LongDiff: Training-Free Long Video Generation in One Go

Supplementary Material

1. More Implementation Details

In our main paper, we equip LaVie [14] and VideoCrafter [2] with our LongDiff to generate 128-frame (i.e., $N = 128$) long videos. Following [9, 10], during sampling, we employ the noise shuffle mechanism and perform DDIM sampling with 50 denoising steps. We set G in Eq.(3) to 16. The weighting factor α in Eq.(10) is set to 2. We set the neighbor range L in Eq.(11) to 8. In addition to the neighbor frames, we also select $n = 8$ key frames for temporal attention computation. Notably, we uniformly sample 50% of the temporal attention layers in the short video model and replace them with our LongDiff module. All experiments are conducted using NVIDIA 6000 Ada GPUs.

2. More Evaluation Metrics Details

Following FreeLong [9], we use metrics from VBench [6] to evaluate video quality. For video consistency, we report: 1) Subject Consistency (SC), measured by DINO [1] feature similarity across frames, to check object appearance stability, and 2) Background Consistency (BC), calculated with CLIP [11] feature similarity across frames. For video fidelity, we assess 1) Motion Smoothness (MS) using AMT [8] motion priors, 2) Temporal Flickering (TF) via mean absolute difference between static frames, and 3) Imaging Quality (IQ), measured by MUSIQ [7]. For video-text consistency, we employ Overall Consistency (OC) from ViCLIP [15] to capture both semantic and style information.

3. Additional Ablation Studies

We here conduct more ablation experiments about our LongDiff based on the short video model LaVie.

3.1. Impact of the Number of Position Groups

In Position Mapping (PM), we map $2N-1$ (from $-(N-1)$ to $(N-1)$) original relative positions into $2G-1$ groups to make the model avoid handling large numbers of distinct positions. In our main paper, we set $G = 16$ for LaVie. Here we evaluate other choices of G, and report the results in Tab. 1. We find that performance improves when we increase G, until G reaches 16, where the improvement tapers off. Thus, we set G = 16.

3.2. Impact of the Number of Key Frames

In our LongDiff, we establish temporal correlations between each frame and its neighboring frames, as well as n key frames selected using a key-frame detection pipeline. Here, we also evaluate other choices of n. As shown in Tab. 2, the model performance reaches optimal results at $n = 8$. Thus, we set $n = 8$ in the experiments to achieve a good result.

Method	SC ↑	BC ↑	MS ↑	TF ↑	IQ ↑	OC ↑
$G = 8$	93.47	95.68	95.72	94.19	66.53	23.77
$G = 12$	96.19	97.17	96.74	95.74	67.88	24.42
$G = 16$	98.10	98.23	97.46	96.84	68.83	25.24
$G = 20$	97.25	97.76	97.14	96.45	68.41	24.98

Table 1. Ablation study for the number of position groups.

Method	SC ↑	BC ↑	MS ↑	TF ↑	IQ ↑	OC ↑
$n = 4$	92.96	95.98	96.97	94.67	67.65	24.65
$n = 6$	97.19	97.52	97.18	95.74	68.11	25.01
$n = 8$	98.10	98.23	97.46	96.84	68.83	25.24
$n = 10$	97.52	97.85	97.29	96.18	68.31	25.13

Table 2. Ablation study for the number of key frames.

3.3. Impact of the Number of Neighbor Frames

Here, we also explore the impact of using different numbers L of neighbor frames for temporal attention computation. As shown in Tab. 3, the model's performance reaches its highest value at $L = 8$. We thus set $L = 8$ in our experiments.

Method	SC ↑	BC ↑	MS ↑	TF ↑	IQ ↑	OC ↑
$L = 2$	94.11	96.18	95.28	94.54	65.77	23.61
$L = 4$	96.90	97.61	96.84	96.15	67.91	24.76
$L = 8$	98.10	98.23	97.46	96.84	68.83	25.24
$L = 16$	96.43	97.37	96.55	95.87	67.58	24.59

Table 3. Ablation study for the number of neighbor frames.

