Free-Lunch Long Video Generation via Layer-Adaptive O.O.D Correction

Supplementary Material

Contents

1. Verification of Context-Length O.O.D

As mentioned in the main paper, we find that naively extending the frame-level context length (with VRPR applied to constrain the frame-level relative position within pre-trained range) leads to increased attention entropy in the video DiT. This diffusion of attention weights correlates with a degradation in generation quality.

To quantify this, we measure the attention entropy. Assume that we generate a video sequence with a total token number of N, where $N = f \times n$ (f is the total number of frames and n is the number of tokens per frame). Let $p_1, p_2, \cdots, p_N$ be the sequence of attention distributions (after softmax) for all tokens. Each p_i is a vector of attention weights for the i-th token, and this vector has a length of N. We compute the standard Shannon entropy for the attention distribution of each token i as:

$$H(p_i) = -\sum_{j=1}^{L} p_{i,j} \ln(p_{i,j}) \tag{1}$$

To obtain the final average attention entropy, we first compute the mean of $H(p_i)$ over all L tokens, and then average this value across all attention heads and all self-attention layers.

Figure 1 plots the resulting average attention entropy as it varies with the context length (i.e., the total number of frames, f). This illustrates that as the context length increases, the attention entropy also rises, confirming the context-length O.O.D phenomenon.

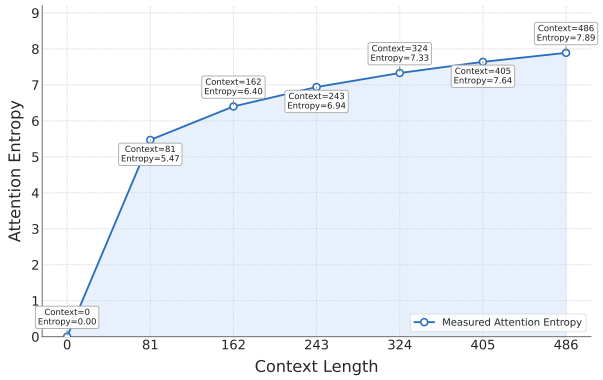


Figure 1. **Layer-wise sensitivity to context-length O.O.D measured via attention entropy differences.**

2. Additional Details of VRPR and TSA

2.1. Implementation Details of VRPR

As stated in the main paper, VRPR is designed to remap O.O.D relative positions back into the pre-trained distribution in a multi-granularity fashion.

The re-encoding functions (Equations 2 and 3 in the main paper) are designed to create a "compressed" representation of relative time. Specifically, for a model pre-trained on L frames (e.g., $L = 81$ for Wan2.1-T2V-1.3B), the pre-trained relative position range is $[-(L-1), L-1]$.

Here, we provide the specific implementation details for Video-based Relative Position Re-encoding (VRPR). Directly modifying the relative position matrix as described in the main paper is practically challenging, To address this, we propose an approximate implementation achieved through the formulation of region-specific position indices. Let $P_q[i]$ denote the position index for the query at frame i after re-encoding, $P_k[j]$ denote the position index for the key at frame j after re-encoding and P_{impl} denote the implemented relative position.

1. Fine-Grained Re-Encoding:
For local interactions within the fine-grained window ($|i - j| \le W_1$), we use standard relative positions:

$$P_q[i] = i, \quad P_k[j] = j \tag{2}$$

In this case, the effective relative position is $P_{\text{impl}} = P_q[i] - P_k[j] = i - j = P$. This provides an exact implementation of the identity mapping $P = P_{\text{ori}}$ described in the main paper.

