

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

## Supplementary Material

- Sec. 1 introduces the experiment that selects consistent features based on the ranking of eigenvalues, which demonstrates that there is no significant correlation between eigenvalue and consistency features;
- Sec. 2 provides an introduction to attention entropy and the derivation process in our method;
- Sec. 4 provides details of experiments for testing our method and more qualitative comparisons.
- Sec. 5 provides details of the ablation experiments and the qualitative results;
- Sec. 6 provides the experimental details for multi-prompt video generation and continuing video generation.

### 1. Selecting Components by Eigenvalues

Since the eigenvector corresponding to the largest eigenvalue contains most of the information, it is difficult to achieve a decoupling of appearance and motion. As shown in Fig. 1 after decoupling video features using the eigenvalues, high eigenvalues struggle to reveal distinct appearance contours. Additionally, components with small eigenvalues do not show significant differences in motion intensity. We also analyze the ranking of the eigenvalues corresponding to the components with the highest cosine similarity. Fig. 2 illustrates that the components with the highest similarity do not correspond to the largest eigenvalues. These results demonstrate that eigenvalues cannot serve as a condition for decoupling, and our cosine similarity selection method also exhibits no significant linear relationship with the eigenvalues.

### 2. Attention Entropy in Long Video Generation

Attention entropy is proposed by [2] to solve variable-sized text-to-image synthesis problems, which has not been used in long video generation. The specific derivation process can be found in the aforementioned paper; here we provide the implementation details in our setting. The calculation process in temporal attention can be represented by the following equation:

$$TempAtten(Q, K, V) = softmax(\lambda \frac{QK^T}{\sqrt{d}})V, \quad (1)$$

where  $Q$ ,  $K$ ,  $V$  represent query, key and value in attention module, respectively. And  $\lambda = 1$  in common attention module. However, due to the variation in the number of frames,

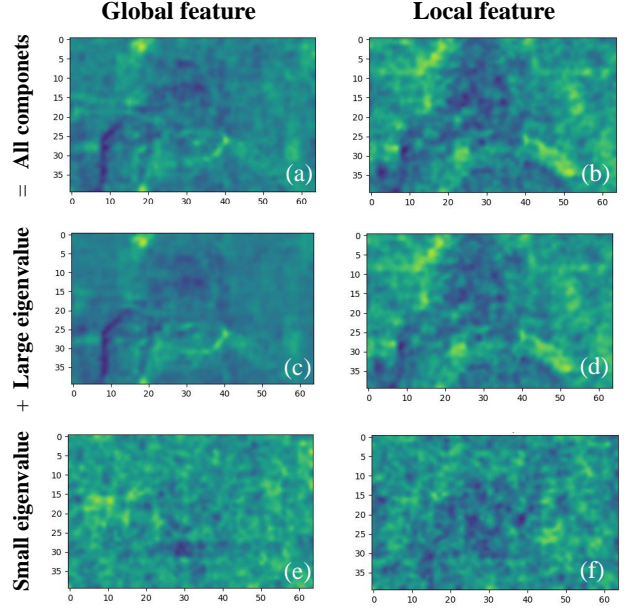


Figure 1. Visualization of features extracted in the principal component space using eigenvalue. We use the same  $k$  value as the cosine similarity method to separate consistency information. It can be seen that (c) and (d) struggle to display distinct contours, while (e) and (f) are quite similar, making it difficult to show significant differences in motion intensity. Therefore, using eigenvalues proves ineffective for the decoupling of video appearance and motion, which further demonstrates the effectiveness of our cosine similarity method for decoupling.

the attention entropy changes, which can be expressed as:

$$Entropy(A) = \log F - \frac{\sigma^2}{2}, \quad (2)$$

where  $A$  represents attention map,  $F$  represents the number of tokens and refers to changing from 16 frames to 64 frames in our experiments.  $\sigma$  is related to  $\lambda$ ,  $Q$  and  $K$ , which can be found in [2] for analytic expression. To ensure the attention entropy unchanged, we set  $\lambda = \sqrt{\log_f F}$  according to [2], where  $f = 16$  and  $F = 64$ .

