FreePCA: Integrating Consistency Information across Long-short Frames in Training-free Long Video Generation via Principal Component Analysis

Jiangtong Tan, Hu Yu, Jie Huang, Jie Xiao, Feng Zhao*
MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition,
University of Science and Technology of China
{jttan, yuhu520, hj0117, ustchbxj}@mail.ustc.edu.cn, fzhao956@ustc.edu.cn

Abstract

*Long video generation involves generating extended videos using models trained on short videos, suffering from distribution shifts due to varying frame counts. It necessitates the use of local information from the original short frames to enhance visual and motion quality, and global information from the entire long frames to ensure appearance consistency. Existing training-free methods struggle to effectively integrate the benefits of both, as appearance and motion in videos are closely coupled, leading to motion inconsistency and visual quality. In this paper, we reveal that global and local information can be precisely decoupled into consistent appearance and motion intensity information by applying Principal Component Analysis (PCA), allowing for refined complementary integration of global consistency and local quality. With this insight, we propose **FreePCA**, a training-free long video generation paradigm based on PCA that simultaneously achieves high consistency and quality. Concretely, we decouple consistent appearance and motion intensity features by measuring cosine similarity in the principal component space. Critically, we progressively integrate these features to preserve original quality and ensure smooth transitions, while further enhancing consistency by reusing the mean statistics of the initial noise. Experiments demonstrate that FreePCA can be applied to various video diffusion models without requiring training, leading to substantial improvements. Code is available at https://github.com/JosephTiTan/FreePCA.*

1. Introduction

As diffusion models have gained prominence in the field of image generation [17, 32], video diffusion generation models [18, 33, 47] have also begun to develop, leading to various applications such as video editing [8, 27, 42] and motion control [13, 15, 39, 43]. Previous video dif-

*Corresponding author.

fusion models [7, 12, 40], trained on large video datasets, have made significant progress in generating high-quality videos. However, training a diffusion model capable of producing long videos remains a resource-intensive task, requiring substantial data collection and annotation [44].

Therefore, researchers have begun to explore methods for generating long videos without training. Directly inputting global noise sequences into models trained on short videos to generate longer videos results in reduced quality, missing objects, and slow motion due to distribution shifts from training data. As shown in Fig. 1(a), the concept of "playing drum kit" is lost, referred to as the global aligned method. To address these issues, some methods [31, 38] maintain the original local frame count using sliding windows to align with the training data, then apply fusion techniques to stitch them together into long videos, referred to as the local stitched method. While these methods can maintain video quality, ensuring consistency between windows remains challenging, as seen in Fig. 1(b). [28] proposes to generate videos by directly using global noise sequences to enhance video consistency, and fuses it with local attention map in the frequency domain to enhance details. But it relies on features with longer frames to generate the entire video, overlooking the flexibility needed in video generation, which leads to the lack of motion and semantic richness like Fig. 1(a). As a summary, how to effectively integrate global information to maintain consistency while ensuring the quality of local information for video generation is significant for this task.

We argue that the local information extracted by sliding windows with original frame count is reasonable for preserving the data distribution of the video generation model and reflects the quality of visual and motion. While global information reflects overall consistency, as its appearance remains unchanged despite the decline in quality in Fig. 1(a). Therefore, we aim to find a strong decoupling space to establish an appropriate complementary relationship between global and local information, thereby achieving high consistency and quality in long video generation.
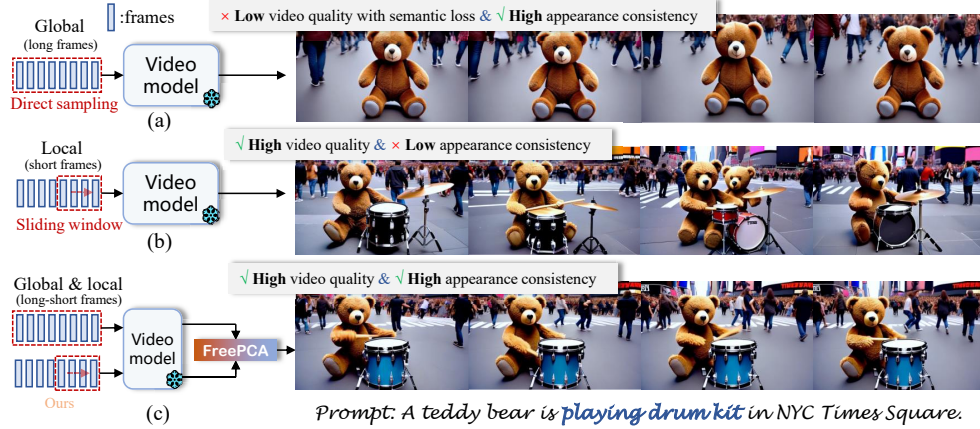
Figure 1. Illustration of different training-free methods for generating long videos. (a) Global aligned method, which inputs the entire video sequence into the model, resulting in lower quality, object loss, and slow motion, but maintains consistency. (b) Local stitched method, which uses a sliding window to extract video segments along the temporal dimension, resulting in poor consistency but retaining the original generation quality. (c) Our FreePCA, which effectively combines the global and local methods via PCA, achieving good consistency while preserving the original generation quality.