3.4. Impact of the Mechanism to Down-Sample Video Features

In our LongDiff, we use a combination of max-pooling, average-pooling, and min-pooling operations to reduce the channel dimension of the video feature F to three channels, aligning it with the input shape required by the key-frame detection pipeline. Here, we evaluate the efficacy of this mechanism by comparing the following variants: 1) **Max**, where only the max-pooling operation is used, and the result is replicated three times along the channel dimension. 2) **Min**, where only the min-pooling operation is used, and

Method	SC ↑	BC ↑	MS ↑	TF ↑	IQ ↑	OC ↑
Max	97.28	97.70	97.02	96.58	68.37	25.06
Min	96.71	97.33	96.71	96.40	68.33	24.93
Average	97.67	97.95	97.23	96.73	68.68	25.14
Ours	98.10	98.23	97.46	96.84	68.83	25.24

Table 4. Ablation study for the mechanism to downsample video features.

the result is replicated three times along the channel dimension. 3) **Average**, where only the average-pooling operation is used, and the result is replicated three times along the channel dimension. As shown in Tab. 4, we observe that using the combination of max-pooling, average-pooling, and min-pooling operations achieves the best performance. Notably, all of these variants of our LongDiff consistently outperform the previous state-of-the-art methods[9, 10].

3.5. Impact of the Weighting Factor α

In our LongDiff, we use two measures—image entropy and frame differencing—to select the most important frame in each shot of the pseudo-video as a key frame. Here, we evaluate the impact of α, which weights these two measures, and report the results in Tab. 5. As shown, the model achieves the best performance with $\alpha = 2$. Therefore, we set α to 2 in our experiments to obtain optimal results.

Method	SC ↑	BC ↑	MS ↑	TF ↑	IQ ↑	OC ↑
$\alpha = 0$	97.10	97.58	96.92	96.53	68.47	25.02
$\alpha = 1$	97.44	97.83	97.21	96.69	68.61	25.11
$\alpha = 2$	98.10	98.23	97.46	96.84	68.83	25.24
$\alpha = 3$	97.94	98.12	97.37	96.79	68.77	25.20

Table 5. Ablation study for the weighting factor α.

3.6. Impact of the Key-Frame Selection Measures

In LongDiff, we use the combination of two measures, image entropy and frame differencing, to select informative key frames by comparing the following variants: 1) **w/o Entropy**, where only image entropy is used as the sole measure to select key frames. 2) **w/o Differencing**, where frames are selected solely based on the frame differencing measure. As shown in Tab. 6, the combination of these two measures yields the best results. In addtion, both variants outperform previous training-free methods[9, 10].

3.7. Impact of the Proportion of LongDiff Modules

In our main experiments, we uniformly replace 50% of the temporal attention layers in the short video model with our LongDiff modules. Here, we explore the impact of varying the proportion of the number of replaced temporal attention layers with LongDiff modules. As shown in Tab. 7, the performance improves noticeably when the proportion of LongDiff is below 50%, and the improvement trend plateaus beyond this point. Based on this observation, we choose to uniformly replace 50% of the temporal attention layers with our LongDiff modules to achieve good results while maintaining efficiency.

Method	SC ↑	BC ↑	MS ↑	TF ↑	IQ ↑	OC ↑
w/o Entropy	97.10	97.58	96.92	96.53	68.47	25.02
w/o Differencing	97.35	97.75	97.06	96.61	68.50	25.02
Ours	98.10	98.23	97.46	96.84	68.83	25.24

Table 6. Ablation study for the key-frame selection measures.

Method	SC ↑	BC ↑	MS ↑	TF ↑	IQ ↑	OC ↑
Proportion 25.0%	94.86	96.14	96.07	95.30	66.53	23.94
Proportion 50.0%	98.10	98.23	97.46	96.84	68.83	25.24
Proportion 75.0%	98.40	98.51	97.63	96.98	68.92	25.25

Table 7. Ablation study for the proportion of LongDiff modules.

4. More Qualitative Results

In this section, we provide more qualitative results regarding the ablation study of the main components of LongDiff (see Fig. 1), longer video generation (see Fig. 2), multi-prompt video generation (see Fig. 3), and more generated videos (see Fig. 4 and Fig. 5).