We also provide both quantitative and qualitative results from comparative experiments on whether to use attention entropy coefficient, as shown in Tab. 1 and Fig. 3. It's important to highlight that our method still surpasses existing approaches, even in the absence of the coefficient for attention entropy.

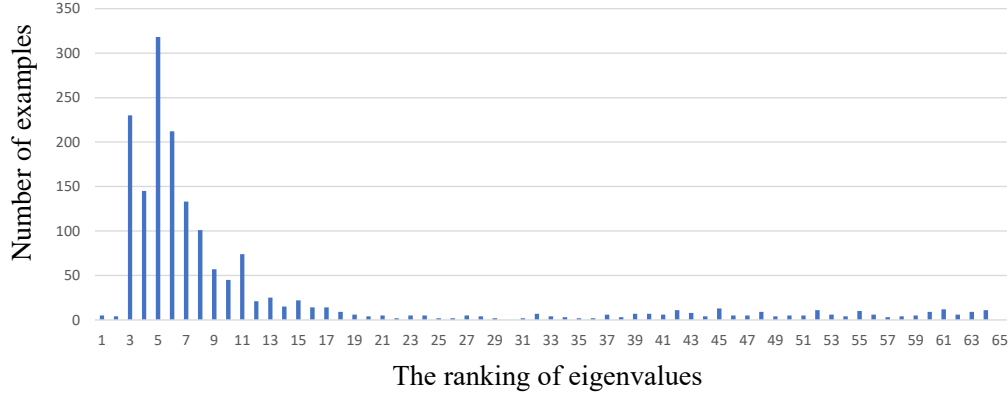


Figure 2. Relationship between the highest cosine similarity with the ranking of eigenvalues. From a statistical perspective, it can be observed that the components with the highest similarity do not correspond to the components with the largest eigenvalues.

Table 1. Quantitative Comparison of whether we use attention entropy. The best values are shown in **bold**. It is worth noting that even without the coefficient for attention entropy, our method still outperforms existing approaches.

Methods	Video Consistency			Video Quality		
	Sub( $\uparrow$ )	Back( $\uparrow$ )	Over( $\uparrow$ )	Motion( $\uparrow$ )	Dynamic( $\uparrow$ )	Imaging( $\uparrow$ )
$\lambda = 1$	94.29	94.81	25.35	95.56	56.94	63.46
Ours( $\lambda = \sqrt{\log_f F}$ )	<b>95.54</b>	<b>95.24</b>	<b>25.69</b>	<b>96.41</b>	<b>59.72</b>	<b>63.70</b>

### 3. Result in DiT framework

In Tab. 2, we evaluate our method with the DiT architecture (Open-Sora) using 64 frames. Our approach enhances long video generation quality and shows generalizability across various backbones.

Table 2. Quantitative Comparison by **VBench-Long**.

Config	Sub $\uparrow$	Back $\uparrow$	Motion $\uparrow$	Imaging $\uparrow$
DiT(Open-Sora)	96.36	96.56	98.09	64.28
DiT+Ours	<b>96.72</b>	<b>97.19</b>	<b>98.67</b>	<b>64.73</b>

### 4. Details of Experiments

In our experiments, we set the DDIM sampling steps to 50, the unconditional guidance scale to 12, and the video output frame rate to 28 fps. The prompts used are selected according to Vbench[1]’s suggestions. Additionally, to make over consistency metric more convincing, we also test the prompts corresponding to subject consistency for over consistency. The experiments are conducted on a single RTX 4090 GPU, and a random seed is set from a uniform distribution for five repeated trials.

In the qualitative results, each frame was selected at equal intervals(i.e., 16 frames). Here, we have included additional qualitative comparison results in Fig. 4, Fig. 5 and Fig. 6, further demonstrating that our FreePCA method out-

performs existing training-free long video generation methods.

### 5. Details of Ablation

For (2) removing the PCA process, we remove the process of mapping features to the principal component space and mapping it back, while retaining all other processes, including the selection of consistent features using cosine similarity. It can be observed that even without using PCA, acceptable results can still be achieved. However, the powerful decoupling ability of PCA can work in conjunction with cosine similarity to achieve better decoupling and thus obtain results with higher quality and consistency.