Inspired by methods in the background subtraction task that use Principal Component Analysis (PCA) in the spatial dimension to segment moving foregrounds from consistent backgrounds [1, 10, 11], we observe that applying PCA in the temporal dimension allows video sequences to be effectively decoupled into consistent appearance and motion intensity. The consistent appearance derived from global information complements local information, enabling effective integration of their advantages (see Sec. 3). This motivates us to integrate this consistency feature generated by the global method into the features of the local method in the principal component space, achieving better consistency and quality, as shown in Fig. 1(c).

In this paper, we propose a training-free method for generating long videos with high consistency and quality via PCA, named FreePCA. Specifically, our approach consists of two steps: 1) **Consistency Feature Decomposition**. As shown in Fig. 4, we use both long and short frames to generate global and local features, then apply PCA to project them into a shared principal component space, leveraging PCA's strong decoupling capacity to extract consistency features. Given the inherent semantic similarity of global and local features, we employ cosine similarity to compare their components and divide them into consistent appearance and motion intensity features based on their similarity levels, which confirms the fact that appearances generated by both methods are similar but their motion intensities differ. Since the consistent appearance in global features is smoother and more stable than that in local features, we utilize high similarity components of global features as consistency features to complement local features. 2) **Progressive Fusion**. We gradually increase the proportion of

consistency features as the window slides to integrate them into the local features. This maintains the model's flexibility in video generation with the original frame count, preserving consistency and ensuring smooth transitions without compromising video quality. We also reuse the mean statistics of the initial noise to enhance video consistency further. Experiments show that FreePCA can be applied to various video diffusion models without training, yielding high-quality results while maintaining consistency. Our method supports multi-prompt generation and continuous video generation, demonstrating excellent performance.

In summary, our contributions are as follows:

- We reveal that PCA can effectively decouple video features into consistent appearance and motion intensity features for the first time, thereby addressing the inconsistency and low quality issues in long video generation.
- Specifically, we introduce a technique to extract the consistency feature from the entire video sequence's global feature in the principal component space and progressively integrate it into the local feature obtained through the sliding window, which ensures video consistency without compromising video quality.
- Extensive experiments demonstrate that our method outperforms existing approaches, achieving state-of-the-art results. Moreover, it can be applied to multiple fundamental video diffusion models without extra training.

2. Related work

2.1. Text-to-video Diffusion Models

In recent years, video generation technology has developed rapidly. Initially, this technology relied on networks based

on Variational Autoencoders (VAEs) [29, 36, 41] and Generative Adversarial Networks (GANs) [9, 30], and it later shifted toward diffusion models [17, 18, 33, 47]. Notably, previous work [3, 4, 42] integrate 3D convolutions and temporal attention to ensure coherent video generation, providing inspiration for subsequent model designs. [12] fine-tunes diffusion models by adding motion modules, demonstrating strong practical applicability and compatibility with LoRA. [16, 45, 46] employ a latent video diffusion model to compress videos into lower dimensions, enhancing generation efficiency. Building on these, [7] utilizes a video latent diffusion process, leveraging high-quality images to improve visual fidelity. However, due to computational constraints and the scarcity of high-quality video datasets, most existing video diffusion models are trained only on fixed-length short videos, limiting their ability to generate longer content. In this paper, we address this limitation by generating high-quality long videos from short video diffusion models without requiring additional training.

2.2. Long Video Generation

Generating long videos has been a challenging task due to the complexities of temporal modeling and resource limitations. Both GAN based [5, 22, 24, 34] and diffusion based [14, 21, 23, 35, 37] methods have proposed solutions for long video generation. Some approaches [44, 48] require extensive training and entail significant computational costs. Conversely, certain methods [26, 28, 31, 38] enable long video generation without training, which we categorize into extrapolation and interpolation. Extrapolation methods [31, 38] maintain the original frame generation effect, but often suffer from lower consistency. In contrast, interpolation methods [28] enhance long video sequences with better consistency, though they may compromise visual quality and motion richness compared to the original frames. Our method combines the benefits of both categories, achieving better consistency while preserving both visual quality and motion richness.

3. Observation and Analysis

In this section, we first introduce the motivation for using PCA and demonstrate that after applying PCA to a video, certain components in the principal component space retain a consistent appearance. We also observe that the proportion of consistency information varies among different long video generation methods statistically. Furthermore, we show how to extract the consistency feature from the video feature in the diffusion model and difference from previous methods.

Rationale for using PCA. Inspired by PCA's information integration capabilities used in video segmentation [1, 10, 11], we find that PCA can measure linear correlations between frames in the temporal dimension and enables us to
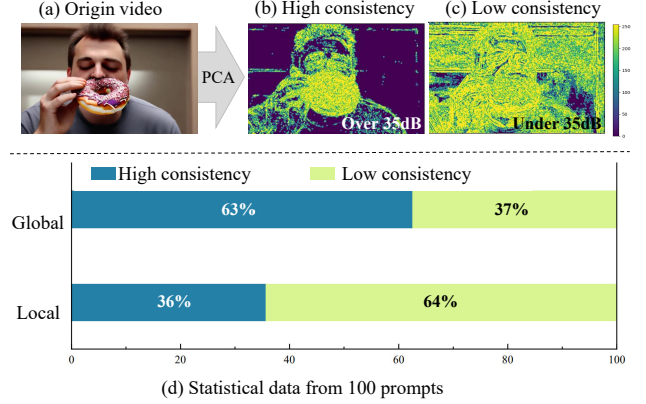

