Memory Process It transforms the random original noise N^f with f frames, and the noise M^p with p frames from noise memory, into modified noise N^s with s frames. First, M^p is randomly selected from M_{noise}, and then rearranged to M^f with f frames. Next, the low-frequency part of M^f is combined with the high-frequency part from N^f to form hybrid noise N^f_*. Finally, N^f_* is rearranged to extend the frame sequence, generating the modified noise N^s as the input for long video generation. It can be formulated as follows:

$$\begin{aligned} M^f &= \mathcal{R}(M^p, p, f), \\ N^f_* &= \text{IFFT}(\text{FFT}(M^f) \odot \mathcal{F} + \text{FFT}(N^f) \odot (1 - \mathcal{F})), \\ N^s &= \mathcal{R}(N^f_*, f, s), \end{aligned} \tag{3}$$

where FFT and IFFT are the 3D Fast Fourier Transform and Inverse Fast Fourier Transform, $\mathcal{F}$ is the Low Pass Filter, $1 - \mathcal{F}$ is the High Pass Filter.

The first rearrange shares low-frequency part of p frames in the noise memory, while the second rearrange shares the fixed noise of f frames from the first rearrange. Together, they not only expand the number of frames but also enhance cross-frame constraints, promote information sharing, and preserve necessary variations.

Token Memory

Fusion The fusion operation is designed to preserve and fuse similar semantic feature tokens. Given the token sets $X = (x_1, x_2, \ldots, x_n)$ and $Y = (y_1, y_2, \ldots, y_n)$, the fusion operation to fuse the top K tokens. First, the cosine similarity matrix $S \in \mathbb{R}^{(n,n)}$ between X and Y is calculated and

then the top K pairs are selected. Finally, take the average of the selected pairs and update them back to X. X_* is the final output. It can be formulated as follows:

$$S[i,j] = \frac{x_i \cdot y_j}{\|x_i\|\|y_j\|}, \quad i = 1, \cdots, n; j = 1, \cdots, n,$$

$$\{(i_r, j_r)\}_{r=1}^{K} = \underset{(i,j)}{\text{top } K \arg\max} S[i,j],$$

$$X_*[l] = \begin{cases} \frac{1}{1+|\mathcal{V}_l|}\left(x_l + \sum_{k \in \mathcal{V}_l} y_{j_k}\right), & \text{if } l \in \{i_r\}_{r=1}^{K} \\ x_l, & \text{otherwise,} \end{cases} \quad (4)$$

where $\mathcal{V}_l = \{k | i_k = l\}$. Fusion considers the total contribution of Y similar to X, which can be defined as:

$$X_* = \mathcal{I}(X, Y, K). \quad (5)$$

Token Memory Block The denoising network can be regarded as a hierarchical composition of multiple blocks, where the feature tokens extracted from each block encode semantic information. We design the Token Memory Block (TMB) to be placed during adjacent blocks, allowing it to process the feature tokens between them.

TMB is divided into the token user and the token updater. The token user refines the origin token from the last block by fusing it with the previous semantic feature token stored in the token memory, producing modified token as input for the next block. While the token updater dynamically updates the token memory with the origin token to facilitate subsequent frame generation.

The token user applies the token merge operation to aggregate feature from the origin token χ_t^b at the t step in the b block into local token, reducing redundant similar features and computation. Specifically, χ_t^b that has f frames is partitioned into chunks with step size s. Tokens in each chunk with s frames are merged through the bipartite soft matching algorithm of ToMe (Bolya et al. 2023), which can reduce the redundancy and token numbers. After recursive merging, χ_t^b is reduced to the local token. Figure 3(a) deltails the process of token merge. Then token memory φ_t^b is fused into the local token by fusion operation, allowing the local token to incorporate features from previous frames and mitigate deviations. Finally, the token unmerge operation decomposes the local token back into its original position and size to get the modified token $\hat{\chi}_t^b$. It can be defined as:

$$\hat{\chi}_t^b = \mathcal{U}(\mathcal{I}(\mathcal{M}(\chi_t^b), \varphi_t^b, K)), \quad (6)$$

where $\mathcal{M}(\cdot)$ and $\mathcal{U}(\cdot)$ are the token merge and unmerge operations, respectively.