For (3) substituting the cosine similarity selection with random selection, We used a uniform distribution to randomly select components as consistent features, while maintaining the same value of  $k$  as in the original experiment.

For (5) replacing the reuse mean statistics with direct reuse, we use noise rescheduling from [5] to show the improvement of our method. The improvement in results shows that reusing only the mean statistics can enhance consistency, while preserving better flexibility in the video generation process, as noise rescheduling is a more stringent reuse method.

For (6) removing the reuse mean statistics, we sample the initial noise for all video sequences according to a standard

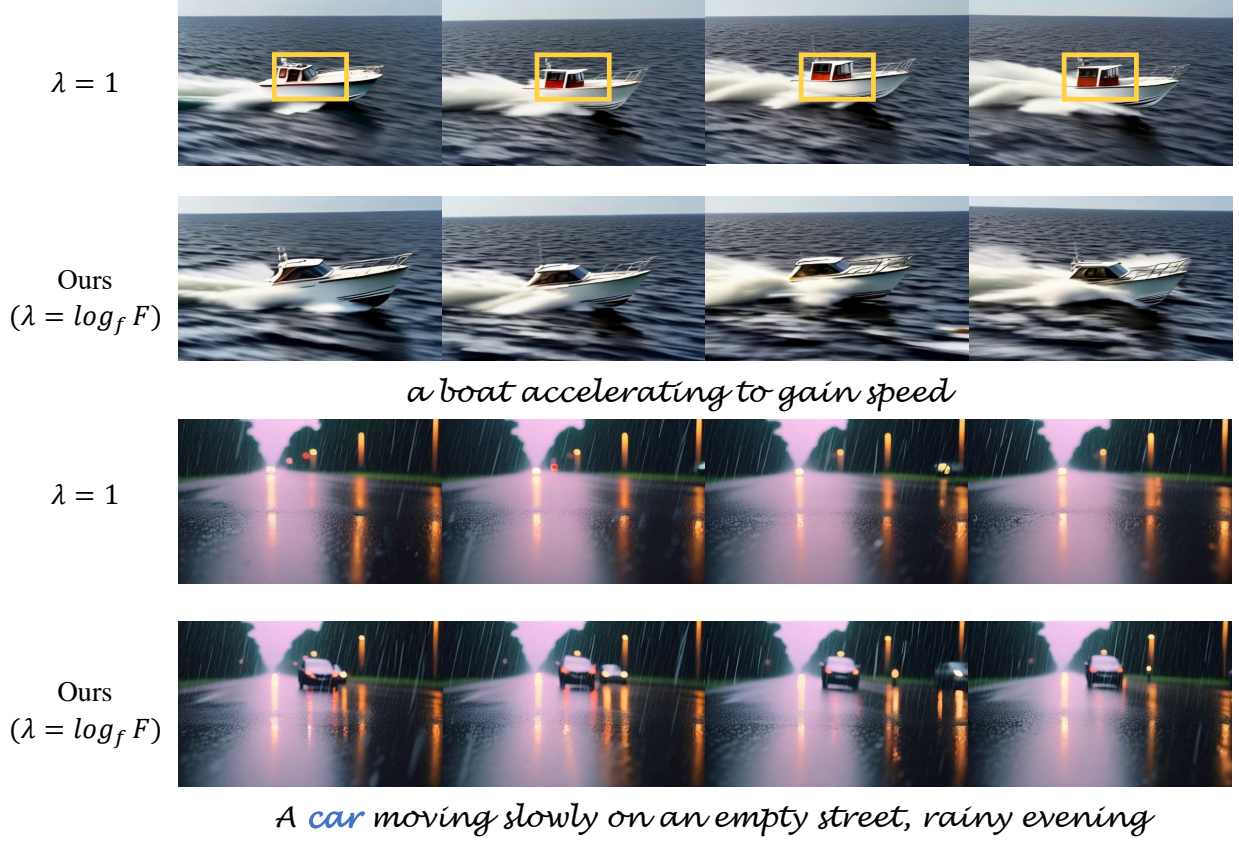


Figure 3. Qualitative results of the attention entropy. If the  $\lambda = \sqrt{\log_f F}$  that maintains a constant attention entropy is not used, inconsistencies between frames and semantic loss may occur due to the shift of distribution.