Figure 2. (a-c) Illustration of consistency information extraction after applying PCA to videos and converting each component's information into edge maps. After applying PCA to video (a), some components' information exhibit high consistency like (b), while others show inconsistency like (c). We use PSNR to assess whether each component's information exhibit consistency as the original video, setting a threshold of 35 dB. (d) shows the statistics of 100 videos generated by the global and local methods. We categorize these videos based on whether they have consistency components into two groups: high consistency, and low consistency. It indicates that PCA can separate consistency information and exhibits statistical patterns, reflecting different consistency degree of different methods in the principal component space.

decouple video features into consistent appearance and motion diversity. After applying PCA to the video in the time dimension, we separate the information of each component in the principal component space and then map it back to the original space individually. We find that although there is significant loss of information in each component, some components still retain a consistent appearance attribute. To represent this consistency, we employ Canny edge detection [2] on each frame and overlaid all frames. If the edges concentrate in a specific area and display a distinct appearance, it indicates better consistency; conversely, if they are dispersed across a wide area, it suggests poorer consistency. To differentiate this consistency variation, we use PSNR [19] to measure the distance between the appearance of the video after applying PCA and the original video to determine whether a component exhibits consistency, setting a threshold of 35 dB. As illustrated in Fig. 2, after applying PCA, some components reflect a consistent appearance attribute, while others are chaotic and inconsistent.

To analyze the probability of these consistency components from a statistical perspective and their relationship with different generation methods, we applied PCA to videos generated using both long frames (global method) and short frames (local method) with 100 prompts. We categorize the 100 examples into two groups by whether they have consistency components, named high consistency and

low consistency. The results indicate that the local method produces a higher number of videos with low consistency compared to the global method statistically. This difference suggests that the degree of consistency of global and local methods can be measured in the principal component space, motivating us to use PCA to address inconsistency issues in local methods and preserve original quality.

Due to the crucial role of temporal attention in long video generation [28, 31], we simultaneously employ both global and local methods in temporal attention and extract their features in the principal component space. We find that after comparing the cosine similarity of global and local features across each component, the components with high cosine similarity exhibit consistent appearance attributes, while other components reflect motion intensity attributes. As visualized in Fig. 3, we subtract the features of adjacent frames one by one, and the resulting values reflect the intensity of changes with time. After visualizing all components' features, we observe that the local feature (b) exhibits greater change intensity compared to the global feature (a), which is due to the stronger consistency of global features. However, both of them struggle to reveal distinct appearance due to coupled appearance and motion. After applying cosine similarity to select components, features with high cosine similarity display distinct structural appearance, indicating their ability to preserve consistent appearance. Additionally, these features are more stable and smooth in the global feature (c), compensating for the unstable and chaotic shortcomings of the local feature (d). In contrast, features with low cosine similarity cannot display distinct appearances but retain the intensity of changes and have larger values for the local feature (f), preserving more rich motion information than the global feature (e). Therefore, in order to achieve precise complementation of their advantages, it is reasonable to integrate features with better consistency (c) into the local features while retaining features (f) with rich motion information to address low quality and inconsistency issues in long video generation.

Difference from previous methods. Our method seems to be similar to [31] and [28], but there are significant differences. We employs PCA that has stronger decoupling capability to decouple video features into consistent appearance and motion diversity and presents clear physical meanings at the feature level. Our method emphasizes how to integrate global consistency with local diversity to achieve superior generation results, an aspect that has not been addressed effectively.

4. Method

Based on the analysis, we propose FreePCA, a train-free method for generating long videos with improved consistency and quality using a pre-trained diffusion model grounded in PCA. As shown in Fig. 4, the pre-trained model
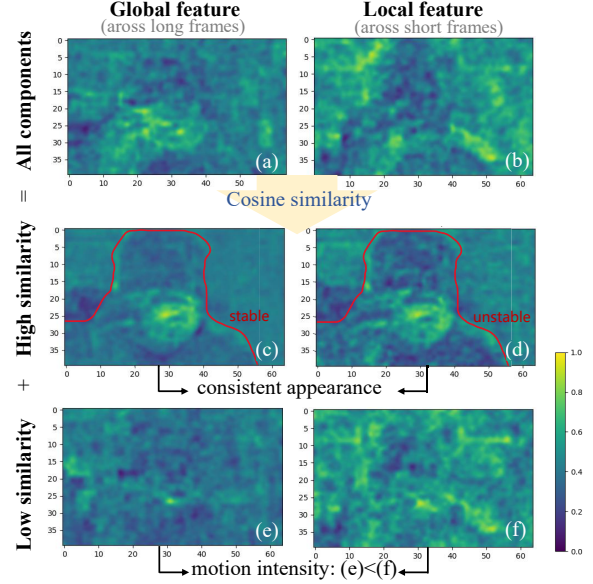

Figure 3. Visualization of consistency features extracted in the principal component space using cosine similarity. After applying both global and local methods in temporal attention and extracting features from the principal component space, we subtract each adjacent frame to exhibit the intensity of changes over time. (a) and (b) show features across all components. (c), (d), (e), and (f) illustrate features selected based on the cosine similarity between global and local features. The distinct character outlines in (c) and (d) indicate that features with high cosine similarity have consistent appearance attributes, while the significant intensity difference between (e) and (f) indicates that features with low cosine similarity exhibit motion intensity attributes.

has a U-net structure with convolutional layers, spatial transformers, and temporal transformers, trained on short video data. FreePCA focuses on the temporal transformer and consists of two steps: Consistency Feature Decomposition and Progressive Fusion. We also leverage mean statistics of initial noise to enhance consistency.