5. Proofs

5.1. Detailed Proof of Theorem 1

Here, we provide a detailed proof of Theorem 1, which is mainly based on [4]. For ease of reading, we restate Theorem 1 from the main paper below.

Theorem 1. *Define the attention logit function in temporal attention as $f(\mathbf{q}, \mathbf{k}, p)$, which maps the query frame $\mathbf{q}$, key frame $\mathbf{k}$, and their relative position p to a scalar value. Consider a video generation task with N frames, where the model categorizes the $2N-1$ relative positions into $g(N)$ groups. Here, $g(N) \in \mathbb{N}$ is a non-decreasing and unbounded function representing the model's capability to differentiate relative positions. Additionally, assume that any two relative positions p and p' within the same group satisfy $d_f(p, p') \leq \epsilon$, where d_f is the distance function associated with the attention logit function f. Then the following holds:*

$$\sup_{-(N-1)\leq p\leq N-1} |f(\mathbf{q}, \mathbf{k}, p)| \geq \left(\frac{g(N)}{2}\right)^{\frac{1}{2r}} \frac{\epsilon}{4e} \quad (1)$$

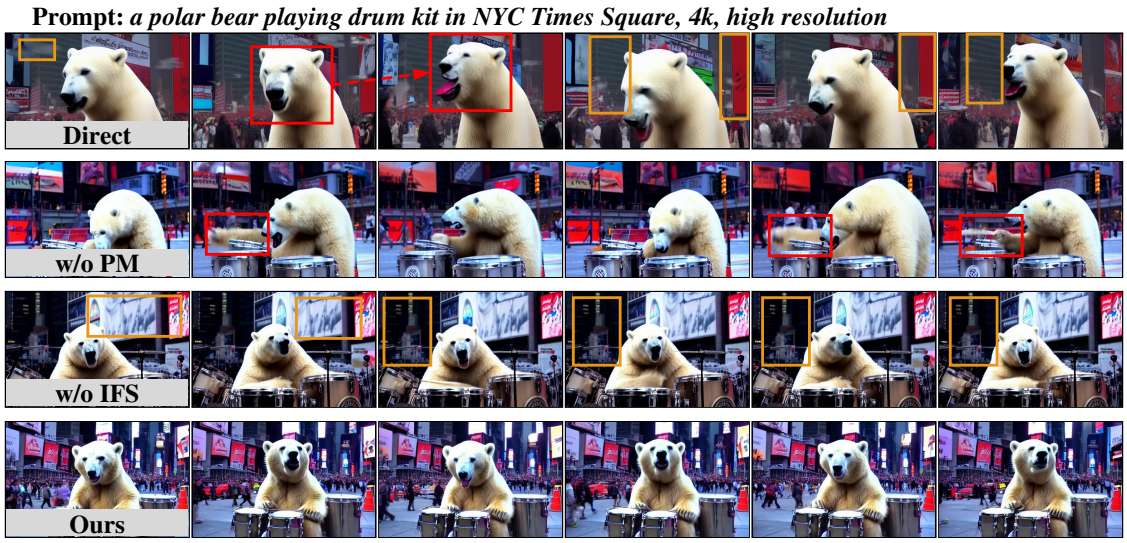

Figure 1. **Ablation Study of the Main Components of LongDiff.** Here, we show the qualitative comparison of LongDiff with two variants: **Ours (w/o PM)**, where we remove the `GROUP` and `SHIFT` operations and thus use the original relative positions to compute temporal attention. 2) **Ours (w/o IFS)**, where we remove the IFS mask in temporal attention computation, requiring each query frame to correlate with all frames for information passing during video generation. As shown, videos generated from the (w/o PM) variant exhibit abrupt temporal transition between frames, particularly noticeable in the bear's hand. On the other hand, videos generated from the (w/o IFS) variant lack some visual details, manifesting as as a blurry "NYC Times Square". We illustrate inferior temporal consistency and the the visual detail issues using red and orange boxes, respectively.