2. Medium-Grained Re-Encoding:

For interactions in the medium range ($W_1 < i - j \leq W_2$), we apply the following mapping:

$$P_q[i] = \left\lfloor \frac{i}{G_1} \right\rfloor + \left(W_1 - \left\lfloor \frac{W_1}{G_1} \right\rfloor \right), \quad P_k[j] = \left\lfloor \frac{j}{G_1} \right\rfloor \tag{3}$$

Conversely, for negative medium-range distances ($-W_2 \leq i - j < -W_1$):

$$P_q[i] = \left\lfloor \frac{i}{G_1} \right\rfloor, \quad P_k[j] = \left\lfloor \frac{j}{G_1} \right\rfloor + \left(W_1 - \left\lfloor \frac{W_1}{G_1} \right\rfloor \right) \tag{4}$$

This formulation approximates the theoretical formula presented in the main paper (Eq. 2). Specifically, for the positive case, the relative position calculated by this implementation is:

$$P_{\text{impl}} = P_q[i] - P_k[j] = \left(\left\lfloor \frac{i}{G_1} \right\rfloor - \left\lfloor \frac{j}{G_1} \right\rfloor \right) + \left(W_1 - \left\lfloor \frac{W_1}{G_1} \right\rfloor \right) \tag{5}$$

Comparing this to the main paper's target formula $P = \lfloor \frac{i-j}{G_1} \rfloor + (W_1 - \lfloor \frac{W_1}{G_1} \rfloor)$, the equivalence relies on the mathematical property that

$$\left\lfloor \frac{i}{G_1} \right\rfloor - \left\lfloor \frac{j}{G_1} \right\rfloor - 1 \leq \left\lfloor \frac{i-j}{G_1} \right\rfloor \leq \left\lfloor \frac{i}{G_1} \right\rfloor - \left\lfloor \frac{j}{G_1} \right\rfloor. \tag{6}$$

So, the error between P_{impl} and P is $P_{\text{impl}} - P = \left\lfloor \frac{i-j}{G_1} \right\rfloor - (\left\lfloor \frac{i}{G_1} \right\rfloor - \left\lfloor \frac{j}{G_1} \right\rfloor) \in \{0, -1\}$. This approximation error is negligible for our case and allows for the decomposition of relative positions into independent query and key indices. Crucially, this approximation preserves the original granularity (i.e., the group size) and the monotonicity of relative position while effectively constraining the relative position within the pre-trained length, preventing O.O.D issues. With this operation, it also guarantees no discontinuity near the boundary of different regions.

3. Coarse-Grained Re-Encoding:

For distant interactions ($i - j > W_2$), we employ a coarser quantization:

$$P_q[i] = \left\lfloor \frac{i}{G_2} \right\rfloor + \left(W_2 - \left\lfloor \frac{W_2}{G_2} \right\rfloor - \left\lfloor \frac{W_2 - W_1}{G_1} \right\rfloor \right), P_k[j] = \left\lfloor \frac{j}{G_2} \right\rfloor \tag{7}$$

For the negative direction ($i - j < -W_2$):

$$P_q[i] = \left\lfloor \frac{i}{G_2} \right\rfloor, P_k[j] = \left\lfloor \frac{j}{G_2} \right\rfloor + \left(W_2 - \left\lfloor \frac{W_2}{G_2} \right\rfloor - \left\lfloor \frac{W_2 - W_1}{G_1} \right\rfloor \right) \tag{8}$$

Importantly, our dynamic mapping strategy preserves monotonicity property of relative position across region boundaries: tokens at greater distances consistently receive larger absolute relative position values. Furthermore, this index-based formulation allows VRPR to seamlessly integrate with efficient attention implementations such as FlashAttention2 following [2], as it only requires modifying the input position indices rather than the full attention matrix.

The specific hyperparameter configurations for the models evaluated in the main paper are as follows. These parameters include the Local Window size (W_1), the Mid-Range Window size (W_2), Group Size 1 (G_1) and Group Size 2 (G_2).