4.1. Consistency Feature Decomposition

Let the features input to the temporal transformer module be $x \in \mathbb{R}^{(b \times h \times w) \times F \times c}$, where b, h, w, F, c represent batch size, height, width, number of frames, and number of channels, respectively. The pre-trained model is trained on video data with f frames, where $f < F$. After inputting x into the temporal attention using both long and short frames, we represent the global feature obtained by inputting the entire video sequence as $x_{global} \in \mathbb{R}^{(b \times h \times w) \times F \times c}$ and the local feature obtained by using the i-th sliding window of size f as $x_{local}^i \in \mathbb{R}^{(b \times h \times w) \times f \times c}$. Since the number of frames in x_{global} is larger than that in x_{local}^i, and to align the features with it, the features of x_{global} are sliced into $x_{global}^i \in \mathbb{R}^{(b \times h \times w) \times f \times c}$ according to the positions of the
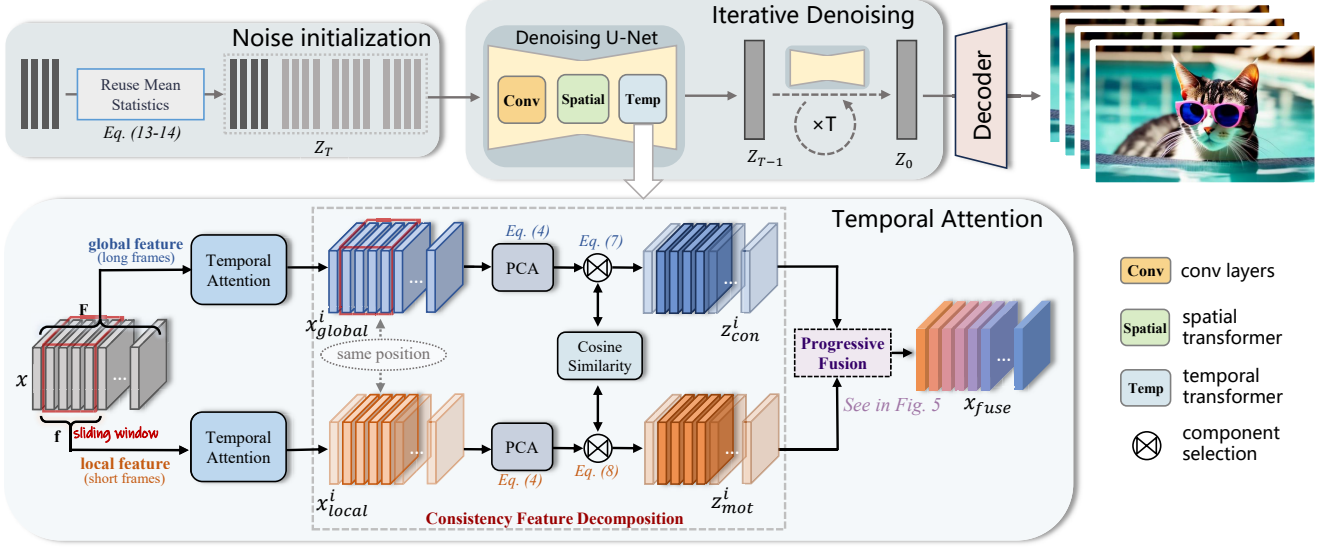
Figure 4. Overview of our method. For noise initialization, we extend short initial noise into long initial noise using a reuse mean statistics approach. In the iterative denoising process, we fuse global and local features in temporal attention through two processes: Consistency Feature Decomposition and Progressive Fusion. For Consistency Feature Decomposition, we crop the global features to the same size as the local features based on their positions. Then, we apply PCA along the temporal dimension to both, using cosine similarity to compare global and local features and decouple them into consistent appearance and motion intensity features. Finally, we perform Progressive Fusion to integrate the global consistent features into the local motion features progressively (see Fig. 5 for details).

i-th sliding window, which can be represented as:

$$x_{global}^i = Slice^i(Temp(x)), \qquad (1)$$

$$x_{local}^i = Temp(Slice^i(x)), \qquad (2)$$

where $Temp$ represents temporal attention and $Slice^i$ represents slicing the video sequence with the i-th window. Due to the deviation of x_{global}^i from the original distribution, we use the principle of attention entropy [25] to amplify its query value by a scaling factor $\lambda = \sqrt{\log_f F}$. Details can be seen in *supplementary materials*.