Figure 2. **Longer Video Generation.** Here, we equip VideoCrafter[2] with our LongDiff to generate 256-frame videos. As shown, these generated videos maintain temporal consistency and visual details. This further demonstrates the efficacy of out method.

where r is the pseudo-dimension of the function class $\mathcal{H} = \{f(\cdot,\cdot,p) \mid p \in \mathbb{Z}\}$, and e is the Euler's number.

The distance function d_f can be rewritten in a more detailed form as:

$$d_f(p,p') = \mathbb{E}_{\mathbf{q}\sim\mathbf{Q},\mathbf{k}\sim\mathbf{K}}(f(\mathbf{q},\mathbf{k},p) - f(\mathbf{q},\mathbf{k},p'))^2 \quad (2)$$

where $\mathbf{Q}$ and $\mathbf{K}$ are the trained distributions for $\mathbf{q}$ and $\mathbf{k}$. To assist in proving the inequality in Theorem 1, the following lemma is introduced from [5].

Lemma 1. *Let $\mathcal{H} = \{h(z)\}$ be a family of functions that map a set $\mathbf{Z}$ into $[0, M]$ with pseudo-dimension $\dim_P(\mathcal{H}) = r$, where $1 \le r < \infty$. Let P be a probability measure on $\mathbf{Z}$. Then, for all $0 < \epsilon \le M$, the ϵ-cover of $\mathcal{H}$ under the metric $d(h_1,h_2) = \mathbb{E}_{z\sim P}(h_1(z) - h_2(z))^2$ is bounded by:*

$$\mathcal{N}_P(\epsilon,\mathcal{H},d) \le 2\left(\frac{2eM}{\epsilon}\ln\frac{2eM}{\epsilon}\right)^r \quad (3)$$

where $N_P(\epsilon,\mathcal{H},d)$ is the cover size, defined as the smallest cardinal of a cover-set $\mathcal{H}'$ such that for every entry $h \in \mathcal{H}$, there exists at least one entry $h' \in \mathcal{H}'$ within ϵ distance from h.

Based on Lemma 1, Theorem 1 can be proven by contradiction as follows.

Proof. First, let the negation of Eq. (1) in Theorem 1 be assumed to hold:

$$\sup_{-(N-1)\le p\le N-1} |f(\mathbf{q},\mathbf{k},p)| < \left(\frac{g(N)}{2}\right)^{\frac{1}{2r}} \frac{\epsilon}{4e} = a \quad (4)$$

This indicates that the function family $\mathcal{H} = \{f(\cdot,\cdot,p)|p \in \mathbb{Z}\}$ maps the input to the range $[-a,a]$. Without loss of generality, all values from the range $[-a,a]$ can be shifted to the range $[0,2a]$ to apply Lemma 1. Then, according

Prompt 1: ***A waterfall flows in the mountains under a clear sky***
Prompt 2: ***A waterfall flows in the fall mountains under a clear sky***

Prompt 1: ***There is a beach where there is no one***
Prompt 2: ***The waves hit the deserted beach***
Prompt 3: ***There is a beach that has been swept away by waves***

Figure 3. **Multi-Prompt Video Generation.** Our LongDiff can be easily adapted for multi-prompt video generation by assigning distinct prompts to each video segment following [9, 10]. As shown, the output of our LongDiff maintains temporal consistency and visual details across different segments.

to Lemma 1[1], the ϵ-cover size $\mathcal{N}_P(\epsilon, \mathcal{H}, d_f)$ of $\mathcal{H}$ satisfies that:

$$\mathcal{N}_P(\epsilon, \mathcal{H}, d_f) \leq 2\left(\frac{4ea}{\epsilon}\ln\frac{4ea}{\epsilon}\right)^r \tag{5}$$

By substituting $a = \left(\frac{g(N)}{2}\right)^{\frac{1}{2r}}\frac{\epsilon}{4e}$ (defined in Eq. (4)) into Eq. (5), the following expression is obtained:

$$\begin{aligned}\mathcal{N}_P(\epsilon, \mathcal{H}, d_f) &\leq 2\left(\left(\frac{g(N)}{2}\right)^{\frac{1}{2r}}\ln\left(\frac{g(N)}{2}\right)^{\frac{1}{2r}}\right)^r \\ &< 2\left(\left(\frac{g(N)}{2}\right)^{\frac{1}{2r}}\left(\frac{g(N)}{2}\right)^{\frac{1}{2r}}\right)^r = g(N)\end{aligned} \tag{6}$$