The token updater directly integrates the local token into the token memory by fusion to maintain global semantic cross-frame token repository to get the updated token memory $\hat{\varphi}_t^b$. It can be defined as:

$$\hat{\varphi}_t^b = \mathcal{I}(\varphi_t^b, \mathcal{M}(\chi_t^b), K). \quad (7)$$

Attention Memory

The attention memory M_{attn} is designed to store and update the cross-frame features at the attention level to model long-term relationships. By incorporating attention memory into the self attention mechanism, the generation of the current frame can consider features from previous frames, providing a corrective effect. And the original self attention mechanism is transferred into Memory Self Attention. Attention memory $\text{cache}_t^l(k_m, v_m)$ build caches to store key-value pairs which are highly correlated.

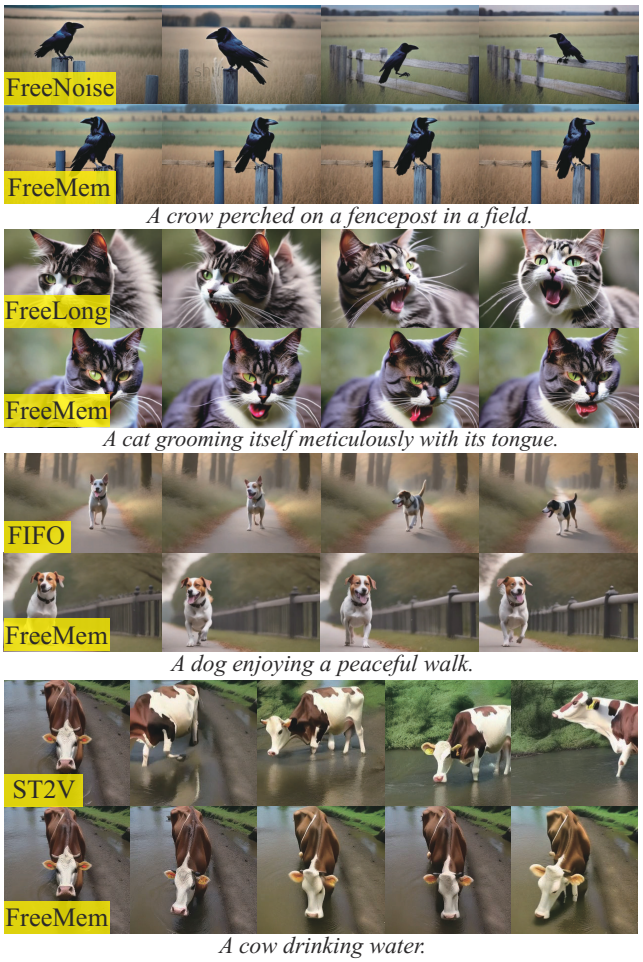

Figure 5: Qualitative comparison on long videos.

Memory Self Attention It can be divided into attention updater and attention user. First, the origin attention tensor y_t^l at the t step in the l layer is transformed to the query q_n, key k_n and value v_n. Initially, the empty cache_t^l is filled with k_n and v_n until full.

For attention updater, k_m and v_m in the attention memory are replaced by k_n and v_n from the origin attention tensor, with rate $\beta \in (0, 1)$ based on the cosine similarity between k_m and k_n. As illustrated in Figure 3(b), the most similar pairs from k_n and v_n are selected to replace the corresponding k_m and v_m to get the updated attention memory $\hat{\text{cache}}_t^l(k_m, v_m)$.