Gaussian distribution as the original way.

Fig. 7 and Fig. 8 presents the qualitative results of the ablation experiments, showing that these settings negatively impact video quality and video consistency.

## 6. Details for Multi-prompt Video Generation and Continuing Video Generation

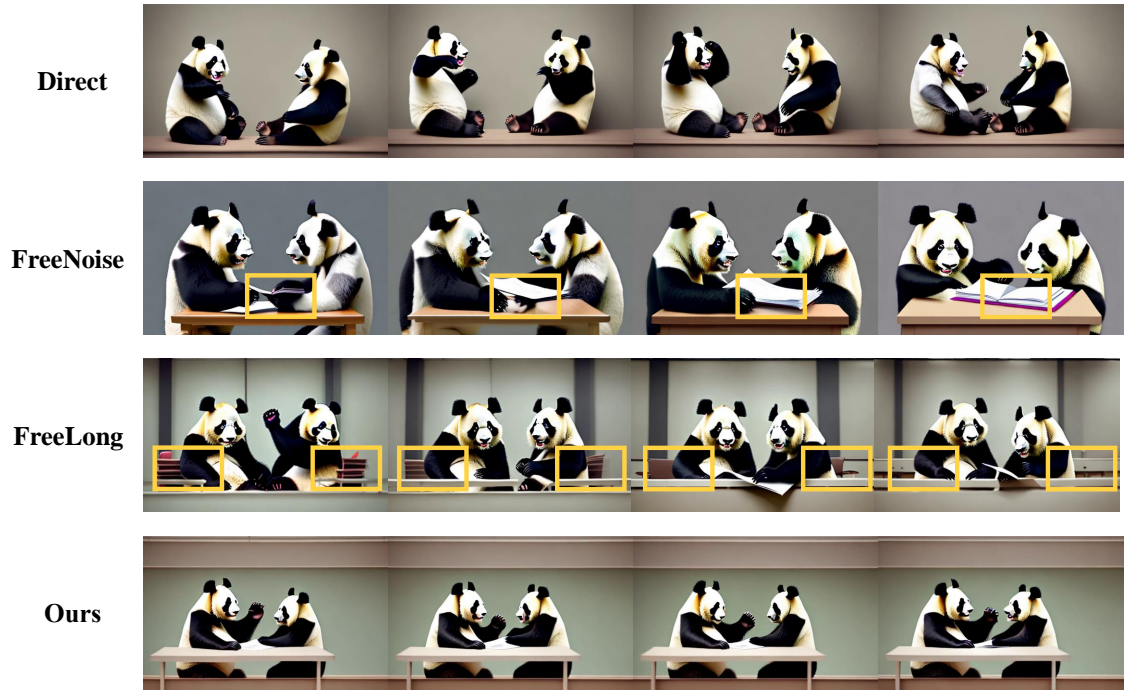
For multi-prompt video generation, we use the same setting as [5]. In most of the denoising steps, the first prompt is used, while the second prompt is applied only at specific steps. This approach allows for the generation of the same layout with different object shapes and motions.

For continuing video generation, we apply DDIM inversion[4] to the given 16-frame video and saved the noise from the last 10 steps, replacing the noise from the first 10 steps of the DDIM process for the first 16 frames of the long video generation (i.e., 64 frames). All other settings remain consistent with the original method.