The principles of hyperparameters selection is that the maximum relative position in the generated video must not exceed the pre-trained context window L. For a target generation length L_{target}, the maximum absolute relative distance is $L_{\text{target}} - 1$. The mapped position $P(L_{\text{target}} - 1)$ must satisfy:

$$P_{\text{impl}}(L_{\text{target}} - 1) \leq L - 1. \tag{9}$$

Substituting the coarse-grained mapping (assuming $D_{max} > W_2$), the criterion is formalized as:

$$\left\lfloor \frac{L_{\text{target}} - 1}{G_2} \right\rfloor + \left(W_2 - \left\lfloor \frac{W_2}{G_2} \right\rfloor - \left\lfloor \frac{W_2 - W_1}{G_1} \right\rfloor \right) \leq L - 1. \tag{10}$$

We select the possible windows W_1, W_2 and group sizes G_1, G_2 that satisfy this inequality to prevent O.O.D issues. The empirical configurations for our experiments are:

For Wan2.1-T2V-1.3B:

- 2×**Extension (161-frame):** Local Window $W_1 = 12$, Mid-Range Window $W_2 = 20$, Group Size 1 $G_1 = 2$ and Group Size 2 $G_2 = 8$
- 4×**Extension (321-frame):** Local Window $W_1 = 10$, Mid-Range Window $W_2 = 14$, Group Size 1 $G_1 = 2$ and Group Size 2 $G_2 = 8$.

For HunyuanVideo:

- 2×**Extension (253-frame):** Local Window $W_1 = 12$, Mid-Range Window $W_2 = 20$, Group Size 1 $G_1 = 2$ and Group Size 2 $G_2 = 4$.
- 4×**Extension (509-frame):** Local Window $W_1 = 12$, Mid-Range Window $W_2 = 20$, Group Size 1 $G_1 = 2$ and Group Size 2 $G_2 = 8$.

These values were determined empirically to provide a good balance between preserving local motion and maintaining global coherence. Noting that we also observed that as long as these parameters satisfy the aforementioned inequality 10, slight adjustments to them (modify a single parameter at a time while maintaining the other three fixed) have minimal impact on the results (performance metrics) as shown on Figure 2, demonstrating relatively good robustness.

2.2. Implementation Details of TSA

TSA is designed to combat the context-length O.O.D problem by preserving attention density. It structures the attention mask into three components. Here, we provide an

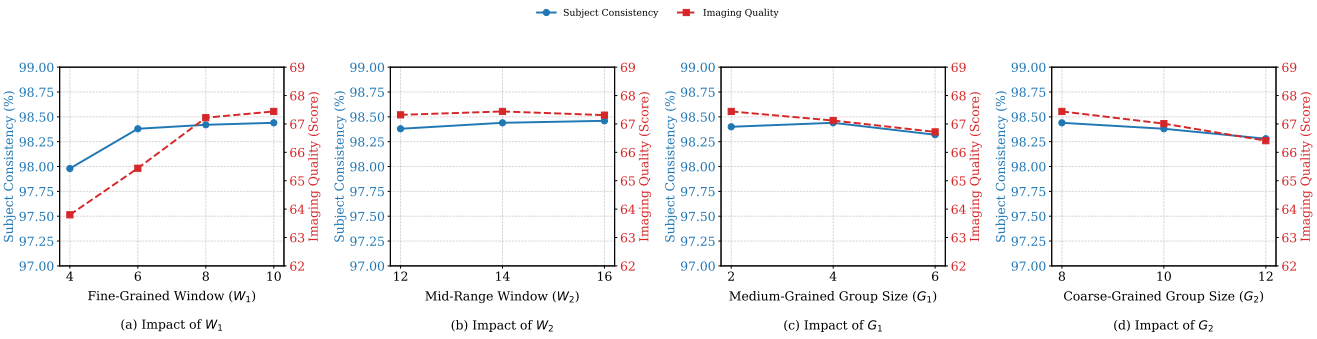


Figure 2. **Impact of W_1, W_2 , G_1 and G_2 on Subject Consistency and Imaging Quality for VRPR**.