For attention user, $\text{cache}_t^l(k_m, v_m)$ are used to compute the modified attention tensor $\hat{y}_t^l$. It performs attention opera-

Method	Temporal Consistency				Video Quality				Inference Efficiency	
	sub_first	sub_pre	sub	back	aesthetic	imaging	flickering	motion	GPU(GB)	time(s)
FreeNoise	85.24	96.66	90.95	95.87	57.95	65.01	93.50	95.52	28.2	350 ± 3
+ **FreeMem**	**95.96**	**99.14**	**97.55**	**98.40**	**58.78**	**67.55**	**97.31**	**98.04**	30.0	500 ± 3
FreeLong	88.60	97.29	92.94	96.46	57.73	66.43	93.77	95.65	28.0	370 ± 2
+ **FreeMem**	**96.33**	**99.42**	**97.87**	**98.56**	**59.79**	**66.91**	**97.24**	**97.95**	32.0	520 ± 2
FIFO-Diffusion	89.71	97.84	93.78	95.35	57.42	66.28	95.41	97.27	9.9	770 ± 5
+ **FreeMem**	**91.54**	**98.54**	**95.04**	**95.94**	**57.75**	**67.30**	**96.42**	**97.73**	14.5	1180 ± 5
StreamingT2V	82.89	96.86	89.88	94.47	53.89	55.45	98.28	98.98	19.4	820 ± 3
+ **FreeMem**	**85.26**	**98.03**	**91.64**	**95.24**	**54.37**	**56.46**	**99.05**	**99.32**	21.8	840 ± 3

Table 1: Quantitative comparison on long videos.

tion by concatenating the k_m, v_m with the origin key k_n and value v_n. The process can be formulated as:

$$\hat{y_t^l} = \text{softmax}\left(\frac{q_n \left[k_n, k_m\right]^T}{\sqrt{d_k}}\right)\left[v_n, v_m\right], \tag{8}$$

where d_k is the dimension of key, $[\cdot, \cdot]$ is the concatenation.

Experiments

Implementation Details

General Setup We select four backbones for long video generation: FreeNoise (Qiu et al. 2024), FreeLong (Lu et al. 2024), FIFO-Diffusion (Kim et al. 2024), and StreamingT2V (Henschel et al. 2025) (ST2V). VideoCrafter2 (Chen et al. 2024) and StreamingT2V are used as the base model. For evaluation, we use 50 prompts, covering diverse entities and scenarios.

Metrics The evaluation is conducted by VBench (Huang et al. 2024). For temporal consistency, we use subject consistency (sub) calculated by DINO (Caron et al. 2021) feature similarity across frames, and background consistency (back) calculated by CLIP (Radford et al. 2021) feature similarity across frames. As foreground is critical, we extend subject consistency to sub_first and sub_pre, denoting the similarity between the first/previous and current frames. For video quality, we use aesthetic quality, imaging quality, temporal flickering, and motion smoothness as metrics. Motion aware warp error (Henschel et al. 2025) (MAWE) evaluates the balance of consistency and motion.

Main Results

Comparison on backbones Table 1 presents the performance of FreeMem across different backbones on 64 frames. FreeMem consistently enhances both temporal consistency and video quality across all backbones. FreeMem improves sub_first and sub_pre by over 10 and 2.5 on FreeNoise, effectively mitigating cumulative deviation. Moreover, FreeMem enhances aesthetics (0.3–2), image quality (0.5–2.5), temporal flickering (0.8–3.8), and motion smoothness (0.3–2.5), resulting in more natural and visually coherent videos.

The additional memory ranging from 2.4-4.6 GB is relatively modest, indicating that FreeMem is lightweight in memory consumption. Additional time expenditure is also limited under 50% compared to backbones, demonstrating a favorable trade-off between efficiency and performance.

Figure 5 presents qualitative comparisons. FreeNoise shows sudden background transitions (row 1), FreeLong alters the cat's face (row 3), FIFO shifts the dog's breed (row 5), and ST2V causes texture inconsistencies (row 7). These cases demonstrate appearance inconsistencies and abrupt