This indicates that, if the assumption in Eq. (4) holds, the ϵ-cover size $\mathcal{N}_P(\epsilon, \mathcal{H}, d_f)$ is smaller than $g(N)$. In other words, we cannot find $g(N)$ distinct functions in the function family $\mathcal{H} = \{f(\cdot,\cdot,p) \mid p \in \mathbb{Z}\}$ such that the pairwise distances (measured by d_f) between them are greater than ϵ. This implies that the number of distinct relative positions differentiated by the model is less than $g(N)$, which contradicts the definition of $g(N)$. Therefore, Eq. (4) does not hold, and thus Eq. (1) in Theorem 1 is proven. □

Pseudo-Dimension of $\mathcal{H}$. As discussed above, Lemma 1 is introduced to assist in proving Theorem 1, which requires that $\mathcal{H}$ has a bounded pseudo-dimension $\dim_P(\mathcal{H}) = r$. Notably, $\mathcal{H} = \{f(\cdot,\cdot,p) \mid p \in \mathbb{Z}\}$ represents the family of attention logit functions, whose form varies depending on the RPE mechanisms. For RoPE, the logit function $f(\cdot,\cdot,p)$ can be expressed as a weighted sum of a finite set of sinusoidal functions $\{\sin(\omega_i p), \cos(\omega_i p)\}$, where the size of this set equals the feature dimension k. Based on the properties of pseudo-dimensions, it follows that $\dim_P(\mathcal{H}_1 + H_2) \leq \dim_P(\mathcal{H}_1) + \dim_P(\mathcal{H}_2)$, and the pseudo-dimension of scaling a single function is at most 2. Therefore, the pseudo-dimension of the whole family is bounded by $\dim_P(\mathcal{H}) \leq 2k$, which satisfies the requirement in Lemma 1.

Analysis of Theorem 1. Theorem 1 implies that, the ability of a video model to distinguish between different relative positions is constrained by the supremum of the model's temporal attention logits. Building on Theorem 1, here, we further analyze whether existing video models can accurately identify frame order during long video generation. Recall that for a video model to correctly identify frame order in a video of length N, it must be capable of distinguishing between $2N-1$ distinct relative positions using its temporal attention logits. According to Theorem 1, this requirement means that a video model capable of correctly identifying frame order must satisfy Eq. (1) when $g(N)$ is set to $2N-1$. Conversely, if the inequality in Eq. (1) fails to hold for $g(N) = 2N-1$, it suggests that the supremum of the temporal attention logits is inadequate for the model to handle $2N-1$ distinct positions. Consequently, the model is unable to correctly identify frame order. Based on the above arguments, here, we perform our analysis taking the LaVie [14] video model as a case study, and use it to directly generate 128-frame (i.e., $N = 128$) videos. Notably, as shown in Eq. (1), to compute it, we need to determine the values of r and ϵ. Below, we then first discuss how we determine the values of r and ϵ for LaVie in our analysis.

Specifically, w.r.t. r, in LaVie, RoPE [12] is employed as the RPE mechanism, and only 32 dimensions (i.e., $k = 32$) of the query and key features are processed by RoPE

[1] A prerequisite for applying Lemma 1 is that $\mathcal{H}$ has a bounded pseudo-dimension $\dim_P(\mathcal{H}) = r$ (i.e., $1 \leq r < +\infty$). It will be shown that $\mathcal{H}$ satisfies this prerequisite later.

in each attention head. Additionally, as discussed earlier, for models using RoPE as the RPE mechanism, $r \leq 2k$ (i.e., $r \leq 64$). Notably, the right-hand side of Eq. (1) is negatively correlated with r. Hence, if $r = 64$ causes the inequality to fail, then the inequality does not hold for any $r < 64$. We then set $r = 64$ for the subsequent analysis here.