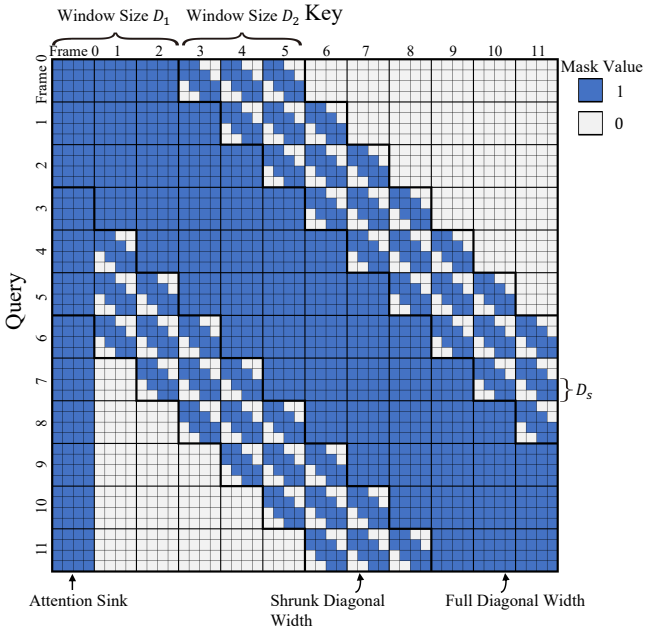


Figure 3. **Attention mask used in TSA**.

enlarged and more detailed view of the attention mask in Figure 3.

Detailed Derivation of Striped Window Width ($D_S = \lfloor \frac{nD_1}{\alpha(D_2-D_1)} \rfloor$). A key innovation of TSA is the adaptive calculation of the spatial window width D_s (which is not explored in detail in the main text). Our theoretical goal is to strictly control the "Attention Mass" (the total number of key tokens involved in calculation) in the mid-range zone to prevent it from dominating the local attention. Let N_{local} be the total number of tokens attended to in the dense local window. Given n tokens per frame and a local window size of D_1, the local token number (attention density) is:

$$N_{\text{local}} = n \times D_1 \tag{11}$$

For the mid-range zone spanning $(D_2 - D_1)$ frames, using dense attention would result in a token count of $n \times (D_2 - D_1)$, which is typically too large. We introduce the α, which defines the decay ratio. Specifically, we mandate that the total attention density in the mid-range (N_{mid}) should be scaled down by α relative to the local density:

$$N_{\text{mid}} = \frac{N_{\text{local}}}{\alpha} \tag{12}$$

Given the spatial strip width D_s, the actual token count in the mid-range is $N_{\text{mid}} = D_s \times (D_2 - D_1)$. Equating these terms allows us to derive the optimal D_s:

$$D_s \times (D_2 - D_1) = \frac{n \times D_1}{\alpha} \implies D_s = \left\lfloor \frac{nD_1}{\alpha(D_2 - D_1)} \right\rfloor \tag{13}$$

The parameter α explicitly controls the trade-off between global context awareness and local detail preservation. α represents the ratio of "Local Attention Density" to "Mid-Range Attention Density". For example, $\alpha = 2$ implies that the aggregate information retrieved from the extended temporal context (D_1 to D_2) is compressed to exactly half the capacity of the immediate local context. By setting $\alpha > 1$, we force the model to prioritize local interactions (which define motion quality) while treating mid-range interactions as supplementary cues.

The specific hyperparameter configurations for the models evaluated in the main paper are as follows. These parameters include the Local Window size (D_1), the Mid-Range Window size (D_2), and the Striped Attention Density (α).

In principle, firstly, the selection of D_1 and α should adhere to the following requirements: $nD_1 + \lfloor \frac{nD_1}{\alpha(D_2-D_1)} \rfloor (D_2 - D_1) \leq nD_1 + \lfloor \frac{nD_1}{\alpha} \rfloor \leq nD_1(1 + \frac{1}{\alpha}) \leq \frac{nL_{\text{pretrained}}}{2}$, i.e., the effective token number stays within the pretrained range, where $L_{\text{pretrained}}$ is the pretrained frame length. Consequently, D_1 and α only need to satisfy $D_1(1 + \frac{1}{\alpha}) \leq \frac{L_{\text{pretrained}}}{2}$. However, an excessively small D_1 would result in an overly limited interaction range for the attention mechanism, which is undesirable for temporal consistency. So we empirically impose the additional constraint that:

$$\frac{L_{\text{pretrained}}}{4} \leq D_1(1 + \frac{1}{\alpha}) \leq \frac{L_{\text{pretrained}}}{2}. \tag{14}$$

For Wan2.1-T2V-1.3B, we set $L_{\text{pretrained}} = 24$, and for HunyuanVideo, we set $L_{\text{pretrained}} = 36$. For the selection of D_2,

it only has to satisfy $D_2 > D_1$, and we will discuss effect of its choice in following ablation study.