StreamingT2V		Temporal Consistency			
		sub_first	sub_pre	sub	back
32F	Base	88.84	97.46	93.15	96.54
	+ FM	**89.57**	**97.93**	**93.75**	**96.68**
	Enhance	**91.66**	97.78	94.72	96.18
	+ FM	91.63	**97.89**	**94.76**	**96.70**
64F	Base	79.96	96.63	88.45	94.59
	+ FM	**84.19**	**98.09**	**91.14**	**95.51**
	Enhance	82.89	96.86	89.88	94.47
	+ FM	**85.26**	**98.03**	**91.64**	**95.24**
128F	Base	67.19	96.75	81.98	91.82
	+ FM	**75.45**	**98.52**	**86.98**	**93.70**
	Enhance	73.13	96.43	84.78	92.25
	+ FM	**79.52**	**98.40**	**88.96**	**93.71**
256F	Base	52.96	96.14	74.55	88.65
	+ FM	**58.81**	**98.72**	**78.76**	**90.21**
	Enhance	62.65	95.84	79.25	90.07
	+ FM	**68.94**	**98.45**	**83.70**	**91.55**

Table 2: Quantitative comparison with different lengths.

AnimateDiff		Temporal Consistency				Dynamic
		sub_first	sub_pre	sub	back	MAWE↓
Realistic Vision	Base	92.07	96.87	94.47	94.5	5.14
	FI	93.72	98.38	96.05	95.64	4.31
	UC	92.43	97.11	94.77	95.09	24.39
	FM	**94.67**	**98.46**	**96.57**	**95.64**	4.48
Majicmix Realistic	Base	95.84	98.65	97.24	94.93	2.11
	FI	97.56	99.14	98.35	96.69	3.02
	UC	95.71	98.46	97.08	95.79	7.44
	FM	**98.00**	**99.38**	**98.69**	**96.83**	2.78
ToonYou	Base	95.69	98.57	97.13	97.05	10.50
	FI	97.83	99.32	98.57	97.69	9.41
	UC	97.16	99.05	98.10	96.58	21.89
	FM	**98.28**	**99.47**	**98.87**	**97.78**	11.94

Table 3: Quantitative comparison with baselines for short video generation on AnimateDiff (Guo et al. 2024) with three fine-tuning models.

changes caused by cumulative deviation. From the remaining rows, FreeMem effectively alleviates these variations, enhancing consistency with multi-level memories.

Comparison on lengths Table 2 presents the performance of ST2V with FreeMem (FM) with different lengths. Base refers to videos using the standard autoregressive pipeline in ST2V. Enhance represents enhanced videos refined through the Streaming Refinement Stage of ST2V. The remaining rows demonstrate the impact of incorporating FreeMem into both the base and enhanced videos. As the length of the video increases, the consistency has a downward trend because the autoregressive pipeline accumulates deviations over time. But FreeMem provides a stable improvement, especially on longer videos. FreeMem improves the sub metric by 5.00 at 128 frames and boosts the sub_first metric by 6.3 at 256 frames.

Comparison with baselines Table 3 compares FreeMem (FM) with FreeInit (Wu et al. 2024) (FI) and UniCtrl (Chen, Xia, and Xu 2024) (UC) on short video generation (16 frames) based on AnimateDiff (Guo et al. 2024) with three fine-tuning models (RealisticVision, MajicmixRealistic, ToonYou). We adapt FreeMem for short videos to ensure a fair comparison with FI and UC. Similar to FreeInit, which iteratively denoises and adds noise during video generation, the stored memory from previous generation processes is used to refine the current video generation and update the memory. FI operates on noise to retain more consistent low-frequency information and UC operates on attention to inject features from the first frame to preserve consistency, and both can improve consistency. However, due to the lack of multi-level operations, FM outperforms FI and UC. From the MAWE results, we can see that FreeMem can maintain dynamics as well.

Ablation Study

We conduct ablation studies to analyze the impact of the three memory modules on FreeNoise with 64 frames. In table 4, the progressive improvements highlight that each memory module contributes to consistency, their integration achieves the highest overall result. Each memory module reduces the cumulative deviation at one hierarchical level.

Noise Memory Analysis First, we refine long videos by FreeInit (Wu et al. 2024), but it focuses solely on sharing low-frequency information across generation rounds, overlooking the sharing between frames. Then we retain the first rearrange in Noise Memory Process (Rearrange1), directly expand it to 64 frames, and combine the low-frequency part with the high-frequency part of 64 random noise. Rearrange1 shares only low-frequency information, leaving high-frequency information independent, thus lacking constraints on high-frequency details of noise memory. Finally, we retain the second rearrange in the Noise Memory Process (Rearrange2), which rearranges 16 frames of random noise into 64 frames. While it enables fixed noise sharing, it lacks the integration of diverse low-frequency and high-frequency parts. In table 5, noise memory achieves the best result, improving the correlation between frames and mitigating deviations at noise level.