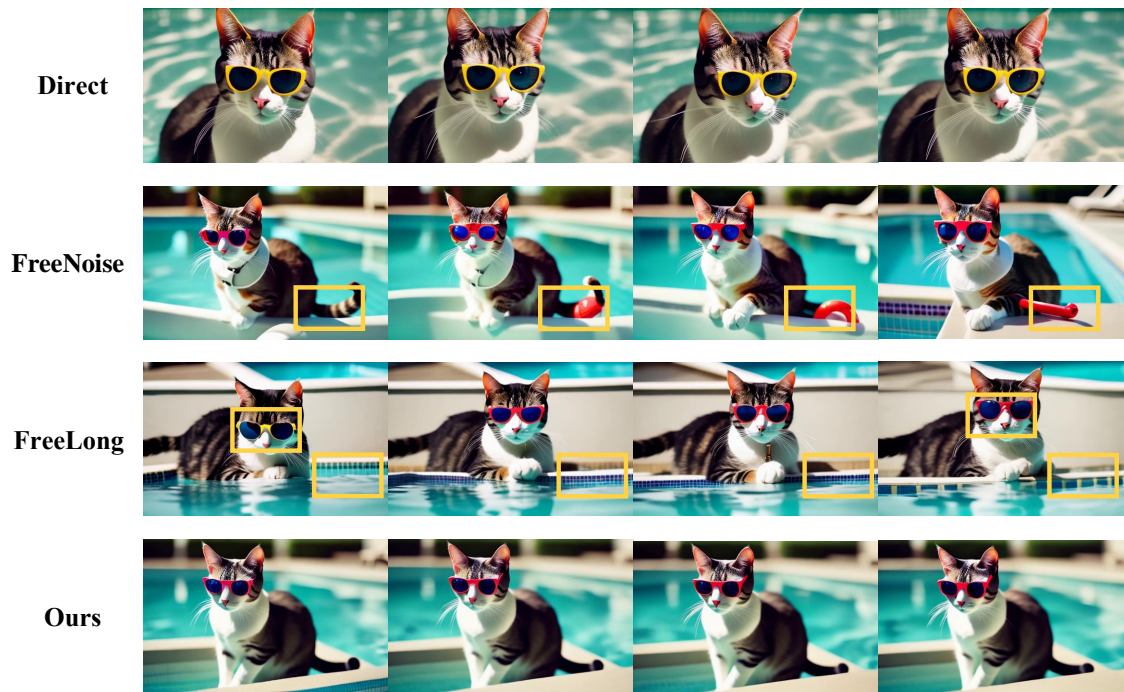
## References

- [1] Ziqi Huang, Yanan 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. 2
- [2] 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. 1
- [3] 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. 4, 5, 6
- [4] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023. 3
- [5] 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*





*Two pandas discussing an academic paper.*



*A cat wearing sunglasses and working as a lifeguard at a pool.*

Figure 4. More qualitative comparisons with existing methods. “Direct” indicates directly sampling 64 frames based on short video generation models. FreeNoise[5] and FreeLong[3] are advanced training-free long video generation methods for comparison.