The original PCA process includes data normalization, calculating the covariance matrix, eigenvalue decomposition, and selecting principal components. To simplify the process, we denote the procedure of calculating the transformation matrix P using eigenvalue decomposition before selecting principal components as $\mathbb{T}_{PCA}$. With the frame dimension as the feature dimension for PCA, we project the two features onto a principal component space using a transformation matrix P from x_{global}^i, which can be represented as:

$$P = \mathbb{T}_{PCA}(x_{global}^i), \qquad (3)$$

$$z_{global}^i, z_{local}^i = P \cdot x_{global}^i, P \cdot x_{local}^i, \qquad (4)$$

where $P \in \mathbb{R}^{f \times f}$. Note that x_{global}^i and x_{local}^i are reshaped from $\mathbb{R}^{(b \times h \times w) \times f \times c}$ to $\mathbb{R}^{f \times (b \times h \times w) \times c}$ before performing the matrix multiplication. And $z_{global}^i, z_{local}^i \in$

$\mathbb{R}^{f \times (b \times h \times w) \times c}$. To extract consistency features, we compare the cosine similarity of each component of global and local features in the principal component space and select the top k most similar components from the z_{global}^i, treating their corresponding features as consistency features, and remove original consistency features in z_{local}^i, which can be represented as:

$$s_{(1)}, s_{(2)}, ..., s_{(f)} = CosSim(z_{global}^i, z_{local}^i), \qquad (5)$$

$$n_{(1)}, n_{(2)}, ..., n_{(f)} = argsort(s_{(1)}, s_{(2)}, ..., s_{(f)}), \qquad (6)$$

$$z_{con}^i = z_{global}^i[n_{(1)}, n_{(2)}, ..., n_{(k)}], \qquad (7)$$

$$z_{mot}^i = z_{local}^i[n_{(k+1)}, n_{(k+2)}, ..., n_{(f)}], \qquad (8)$$

where $CosSim(\cdot, \cdot)$ represents the computation of cosine similarity s for each of the f components. $argsort(\cdots)$ provides the indices n of the ascending sorted list for cosine similarity. Eq. (7) and Eq. (8) indicate "component selection" to obtain the consistent appearance features z_{con}^i selected from the top k highest cosine similarity components in z_{global}^i and the motion intensity features z_{mot}^i from z_{local}^i after removing the consistency features.

4.2. Progressive Fusion

To avoid affecting the original generation quality of the video model, we gradually add consistency features during the sliding window process. Specifically, we use different values of k for each window to control the proportion of
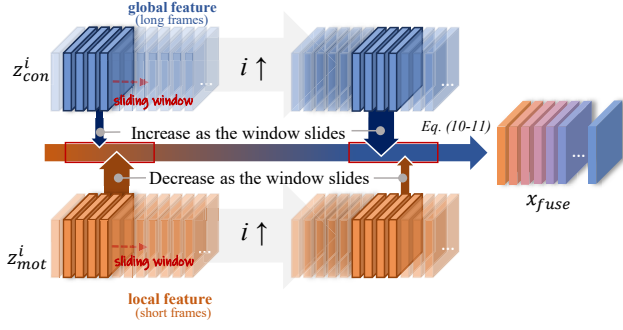
Figure 5. Illustration of Progressive Fusion. As the window slides, the proportion of consistent appearance features from the global features is gradually increased, while the proportion of local motion intensity features is decreased. Finally, the overlapping portions within the window are averaged to obtain the final result.

consistency features added. The relationship between k and the i-th sliding window can be expressed as:

$$k = min(i, K_{max}), i \in \mathbb{N}. \quad (9)$$

To preserve the original video generation quality as much as possible, we set a maximum value $K_{max} = 3$. Then, the selected consistency features can be progressively integrated into local features, and finally the transposed matrix P^T is used to map them back to the original space:

$$z_{fuse}^i = Concat(z_{con}^i, z_{mot}^i), \quad (10)$$

$$x_{fuse}^i = P^T \cdot z_{fuse}^i, \quad (11)$$

where $Concat$ indicates concatenation along the temporal dimension. To get the whole video sequences, we average the values in the overlapping windows and reshape it back to $\mathbb{R}^{(b \times h \times w) \times F \times c}$ to get the final x_{fuse}. Recent research[6] has suggested that diffusion models initially generate scene layouts and object shapes, followed by fine details in later steps. Therefore, in the 50-step denoising process of DDIM, we use the complete FreePCA for the first 25 steps and employ the local method for the remaining 25 steps to ensure better generation results.

4.3. Reuse Mean Statistics

Previous method [31] employ noise rescheduling technique to ensure video consistency. However, this approach imposes strict limitations on the input, which hinder the generation of richer scenes. Earlier work [43] has suggested that the mean extracted from the video sequence in the temporal dimension can reflect appearance information. Inspired by this, we extract the mean of the first f frames and replace the mean of the subsequent $F - f$ frames of noise. We find that this approach not only maintains the appearance consistency of the video but also enhances the flexibility of video

generation, which can be expressed as:

$$\epsilon_t \sim \mathcal{N}(0, 1), t = 1, 2, ..., F, \quad (12)$$

$$\epsilon_{j:j+f}' = \epsilon_{j:j+f} - mean(\epsilon_{j:j+f}) + mean(\epsilon_{1:f}), \quad (13)$$

where $j = nf + 1, n \in \mathbb{Z}^+$, and $j < F - f$. Followed [31], we also use the same shuffle way, then all the initial noise can be represented as:

$$[\epsilon_1, \epsilon_2, ...,, \epsilon_f, sh(\epsilon_{f+1:2f+1}'), ..., sh(\epsilon_{j:j+f}'), ...], \quad (14)$$

where $sh(\cdot)$ denotes shuffling the order of frame sequences.