Token Memory Analysis First, we apply only merge and unmerge operations (Token Merge) to combine similar tokens. It outperforms the Base by merging similar tokens between adjacent frames, enhancing feature consistency across short-interval frames. Next, we utilize token memory with initialization and usage but without updates (Token Use), limiting improvements as it relies solely on initial frame features and ignores global long-term features. Finally, instead of averaging similar tokens, we replace them with the top token during fusion for token memory (Token Replace). Using individual features instead of the overall average features will reduce performance. In table 5, token memory shows great improvements, reducing semantic deviations at the token level compared to other alternatives.

Attention Memory Analysis First, following UniCtrl (Chen, Xia, and Xu 2024), we replace the current frame's key and value with those of the first frame for self attention (Attn1). Then, we concat the current frame's key and value with those of the first frame for self attention (Attn2). These alternatives add first-frame references differently but remain limited in feature, falling short of attention memory with dynamic updates. In table 5, attention memory achieves the best results, enhancing the long-range relationships modeling and mitigating deviations at the attention level.

Ablation			Temporal Consistency			
Noise	Token	Attn	sub_first	sub_pre	sub	back
			85.24	96.66	90.95	95.87
✓			93.92	98.84	96.38	97.95
✓	✓		94.68	99.25	96.96	98.15
✓	✓	✓	**95.96**	**99.14**	**97.55**	**98.40**

Table 4: Ablation study on noise/token/attention memory.

Condition	sub_first	sub_pre	sub	back
Base	85.24	96.66	90.95	95.87
FreeInit	91.65	98.13	94.89	97.58
Rearrange1	89.80	97.44	93.62	96.75
Rearrange2	93.58	98.31	95.94	97.88
Noise Memory	**93.92**	**98.84**	**96.38**	**97.95**
Base	85.24	96.66	90.95	95.87
Token Merge	85.47	96.76	91.12	95.93
Token Use	86.41	96.41	91.41	95.94
Token Replace	84.70	95.97	90.34	95.47
Token Memory	**86.10**	**97.09**	**91.60**	**95.97**
Base	85.24	96.66	90.95	95.87
Attn1	86.05	97.03	91.54	95.96
Attn2	86.50	95.78	91.14	95.07
Attn Memory	**87.68**	**96.80**	**92.24**	**95.99**

Table 5: Comparisons on noise/token/attention memory.

Conclusion

FreeMem boosts long video consistency via tuning-free noise/token/attention memory modules with hierarchical update/injection. It reduces multi-level deviations, enhancing subject/background consistency and quality. Validated by experiments, it integrates easily into models for cost-effective long video generation.

Acknowledgements

We thank the anonymous reviewers for their constructive comments. This research was supported by the National Natural Science Foundation of China (No.62476192) and the Natural Science Foundation of Tianjin (No.23JCQNJC02010).

References

Blattmann, A.; Dockhorn, T.; Kulal, S.; Mendelevitch, D.; Kilian, M.; Lorenz, D.; Levi, Y.; English, Z.; Voleti, V.; Letts, A.; et al. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*.

Bolya, D.; Fu, C.-Y.; Dai, X.; Zhang, P.; Feichtenhofer, C.; and Hoffman, J. 2023. Token Merging: Your ViT but Faster. In *International Conference on Learning Representations*.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging Properties in Self-Supervised Vision Transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

Chen, H.; Zhang, Y.; Cun, X.; Xia, M.; Wang, X.; Weng, C.; and Shan, Y. 2024. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7310–7320.

Chen, J.; Long, F.; An, J.; Qiu, Z.; Yao, T.; Luo, J.; and Mei, T. 2025. Ouroboros-diffusion: Exploring consistent content generation in tuning-free long video diffusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 2079–2087.