Meanwhile, to determine the value of ϵ, we first examine a scenario where these $2N-1$ positions are uniformly distributed to $2N-1$ groups (given $g(N) = 2N-1$). And the boundaries of these clusters (groups) are precisely situated in the middle of two adjacent positions. Consequently, the maximum intra-cluster (group) distance for the cluster that includes position p can be determined by calculating $d_f(p-0.5, p+0.5)$. According to the definition in Theorem 1, ϵ is greater than the maximum intra-cluster distance (measured by d_f) across all position clusters. With the maximum intra-distance of each group, we can determine the lower bound of ϵ, denoted as $\Omega(\epsilon_{\text{uni}})$. Notably, though $\Omega(\epsilon_{\text{uni}})$ is obtained based on the assumption that these $2N-1$ positions are uniformly clustered, for any non-uniformly distributed scenarios, there must exist at least one position cluster of larger size with a greater maximum intra-cluster distance. This means the true lower bound of ϵ is greater than $\Omega(\epsilon_{\text{uni}})$. Additionally, the right-hand side of Eq. (1) is positively correlated with ϵ. Hence, if setting ϵ to $\Omega(\epsilon_{\text{uni}})$ causes the Eq. (1) to fail, then the inequality does not hold for any ϵ. Thus, we here set $\epsilon = \Omega(\epsilon_{\text{uni}})$ for subsequent analysis.

After determining the values of r and ϵ, we extract query and key features from all the temporal attention heads to compute both the left and right sides of Eq. (1). We find that when generating 128-frame videos, only query and key features in 40% of attention heads satisfy the inequality in Eq. (1), and this percentage decreases to 34% when setting $N = 256$. This suggests that the supremum of the temporal attention logits is insufficient for the model to achieve $g(N) = 2N-1$. In other words, the existing video model can struggle in identifying correct frame order.

5.2. Detailed Proof of Theorem 2

Here, we provide a detailed proof of Theorem 2. For ease of reading, we restate Theorem 2 from the main paper below.

Theorem 2. *When generating a video with N frames, the information entropy H of temporal correlations over frames of the video sequence, is lower bounded by [4]:*

$$H\left(\frac{e^{a_i}}{\Sigma_{j=1}^{N} e^{a_j}} | 1 \leq i \leq N\right) \geq \ln N - 2B, \tag{7}$$

where $\{a_i\}_{i=1}^{N}$ are the attention logits with boundary $[-B, B]$.

Proof. The information entropy H of a discrete distribution P is given as $H(P) = -\Sigma_i p_i \ln p_i$. Hence, the information entropy of temporal correlation is computed as follows [4]:

$$\begin{aligned}
&H\left(\frac{e^{a_i}}{\Sigma_{j=1}^{N} e^{a_j}} | 1 \leq i \leq N\right) \\
&\quad = -\sum_i \frac{e^{a_i}}{\sum_j e^{a_j}} \ln \frac{e^{a_i}}{\sum_j e^{a_j}} \\
&\quad = -\sum_i \frac{e^{a_i}}{\sum_j e^{a_j}} \left(a_i - \ln \sum_j e^{a_j}\right) \\
&\quad = -\sum_i \frac{e^{a_i}}{\sum_j e^{a_j}} a_i + \ln \sum_j e^{a_j} \\
&\quad \geq -\max_i a_i + \ln(N e^{-B}) \\
&\quad \geq \ln N - 2B
\end{aligned} \tag{8}$$

□

6. More Experiment Results

6.1. User Study

Following [10], we carried out a user study to assess our results based on human subjective judgment. In this study, participants were shown generated long videos using LaVie as the short video model from all methods (a total of 250 videos), with the examples presented in a random order to eliminate potential bias. Participants were then asked to score the generated videos on a scale of 1 to 5 according to three evaluation criteria: content consistency, video quality, and video-text alignment. The average scores for each method are reported in Tab. 8. As shown, our method received the highest ratings across all metrics.

Method	Content Consistency ↑	Video Quality	Video-Text Alignment
Direct	2.8	1.9	2.3
Sliding	1.8	3.1	2.5
FreeNoise [10]	3.3	3.6	3.5
FreeLong [9]	3.7	3.8	3.9
Ours	4.7	4.6	4.7

Table 8. Comparison based on user study.