5. Experiments

5.1. Implementation Details

Setting up. To validate the effectiveness and generalizability of our method, we apply FreePCA to the publicly available diffusion-based text-to-video models, VideoCrafter2 [7] and LaVie [40], which are trained on 16-frame video data. Our goal is to enable these models to generate longer videos (i.e., 64 frames) while maintaining the original video generation quality as much as possible. Our method requires no training and can be used directly during the inference phase.

Test Prompts. We use 326 prompts from Vbench [20] to test the effectiveness of our method.

Evaluation Metrics. We employ the metrics available in Vbench[20] to evaluate our approach. We test on two aspects: video consistency and video quality. For video consistency, we use three metrics: 1) Subject Consistency, which assesses whether objects remain consistent throughout the video by evaluating the similarity of DINO features between frames. 2) Background Consistency, which measures the consistency of the background scene by calculating the similarity of CLIP features between frames. 3) Overall Consistency, which evaluates semantic and stylistic consistency by using ViCLIP features to compute the similarity between frames. For video quality, we test from two perspectives: motion and appearance, using three metrics: 1) Motion Smoothness, which assesses motion smoothness using the AMT video interpolation model. 2) Dynamic Degree, which estimates the optical flow intensity between consecutive frames using RAFT to determine whether the video is static. 3) Imaging Quality, using the MUSIQ image quality predictor trained on the SPAQ dataset.

Baseline. We compare our method FreePCA with other training-free long video generation methods, including: 1) Direct Sampling, which directly uses a short video model to generate a 64-frame video. 2) FreeNoise [31], which introduce noise rescheduling to maintain consistency between video frames. 3) FreeLong [28], which blends low-frequency global feature with high-frequency local attention map to improve video quality.

Table 1. Quantitative Comparison. "Direct sampling" indicates directly sampling 64 frames based on short video generation models. The best values are shown in **bold**.

Methods	Video Consistency			Video Quality			Inference Time (min)
	Sub($\uparrow$)	Back($\uparrow$)	Over($\uparrow$)	Motion($\uparrow$)	Dynamic($\uparrow$)	Imaging($\uparrow$)	
Direct sampling	93.38	95.16	23.52	92.89	44.45	60.02	**3.6**
FreeNoise [31]	91.98	93.86	25.62	94.83	52.77	63.49	4.3
FreeLong [28]	93.77	93.79	24.76	94.49	45.83	63.35	4.1
Ours	**95.54**	**95.24**	**25.69**	**96.41**	**59.72**	**63.70**	4.7

Figure 6. Qualitative comparison using VideoCrafter2 as base model. Direct sampling leads to a loss of detail and semantics, while FreeNoise and FreeLong exhibit inconsistencies. Our method performs best in terms of both quality and consistency.

Table 2. Result of ablation study.

Config	Sub$\uparrow$	Over$\uparrow$	Motion$\uparrow$	Imaging$\uparrow$
(1)$K_{max} = 1$	89.13	25.35	91.46	60.20
(1)$K_{max} = 5$	89.40	25.14	91.86	60.56
(2)	94.19	25.40	95.60	61.36
(3)	94.11	25.44	95.40	59.86
(4)	88.98	25.39	91.69	59.10
(5)	94.21	25.18	96.29	62.49
(6)	89.54	25.16	92.23	60.19
Ours($K_{max} = 3$)	**95.54**	**25.69**	**96.41**	**63.70**

5.2. Comparison with the Baseline

Table 1 presents the quantitative results. Directly generating long videos faces issues with domain generalization, leading to a decline in both appearance and motion quality, although its consistency remains relatively acceptable. It also has the worst overall consistency due to its semantic accuracy. FreeNoise does not exhibit significant declines in quality metrics due to sliding windows, but its consistency worsens. FreeLong struggles to further improve quality due to its simplistic frequency fusion way. In contrast, our FreePCA not only achieves superior video quality but also maintains the best consistency due to the use of PCA and the progressive fusion approach. Additionally, we test

the inference time of our method on an NVIDIA RTX 4090 and find that our approach achieves better generation results with an acceptable increase in inference time. We also present the results under the DiT framework in the *supplementary materials*.

Qualitative comparisons are presented in Fig. 6 and Fig. 7. It is clear that directly generating long videos results in significant quality degradation, including missing objects, slow motion, and lacking details. FreeNoise exhibits poor appearance consistency. While FreeLong slightly improves consistency, inconsistencies and loss of semantics still exist. In contrast, our FreePCA maintains excellent consistency while ensuring high quality in both appearance and motion.