Chen, X.; Xia, T.; and Xu, S. 2024. UniCtrl: Improving the Spatiotemporal Consistency of Text-to-Video Diffusion Models via Training-Free Unified Attention Control. *arXiv preprint arXiv:2403.02332*.

Deng, H.; Pan, T.; Diao, H.; Luo, Z.; Cui, Y.; Lu, H.; Shan, S.; Qi, Y.; and Wang, X. 2024. Autoregressive Video Generation without Vector Quantization. *arXiv preprint arXiv:2412.14169*.

Dong, H.; Wang, X.; Lin, D.; Wu, Y.; Chen, Q.; Liu, R.; Yang, K.; Li, P.; and Guo, Q. 2025. NoiseController: Towards Consistent Multi-view Video Generation via Noise Decomposition and Collaboration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 14443–14452.

Duan, Z.; Zhou, W.; Chen, C.; Li, Y.; and Qian, W. 2024. ExVideo: Extending Video Diffusion Models via Parameter-Efficient Post-Tuning. *arXiv preprint arXiv:2406.14130*.

Fan, J.; Xue, H.; Zhang, Q.; and Chen, Y. 2024. Refdrop: Controllable consistency in image or video generation via reference feature guidance. *Advances in Neural Information Processing Systems*, 37: 33602–33637.

Ge, S.; Nah, S.; Liu, G.; Poon, T.; Tao, A.; Catanzaro, B.; Jacobs, D.; Huang, J.-B.; Liu, M.-Y.; and Balaji, Y. 2023. Preserve your own correlation: A noise prior for video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22930–22941.

Guo, Y.; Yang, C.; Rao, A.; Liang, Z.; Wang, Y.; Qiao, Y.; Agrawala, M.; Lin, D.; and Dai, B. 2024. AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning. *International Conference on Learning Representations*.

HaCohen, Y.; Chiprut, N.; Brazowski, B.; Shalem, D.; Moshe, D.; Richardson, E.; Levin, E.; Shiran, G.; Zabari, N.; Gordon, O.; Panet, P.; Weissbuch, S.; Kulikov, V.; Bitterman, Y.; Melumian, Z.; and Bibi, O. 2024. LTX-Video: Realtime Video Latent Diffusion. *arXiv preprint arXiv:2501.00103*.

He, B.; Li, H.; Jang, Y. K.; Jia, M.; Cao, X.; Shah, A.; Shrivastava, A.; and Lim, S.-N. 2024. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13504–13514.

Henschel, R.; Khachatryan, L.; Poghosyan, H.; Hayrapetyan, D.; Tadevosyan, V.; Wang, Z.; Navasardyan, S.; and Shi, H. 2025. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2568–2577.

Hong, Y.; Liu, B.; Wu, M.; Zhai, Y.; Chang, K.-W.; Li, L.; Lin, K.; Lin, C.-C.; Wang, J.; Yang, Z.; et al. 2024. SlowFast-VGen: Slow-Fast Learning for Action-Driven Long Video Generation. *arXiv preprint arXiv:2410.23277*.

Huang, Z.; He, Y.; Yu, J.; Zhang, F.; Si, C.; Jiang, Y.; Zhang, Y.; Wu, T.; Jin, Q.; Chanpaisit, N.; Wang, Y.; Chen, X.; Wang, L.; Lin, D.; Qiao, Y.; and Liu, Z. 2024. VBench: Comprehensive Benchmark Suite for Video Generative Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Kara, O.; Kurtkaya, B.; Yesiltepe, H.; Rehg, J. M.; and Yanardag, P. 2024. RAVE: Randomized Noise Shuffling for Fast and Consistent Video Editing with Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Kim, J.; Kang, J.; Choi, J.; and Han, B. 2024. Fifo-diffusion: Generating infinite videos from text without training. *Advances in Neural Information Processing Systems*, 37: 89834–89868.

Li, J.; Feng, W.; Fu, T.-J.; Wang, X.; Basu, S.; Chen, W.; and Wang, W. Y. 2024a. T2V-Turbo: Breaking the Quality Bottleneck of Video Consistency Model with Mixed Reward Feedback. *arXiv preprint arXiv:2405.18750*.