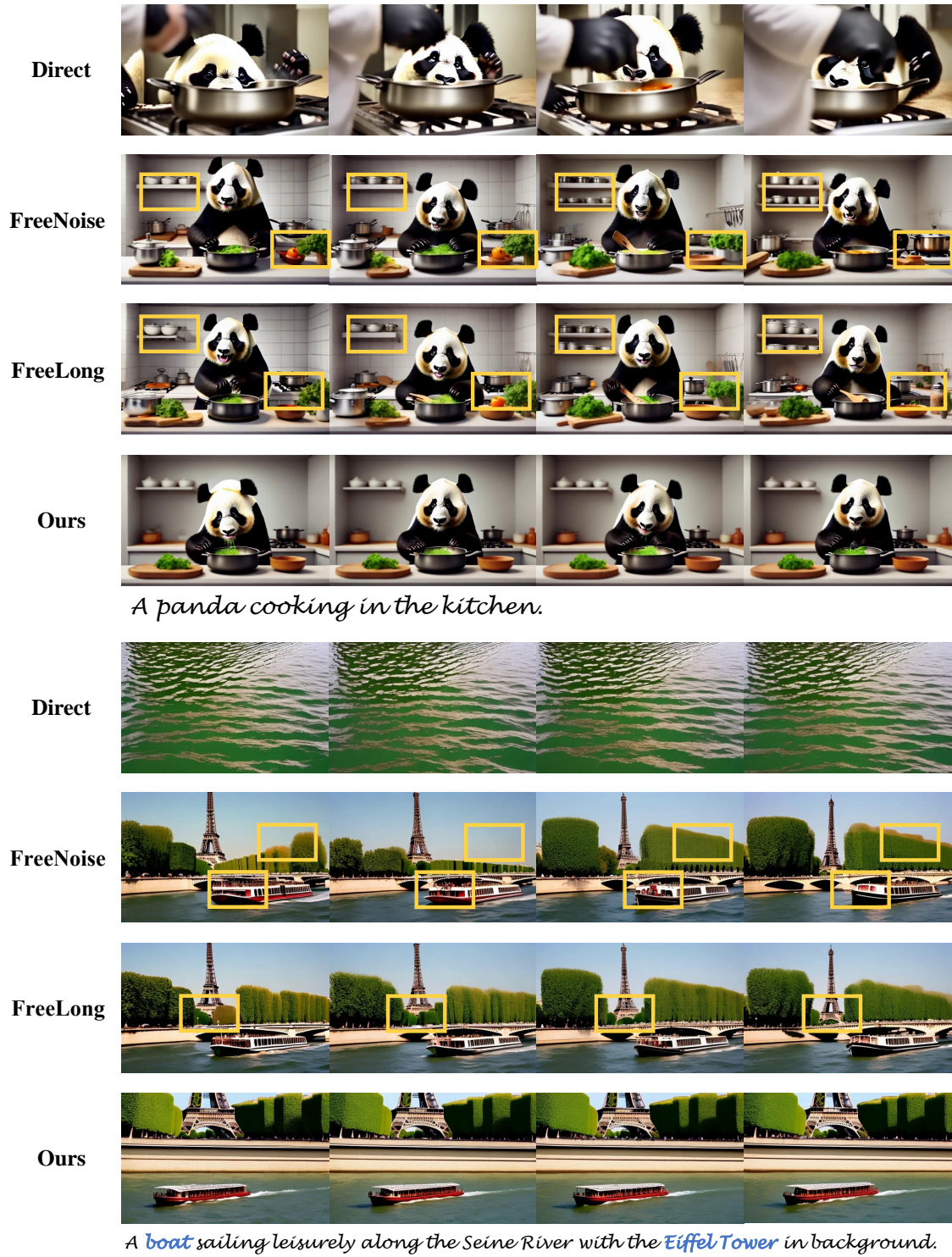
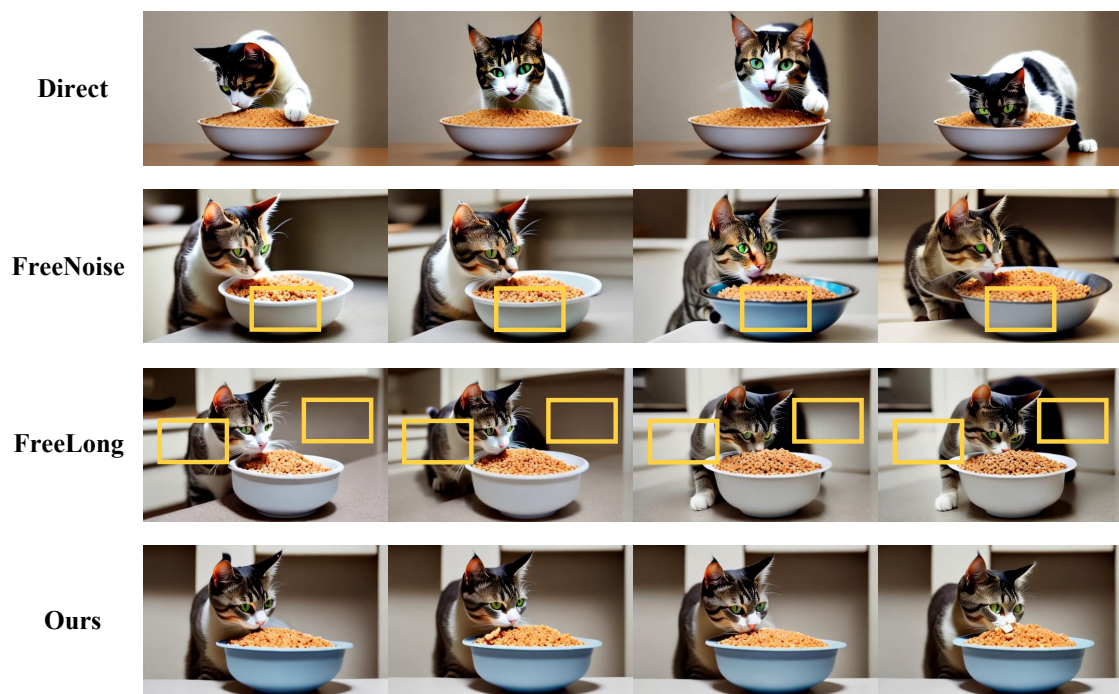