5.3. Ablation

We conduct ablation experiments on: 1) The choice of K_{max}. 2) Removing the PCA process. 3) Substituting the cosine similarity selection with random selection. 4) Setting $k = 3$ as a fixed value. 5) Replacing the reuse mean statistics with direct reuse. 6) Removing the reuse mean statistics. The results in Table 2 show that the overall performance is best when $K_{max} = 3$. It is also evident that optimal results are achieved only when all components of our method are fully utilized. The qualitative comparisons
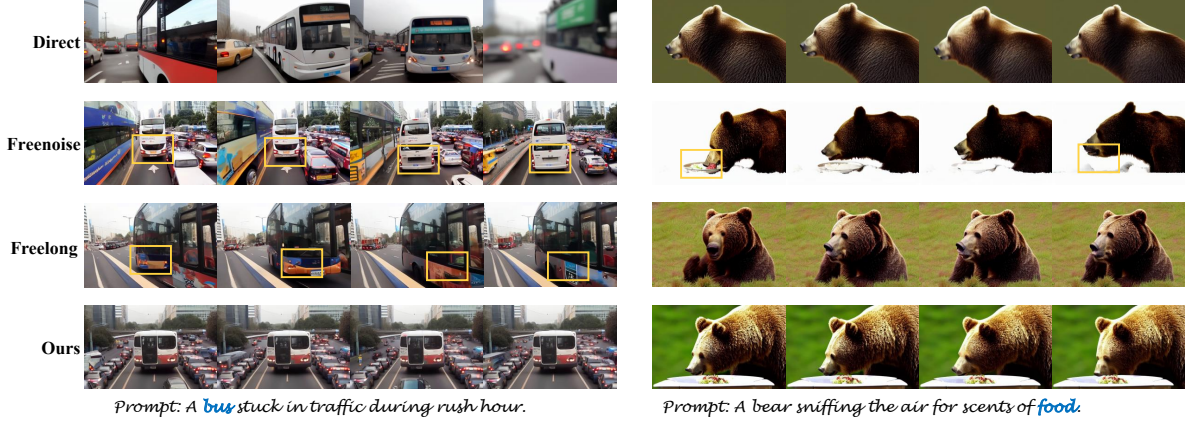
Figure 7. Qualitative comparison using LaVie as base model. Direct sampling results in content blurriness and slow motion, while FreeNoise and FreeLong exhibit inconsistencies and loss of semantics. Our method has the best quality and consistency.

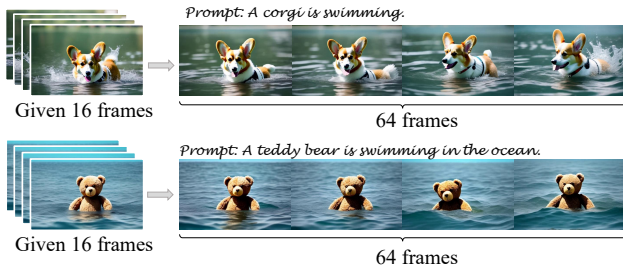Figure 8. Result of multi-prompt video generation.

Figure 9. Result of continuing generation based on a given video.

and details are presented in the *supplementary materials*.

5.4. Enhance Consistency in Other Applications

Our method not only generates long videos but can also be applied in other scenarios to enhance the consistency of generated videos, such as multi-prompt video generation and continuing video generation based on a given video.

Multi-prompt Video Generation. Our FreePCA can be seamlessly extended to multi-prompt video generation, where different prompts are provided for various video seg-

ments. As illustrated in Fig. 8, our approach enhances the consistency of video generation, maintaining a cohesive appearance even under different prompts.

Continuing Video Generation. This task involves expanding a given short video into a longer video with the same content. By simply using DDIM inversion to the given video before performing FreePCA, we can generate richer, longer video content while keeping the original video unchanged, as illustrated in Fig. 9. The experiments above thoroughly demonstrate the generalizability and practicality of our approach, establishing it as a viable paradigm for maintaining video consistency across various scenarios. Details can be found in the *supplementary materials*.

6. Conclusion

In this paper, we introduce FreePCA, a training-free method for generating high quality and consistency long videos from short video diffusion models. Leveraging the strong decoupling capacity of PCA to extract consistency features from video features, we propose Consistency Feature Decomposition, applying cosine similarity after PCA process to identify consistency features. We also design Progressive Fusion, gradually increasing the proportion of consistency features as the window slides to ensure video consistency without sacrificing quality. Additionally, we introduce Reuse Mean Statistics to further enhance consistency. Experiments demonstrate that FreePCA significantly outperforms existing models, achieving high fidelity and consistency, and establishes a training-free paradigm for enhancing consistency in other applications.

References

[1] Basit Alawode and Sajid Javed. Learning spatial-temporal regularized tensor sparse rpca for background subtraction. *arXiv preprint arXiv:2309.15576*, 2023. 2, 3

[2] Paul Bao, Lei Zhang, and Xiaolin Wu. Canny edge detection enhancement by scale multiplication. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(9): 1485–1490, 2005. 3

[3] 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. 3

[4] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 3

[5] Tim Brooks, Janne Hellsten, Miika Aittala, Ting-Chun Wang, Timo Aila, Jaakko Lehtinen, Ming-Yu Liu, Alexei Efros, and Tero Karras. Generating long videos of dynamic scenes. *Advances in Neural Information Processing Systems*, 35:31769–31781, 2022. 3

[6] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22560–22570, 2023. 6

[7] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7310–7320, 2024. 1, 3, 6

[8] Nathaniel Cohen, Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Slicedit: Zero-shot video editing with text-to-image diffusion models using spatio-temporal slices. *arXiv preprint arXiv:2405.12211*, 2024. 1

[9] Kangle Deng, Tianyi Fei, Xin Huang, and Yuxin Peng. Ircgan: Introspective recurrent convolutional gan for text-to-video generation. In *International Joint Conference on Artificial Intelligence*, pages 2216–2222, 2019. 3

[10] Zhi Gao, Loong-Fah Cheong, and Yu-Xiang Wang. Block-sparse rpca for salient motion detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(10):1975–1987, 2014. 2, 3