Specifically,

For Wan2.1-T2V-1.3B:

- $2\times$**Extension (161-frame):** Local Window $D_1 = 8$, Mid-Range Window $D_2 = 16$, and Striped Attention Density $\alpha = 4$.
- $4\times$**Extension (321-frame):** Local Window $D_1 = 8$, Mid-Range Window $D_2 = 24$, and Striped Attention Density $\alpha = 4$.

For HunyuanVideo:

- $2\times$**Extension (253-frame):** Local Window $D_1 = 12$, Mid-Range Window $D_2 = 24$, and Striped Attention Density $\alpha = 4$.
- $4\times$**Extension (509-frame):** Local Window $D_1 = 12$, Mid-Range Window $D_2 = 36$, and Striped Attention Density $\alpha = 4$.

The **Attention Sink** component is implemented by ensuring that all tokens (queries) from all frames are permitted to attend to the tokens from the initial frame.

3. Layer-wise Sensitivity Profile and Strategy

To rigorously determine the layer-wise application of FreeLOC, we conducted a comprehensive probing experiment. The layer-wise probing was conducted using $N = 10$ diverse prompts, generating $M = 3$ videos per prompt for each perturbation configuration to ensure statistical reliability.

Sensitivity Determination. We identify layers sensitive to frame-level relative position O.O.D based on VisionReward and Attention Logits Difference (ALD), and layers sensitive to context-length O.O.D based on the Context-Length Sensitivity Score. To establish a definitive binary sensitivity classification, we employ a thresholding strategy where the top two-thirds of layers exhibiting the most significant metric degradation (or score increase) are designated as sensitive to each respective O.O.D type.

Profile for Wan2.1-T2V-1.3B. Based on this protocol, the specific sensitivity profile for the 30 layers of Wan2.1-T2V-1.3B is identified as follows:

- **Layers sensitive to Frame-level Relative Position O.O.D (20 layers):**
  $\{28, 1, 29, 0, 27, 26, 21, 17, 20, 12, 16, 14, 8, 4, 13, 19, 10, 23, 5, 3\}$.
- **Layers sensitive to Context-Length O.O.D (20 layers):**
  $\{18, 24, 13, 23, 22, 10, 15, 6, 11, 25, 2, 4, 16, 29, 7, 8, 12\ 9, 3, 21\}$.

The sequence here is arranged in descending order of sensitivity. Notably, we conclude that in this configuration, *every single layer* is sensitive to at least one type of O.O.D issue, necessitating a comprehensive correction strategy.

Layer-wise Strategy Given these overlapping sensitivities, we devised a balanced allocation strategy to optimize performance.

1. For layers identified as sensitive to **Frame-level Relative Position O.O.D**, we apply the **VRPR** strategy.
2. For layers identified as sensitive to **Context-Length O.O.D**, we apply the combined **VRPR + TSA** strategy.

We found that an even distribution—allocating the top 50% of layers (specifically, the 15 layers most sensitive to context length) to the VRPR+TSA strategy, and the remaining 50% to the VRPR-only strategy—yields the relatively optimal balance. Interestingly, the 16 layers identified as context-sensitive in our probing aligns almost perfectly with this 50% allocation.

To validate this choice, we compared this balanced (1/2 VRPR, 1/2 VRPR+TSA) distribution against other ratios:

- **1/3 VRPR, 2/3 VRPR+TSA**: Resulted in over-sparsification, harming local details.
- **2/3 VRPR, 1/3 VRPR+TSA**: Insufficient context correction, leading to visual details degradation in long videos.

The 50%/50% split proved to be the robust "sweet spot". The detailed visualization of Qualitative comparison is shown in Figure 4.