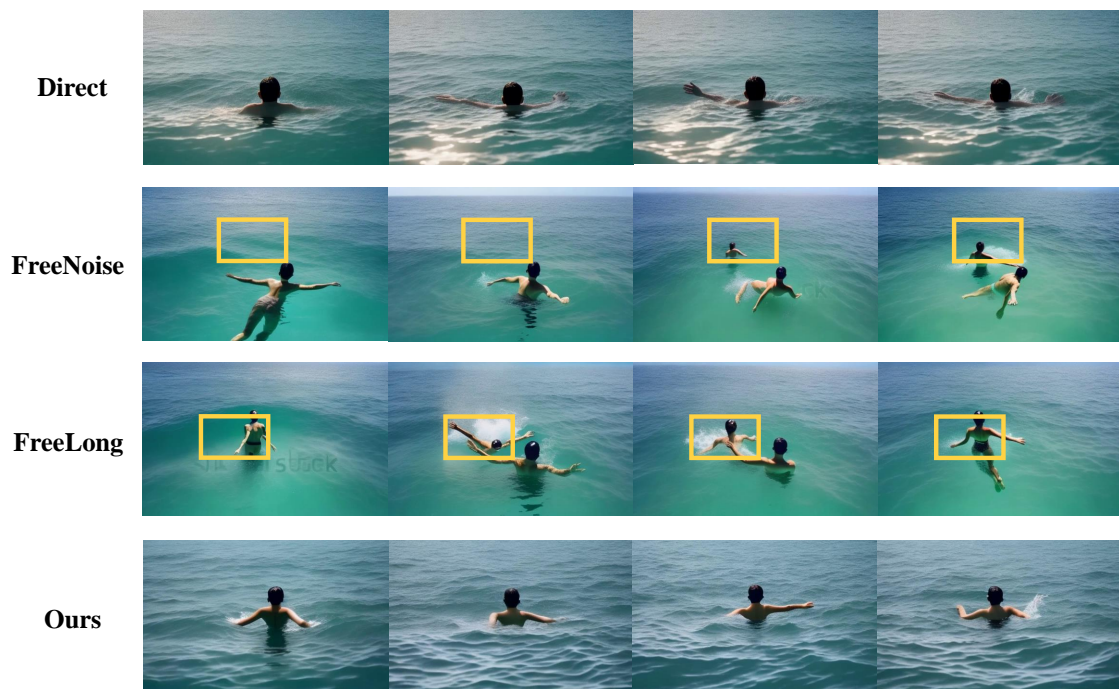