[11] Donald Goldfarb and Zhiwei Qin. Robust low-rank tensor recovery: Models and algorithms. *SIAM Journal on Matrix Analysis and Applications*, 35(1):225–253, 2014. 2, 3

[12] 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. *arXiv preprint arXiv:2307.04725*, 2023. 1, 3

[13] Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Sparsectrl: Adding sparse controls to text-to-video diffusion models. In *European Conference on Computer Vision*, pages 330–348. Springer, 2025. 1

[14] William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. Flexible diffusion modeling of long videos. *Advances in Neural Information Processing Systems*, 35:27953–27965, 2022. 3

[15] 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. 1

[16] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *arXiv preprint arXiv:2211.13221*, 2(3):4, 2022. 3

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1, 3

[18] 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. 1, 3

[19] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th International Conference on Pattern Recognition*, pages 2366–2369. IEEE, 2010. 3

[20] 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*, pages 21807–21818, 2024. 6

[21] Deyi Ji, Haoran Wang, Mingyuan Tao, Jianqiang Huang, Xian-Sheng Hua, and Hongtao Lu. Structural and statistical texture knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16876–16885, 2022. 3

[22] Deyi Ji, Feng Zhao, Hongtao Lu, Mingyuan Tao, and Jieping Ye. Ultra-high resolution segmentation with ultra-rich context: A novel benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23621–23630, 2023. 3

[23] Deyi Ji, Wenwei Jin, Hongtao Lu, and Feng Zhao. Ppt-former: Pseudo multi-perspective transformer for uav segmentation. *International Joint Conference on Artificial Intelligence*, pages 893–901, 2024. 3

[24] Deyi Ji, Feng Zhao, Lanyun Zhu, Wenwei Jin, Hongtao Lu, and Jieping Ye. Discrete latent perspective learning for segmentation and detection. In *International Conference on Machine Learning*, pages 21719–21730, 2024. 3

[25] Zhiyu Jin, Xuli Shen, Bin Li, and Xiangyang Xue. Training-free diffusion model adaptation for variable-sized text-to-image synthesis. *Advances in Neural Information Processing Systems*, 36:70847–70860, 2023. 5

[26] Jihwan Kim, Junoh Kang, Jinyoung Choi, and Bohyung Han. Fifo-diffusion: Generating infinite videos from text without training. *arXiv preprint arXiv:2405.11473*, 2024. 3

[27] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8599–8608, 2024. 1

[28] Yu Lu, Yuanzhi Liang, Linchao Zhu, and Yi Yang. Freelong: Training-free long video generation with spectralblend temporal attention. *arXiv preprint arXiv:2407.19918*, 2024. 1, 3, 4, 6, 7

[29] Gaurav Mittal, Tanya Marwah, and Vineeth N Balasubramanian. Sync-draw: Automatic video generation using deep recurrent attentive architectures. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1096–1104, 2017. 3

[30] Yingwei Pan, Zhaofan Qiu, Ting Yao, Houqiang Li, and Tao Mei. To create what you tell: Generating videos from captions. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1789–1798, 2017. 3

[31] Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. Freenoise: Tuning-free longer video diffusion via noise rescheduling. *arXiv preprint arXiv:2310.15169*, 2023. 1, 3, 4, 6, 7

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1

[33] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 1, 3

[34] 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*, pages 3626–3636, 2022. 3

[35] Hung-Yu Tseng, Qinbo Li, Changil Kim, Suhib Alsisan, Jia-Bin Huang, and Johannes Kopf. Consistent view synthesis with pose-guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16773–16783, 2023. 3

[36] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *arXiv preprint arXiv:1711.00937*, 2017. 3

[37] Vikram Voleti, Alexia Jolicoeur-Martineau, and Chris Pal. Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. *Advances in Neural Information Processing Systems*, 35:23371–23385, 2022. 3

[38] Fu-Yun Wang, Wenshuo Chen, Guanglu Song, Han-Jia Ye, Yu Liu, and Hongsheng Li. Gen-l-video: Multi-text to long video generation via temporal co-denoising. *arXiv preprint arXiv:2305.18264*, 2023. 1, 3

[39] 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. *arXiv preprint arXiv:2306.02018*, 2024. 1

[40] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023. 1, 6

[41] Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. Godiva: Generating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806*, 2021. 3

[42] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023. 1, 3

[43] Zeqi Xiao, Yifan Zhou, Shuai Yang, and Xingang Pan. Video diffusion models are training-free motion interpreter and controller. *arXiv preprint arXiv:2405.14864*, 2024. 1, 6

[44] Shengming Yin, Chenfei Wu, Huan Yang, Jianfeng Wang, Xiaodong Wang, Minheng Ni, Zhengyuan Yang, Linjie Li, Shuguang Liu, Fan Yang, et al. Nuwa-xl: Diffusion over diffusion for extremely long video generation. *arXiv preprint arXiv:2303.12346*, 2023. 1, 3

[45] Yan Zeng, Guoqiang Wei, Jiani Zheng, Jiaxin Zou, Yang Wei, Yuchen Zhang, and Hang Li. Make pixels dance: High-dynamic video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8850–8860, 2024. 3

[46] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *arXiv preprint arXiv:2309.15818*, 2023. 3

[47] 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. 1, 3

[48] Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *arXiv preprint arXiv:2405.01434*, 2024. 3