6.2. Inference Time

In this section, we compare the inference times (time required for each denoising step) of our LongDiff with other training-free methods [9, 10] and two basic methods, **Direct** and **Sliding**, on the NVIDIA A6000 Ada

Method	Inference Time ↓
Direct	4.0s
Sliding	5.4s
FreeNoise [10]	5.4s
FreeLong [9]	4.7s
Ours	5.5s

Table 9. Comparison of inference time.

GPU. We apply all methods to LaVie and generate 128-frame videos for comparison. As shown in Tab. 9, Our LongDiff significantly improves the quality of long videos generated by the short video model and achieves state-of-the-art results with only a modest increase in inference time compared to the Direct method.

6.3. Evaluation on Video Models with Absolute Positional Encoding

Our LongDiff can also be adapted to video models utilizing absolute positional encoding mechanisms, such as sinusoidal position encoding [13]. This is achieved by performing the `GROUP` and `SHIFT` operations directly on the frame position rather than the relative positions among frames. Here, we take Animatediff [3], which uses sinusoidal position encoding for temporal attention computation, as a case study to evaluate the efficacy of our LongDiff. Specifically, we adapt Animatediff to generate 128-frame long videos with a resolution of 255×255. As shown in Tab. 10, compared to other training-free methods, LongDiff achieves the best performance across all the metrics.

Method	SC ↑	BC ↑	MS ↑	TF ↑	IQ ↑	OC ↑
Direct	92.25	94.35	97.42	96.75	49.27	20.01
Sliding	86.62	92.68	97.86	96.95	60.51	23.42
FreeNoise [10]	95.84	96.75	98.92	98.61	64.69	24.78
FreeLong [9]	95.11	95.86	97.72	98.10	60.23	23.51
Ours	97.54	97.39	98.98	98.70	65.14	25.11

Table 10. Quantitative comparisons of longer video generation (128 frames) on the Animatediff.

References

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 1

[2] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023. 1, 3, 7

[3] 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. 6

[4] Chi Han, Qifan Wang, Hao Peng, Wenhan Xiong, Yu Chen, Heng Ji, and Sinong Wang. Lm-infinite: Zero-shot extreme length generalization for large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3991–4008, 2024. 2, 5

[5] David Haussler. Decision theoretic generalizations of the pac model for neural net and other learning applications. *Information and computation*, 100(1):78–150, 1992. 3

[6] 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. 1

[7] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021. 1

[8] Zhen Li, Zuo-Liang Zhu, Ling-Hao Han, Qibin Hou, Chun-Le Guo, and Ming-Ming Cheng. Amt: All-pairs multi-field transforms for efficient frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9801–9810, 2023. 1

[9] Yu Lu, Yuanzhi Liang, Linchao Zhu, and Yi Yang. Freelong: Training-free long video generation with spectralblend temporal attention. *Advances in Neural Information Processing Systems*, 2024. 1, 2, 4, 5, 6

[10] Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. Freenoise: Tuning-free longer video diffusion via noise rescheduling. In *International Conference on Learning Representations (ICLR)*, 2024. 1, 2, 4, 5, 6

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[12] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 4