Li, W.; Cao, Y.; Su, X.; Lin, X.; You, S.; Zheng, M.; Chen, Y.; and Xu, C. 2024b. Training-free Long Video Generation with Chain of Diffusion Model Experts. *arXiv preprint arXiv:2408.13423*.

Li, X.; Zhang, F.; Pan, J.; Hou, Y.; Tan, V. Y.; and Yang, Z. 2024c. Enhancing Multi-Text Long Video Generation Consistency without Tuning: Time-Frequency Analysis, Prompt Alignment, and Theory. *arXiv preprint arXiv:2412.17254*.

Li, Z.; Hu, S.; Liu, S.; Zhou, L.; Choi, J.; Meng, L.; Guo, X.; Li, J.; Ling, H.; and Wei, F. 2024d. Arlon: Boosting diffusion transformers with autoregressive models for long video generation. *arXiv preprint arXiv:2410.20502*.

Lin, Z.; Liu, W.; Chen, C.; Lu, J.; Hu, W.; Fu, T.-J.; Allardice, J.; Lai, Z.; Song, L.; Zhang, B.; et al. 2025. Stiv: Scalable text and image conditioned video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16249–16259.

Lu, Y.; Liang, Y.; Zhu, L.; and Yang, Y. 2024. Freelong: Training-free long video generation with spectralblend temporal attention. *Advances in Neural Information Processing Systems*, 37: 131434–131455.

Ma, Y.; Chen, J.; Di, D.; Xie, Q.; Fan, L.; Chen, W.; Gou, X.; Zhao, N.; and Yang, X. 2025. Tuning-Free Long Video Generation via Global-Local Collaborative Diffusion. *arXiv preprint arXiv:2501.05484*.

Polyak, A.; Zohar, A.; Brown, A.; Tjandra, A.; Sinha, A.; Lee, A.; Vyas, A.; Shi, B.; Ma, C.-Y.; Chuang, C.-Y.; Yan, D.; Choudhary, D.; Wang, D.; Sethi, G.; Pang, G.; Ma, H.; Misra, I.; Hou, J.; Wang, J.; Jagadeesh, K.; Li, K.; Zhang, L.; Singh, M.; Williamson, M.; Le, M.; Yu, M.; Singh, M. K.; Zhang, P.; Vajda, P.; Duval, Q.; Girdhar, R.; Sumbaly, R.; Rambhatla, S. S.; Tsai, S.; Azadi, S.; Datta, S.; Chen, S.; Bell, S.; Ramaswamy, S.; Sheynin, S.; Bhattacharya, S.; Motwani, S.; Xu, T.; Li, T.; Hou, T.; Hsu, W.-N.; Yin, X.; Dai, X.; Taigman, Y.; Luo, Y.; Liu, Y.-C.; Wu, Y.-C.; Zhao, Y.; Kirstain, Y.; He, Z.; He, Z.; Pumarola, A.; Thabet, A.; Sanakoyeu, A.; Mallya, A.; Guo, B.; Araya, B.; Kerr, B.; Wood, C.; Liu, C.; Peng, C.; Vengertsev, D.; Schonfeld, E.; Blanchard, E.; Juefei-Xu, F.; Nord, F.; Liang, J.; Hoffman, J.; Kohler, J.; Fire, K.; Sivakumar, K.; Chen, L.; Yu, L.; Gao, L.; Georgopoulos, M.; Moritz, R.; Sampson, S. K.; Li, S.; Parmeggiani, S.; Fine, S.; Fowler, T.; Petrovic, V.; and Du, Y. 2024. Movie Gen: A Cast of Media Foundation Models. arXiv:2410.13720.

Qiu, H.; Xia, M.; Zhang, Y.; He, Y.; Wang, X.; Shan, Y.; and Liu, Z. 2024. FreeNoise: Tuning-Free Longer Video Diffusion via Noise Rescheduling. In *ICLR*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ren, W.; Yang, H.; Zhang, G.; Wei, C.; Du, X.; Huang, W.; and Chen, W. 2024. ConsistI2V: Enhancing Visual Consistency for Image-to-Video Generation. *Transactions on Machine Learning Research*.