Figure 5. More qualitative comparisons with existing methods. “Direct” indicates directly sampling 64 frames based on short video generation models. FreeNoise[5] and FreeLong[3] are advanced training-free long video generation methods for comparison.





*A cat eating food out of a bowl.*



*A person swimming in ocean.*

Figure 6. More qualitative comparisons with existing methods. “Direct” indicates directly sampling 64 frames based on short video generation models. FreeNoise[5] and FreeLong[3] are advanced training-free long video generation methods for comparison.



Figure 7. Qualitative results of the ablation. (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. Prompts: a motorcycle cruising along a coastal highway.



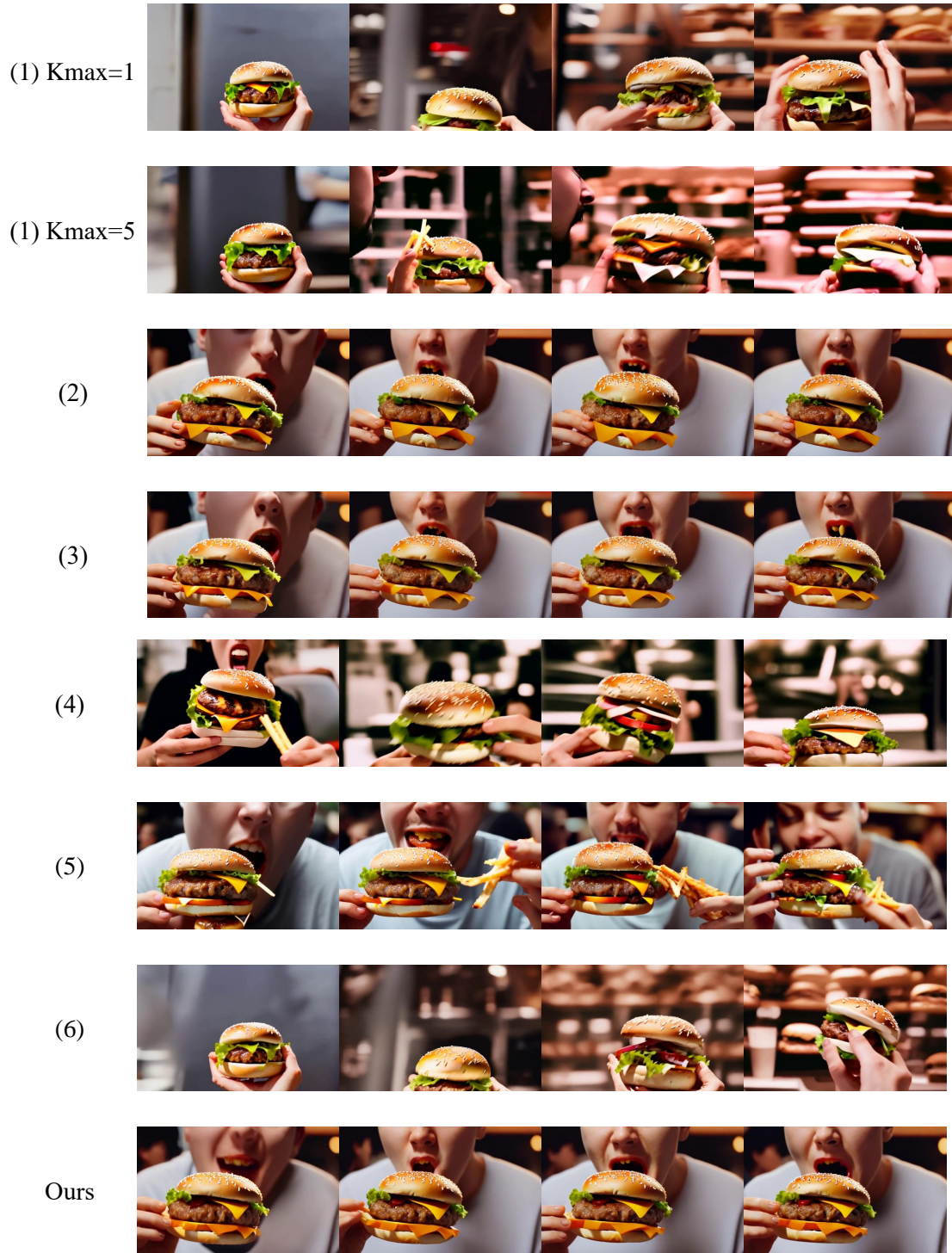


Figure 8. Qualitative results of the ablation. (1) The choice of  $K_{max}$ . (2) Removing the PCA process. (3) Substituting the cosine similarity selection with random selection. (4) Setting  $k = 3$  as a fixed value. (5) Replacing the reuse mean statistics with direct reuse. (6) Removing the reuse mean statistics. Prompts: a person eating a burger.