Profile for Hunyuan Video. Resembling the probing procdedure of Wan2.1-T2V-1.3B, the specific sensitivity profile for the 60 layers of Hunyuan Video is identified as follows:

- **Context-length sensitive layers (40 layers):** These layers exhibit high sensitivity to context length and first 30 layers are assigned the **VRPR+TSA** strategy. The identified layers are:
  $\{1, 2, 45, 29, 5, 13, 38, 59, 11, 34, 47, 16, 0, 53, 21, 41,$
  $8, 30, 55, 19, 4, 49, 26, 42, 12, 57, 23, 25, 36, 51, 15, 39, 7,$
  $22, 44, 18, 32, 9, 27, 35\}$.
- **Frame-level Relative Position-sensitive layers (40 layers):** These layers exhibit sensitivity to relative position shifts. Layers that appear in this list *but not* in the context-length sensitive list are assigned the **VRPR only** strategy. The identified layers are:
  $\{58, 59, 31, 6, 54, 21, 37, 29, 47, 18, 28, 10, 1, 24, 3,$
  $35, 41, 14, 52, 27, 9, 43, 17, 33, 48, 2, 23, 50, 11, 38, 4, 25,$
  $40, 15, 0, 7, 56, 20, 46, 32\}$.

4. More Implementation Details

All experiments were conducted on a single NVIDIA A100 (80GB) GPU. Basic settings for generation are listed in Table 1. For test prompt, we randomly sample 100 prompts from Vbench-long [1] following prior works [3, 4, 6].

5. Additional Ablation Study Results

For efficiency, we conduct following ablation studies on Wan2.1-T2V-1.3B [5] with $4\times$ length extension (similar results can also be obtained on HunyuanVideo).

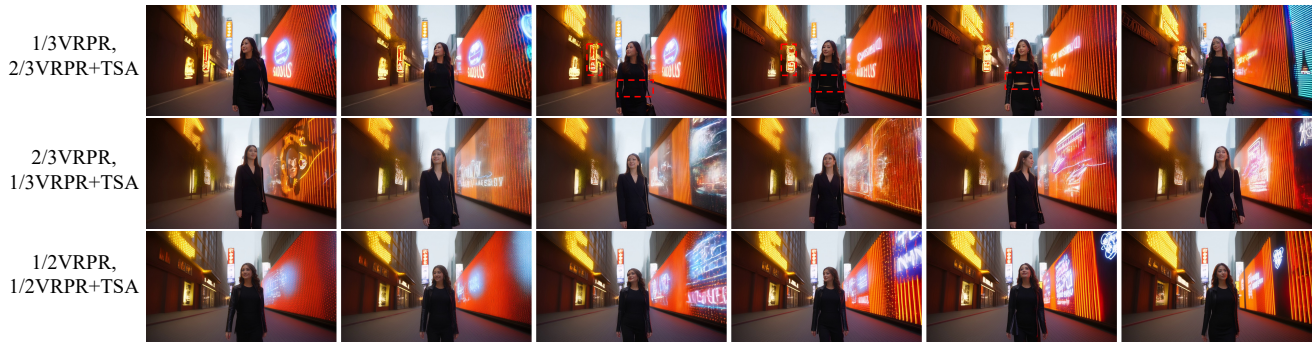


Figure 4. **Qualitative comparison of different allocation strategy**. The (1/2, 1/2) split proved to be the robust "sweet spot" which balance detail preservation and temporal consistency.

Table 1. **Generation Settings.** Basic settings used for Wan2.1-T2V-1.3B and Hunyuan Video for quantitative and qualitative evaluation.