Si, C.; Fan, W.; Lv, Z.; Huang, Z.; Qiao, Y.; and Liu, Z. 2025. RepVideo: Rethinking Cross-Layer Representation for Video Generation. *arXiv 2501.08994*.

Tan, Z.; Yang, X.; Liu, S.; and Wang, X. 2024. Videoinfinity: Distributed long video generation. *arXiv preprint arXiv:2406.16260*.

Tian, Y.; Yang, L.; Yang, H.; Gao, Y.; Deng, Y.; Chen, J.; Wang, X.; Yu, Z.; Tao, X.; Wan, P.; et al. 2024. Videotetris: Towards compositional text-to-video generation. *Advances in Neural Information Processing Systems*, 37: 29489–29513.

Wang, F.-Y.; Chen, W.; Song, G.; Ye, H.-J.; Liu, Y.; and Li, H. 2023a. Gen-l-video: Multi-text to long video generation via temporal co-denoising. *arXiv preprint arXiv:2305.18264*.

Wang, J.; Yuan, H.; Chen, D.; Zhang, Y.; Wang, X.; and Zhang, S. 2023b. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*.

Wang, X.; Li, X.; and Chen, Z. 2024. CoNo: Consistency Noise Injection for Tuning-free Long Video Diffusion. *arXiv preprint arXiv:2406.05082*.

Wang, Y.; Chen, X.; Ma, X.; Zhou, S.; Huang, Z.; Wang, Y.; Yang, C.; He, Y.; Yu, J.; Yang, P.; et al. 2023c. LAVIE: High-Quality Video Generation with Cascaded Latent Diffusion Models. *arXiv preprint arXiv:2309.15103*.

Wu, T.; Si, C.; Jiang, Y.; Huang, Z.; and Liu, Z. 2024. Freeinit: Bridging initialization gap in video diffusion models. In *European Conference on Computer Vision*, 378–394. Springer.

Yang, K.; Ma, E.; Peng, J.; Guo, Q.; Lin, D.; and Yu, K. 2023. Bevcontrol: Accurately controlling street-view elements with multi-perspective consistency via bev sketch layout. *arXiv preprint arXiv:2308.01661*.

Yang, Q.; Guan, J.; Wang, K.; Yu, L.; Chu, W.; Zhou, H.; Feng, Z.; Feng, H.; Ding, E.; Wang, J.; et al. 2024a. Showmaker: Creating high-fidelity 2d human video via fine-grained diffusion modeling. *Advances in Neural Information Processing Systems*, 37: 51039–51062.

Yang, Z.; Teng, J.; Zheng, W.; Ding, M.; Huang, S.; Xu, J.; Yang, Y.; Hong, W.; Zhang, X.; Feng, G.; et al. 2024b. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. *arXiv preprint arXiv:2408.06072*.

Yin, T.; Zhang, Q.; Zhang, R.; Freeman, W. T.; Durand, F.; Shechtman, E.; and Huang, X. 2025. From slow bidirectional to fast autoregressive video diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 22963–22974.

Yu, J.; Liu, J.; Bo, L.; and Mei, T. 2022. Memory-augmented non-local attention for video super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 17834–17843.

Zheng, L.; Zhang, Y.; Guo, H.; Pan, J.; Tan, Z.; Lu, J.; Tang, C.; An, B.; and Yan, S. 2024a. MEMO: Memory-Guided Diffusion for Expressive Talking Video Generation. *arXiv preprint arXiv:2412.04448*.

Zheng, Z.; Peng, X.; Yang, T.; Shen, C.; Li, S.; Liu, H.; Zhou, Y.; Li, T.; and You, Y. 2024b. Open-Sora: Democratizing Efficient Video Production for All.

Zhou, Y.; Zhou, D.; Cheng, M.-M.; Feng, J.; and Hou, Q. 2024. StoryDiffusion: Consistent Self-Attention for Long-Range Image and Video Generation. *Advances in Neural Information Processing Systems*.

Zhu, M.; Pan, P.; Chen, W.; and Yang, Y. 2019. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5802–5810.