[13] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 6

[14] 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, 4, 8

[15] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023. 1

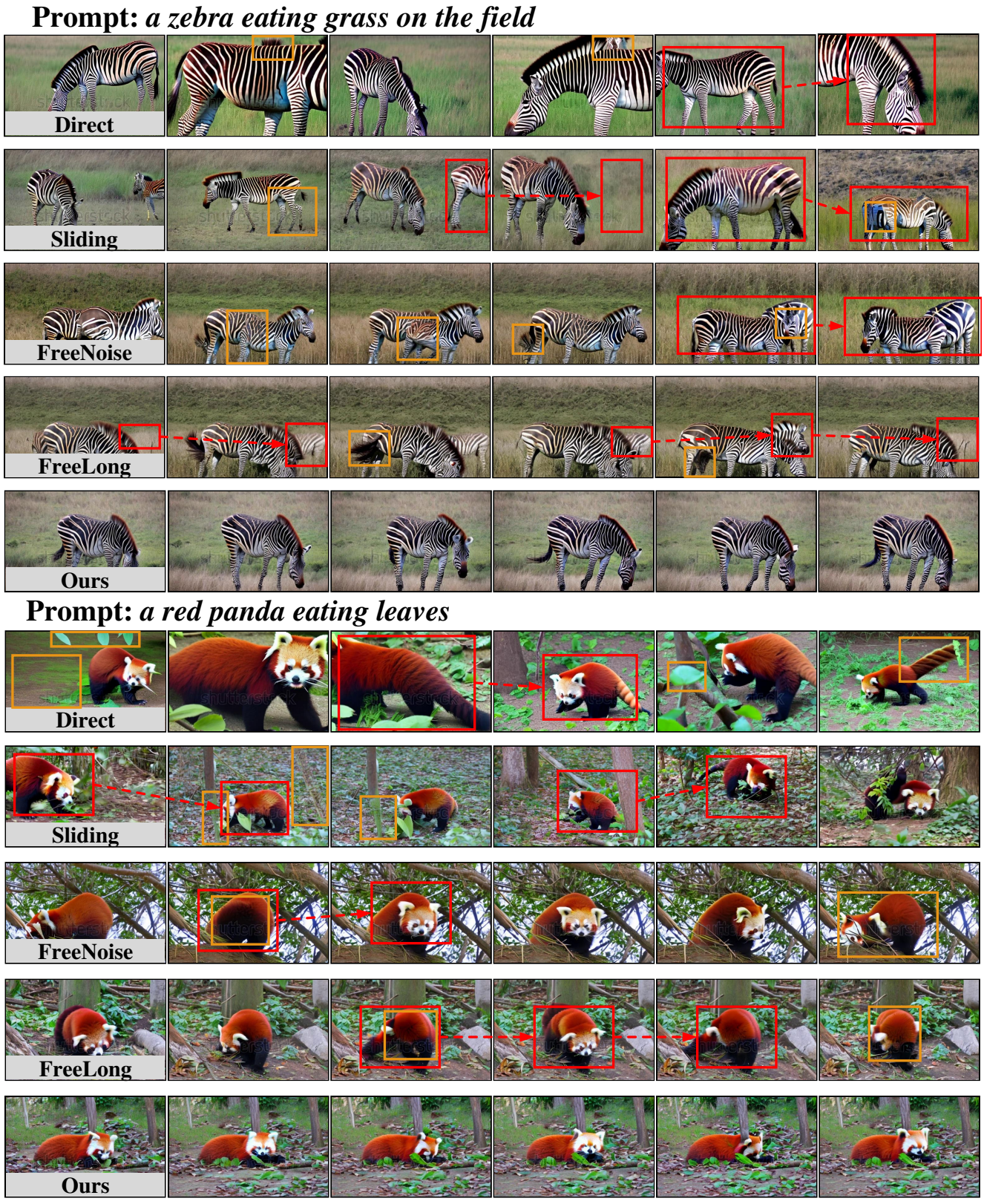

Figure 4. Qualitative comparisons of long video generation (128 frames) based on VideoCrafter[2]. Compared to our LongDiff, videos generated by other methods lack temporal consistency to some extent (e.g., zebras that suddenly appear and disappear in the videos generated from the first prompt; drastic motion changes of the red panda in the videos generated from the second prompt), and suffer from visual detail issues (e.g., blurred zebra bodies in the videos generated from the first prompt; fuzzy leaves and red pandas in the videos generated from the second prompt). We illustrate inferior temporal consistency and visual detail issues using red and orange boxes, respectively.

Figure 5. Qualitative comparisons of long video generation (128 frames) based on LaVie[14]. Compared to our LongDiff, videos generated by other methods lack temporal consistency to some extent (e.g., altered yacht structures in the videos generated from the first prompt; changing river surfaces and trees that suddenly appear and disappear in the videos generated from the second prompt), and suffer from visual detail issues (e.g., the fuzzy yacht in the videos generated from the first prompt; the blurred forest in the videos generated from the second prompt) . We illustrate inferior temporal consistency and the visual detail issues using red and orange boxes, respectively.