Parameter	Wan2.1-T2V-1.3B	Hunyuan Video
Sampler	Unipc	Euler
Denoising Steps	50	50
CFG Scale	6.0	6.0
Resolution	832×480 (480p)	832×480 (480p)
Base Length (1x)	81 frames	127 frames
2x Length	161 frames	253 frames
4x Length	321 frames	509 frames

5.1. Impact of D_1, D_2 and α for TSA

To investigate the individual contributions of the TSA hyperparameters—local window size D_1, mid-range window boundary D_2, and attention decay factor α—we conducted controlled ablation studies measuring Subject Consistency (SC) to measure temporal consistency and Imaging Quality (IQ) to measure video quality. We conduct following experiments: (1) **Impact of D_1:** Fixing $D_2 = 24$ and $\alpha = 4$, we varied $D_1 \in [5, 9]$ according to Eq 14. (2) **Impact of D_2:** Fixing $D_1 = 8$ and $\alpha = 4$, we varied $D_2 \in [8, 16, 24, 32, 40, 68]$. (3) **Impact of α:** Fixing $D_1 = 8$ and $D_2 = 24$, we varied $\alpha \in [1, 2, 4, 6, 8]$.

Results Analysis: Figure 5 illustrates the trade-offs observed:

- **Varying D_1:** As D_1 increases, SC improves continuously due to stronger local temporal coupling. However, IQ initially improves but subsequently declines, likely because an overly large dense window introduces excessive attention entropy.
- **Varying D_2:** Increasing D_2 initially boosts SC, which then plateaus as the benefit of mid-range context saturates. IQ follows an inverted-U pattern, increasing at first but dropping if the mid-range window becomes too computationally diffuse.
- **Varying α:** As α increases (sparser attention), SC remains stable initially but eventually drops as temporal connections become too sparse. Conversely, IQ improves continuously and then plateaus, as sparsity effectively reduces noise and attention blurring.

Table 2. **Ablation study of impact of attention sink**. Attention could significantly improve overall video consistency and video quality.

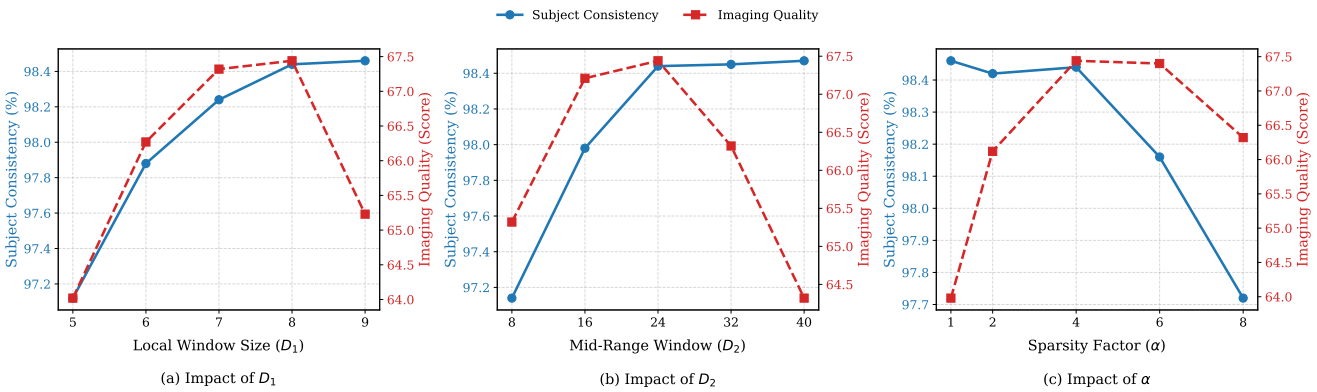
Method	SC (↑)	BC (↑)	MS (↑)	IQ (↑)	AQ (↑)	DD (↑)
Mid Frame	98.14	97.65	98.91	67.12	59.83	34.65
Last Frame	98.19	97.68	98.87	67.48	60.98	35.72
w/o Attention Sink	98.05	97.67	98.92	**67.53**	60.92	34.59
Our First Frame	**98.44**	**97.78**	**98.97**	67.44	**61.21**	**36.27**

5.2. Impact of Attention Sink

Attention sink is not unique to LLMs and prior analyses of video DiTs' attention maps also show sink behavior that serves as a global anchor to prevent identity drift instead of introducing a static bias. We run simple ablations with mid-frame sink, last-frame sink and without attention sink in Table 2, and first-frame sink performs best. For videos with large scene changes, such sink may be suboptimal, but our method can extend to adaptive sink to support this case and we will improve it in the future.

5.3. More Comparison with Other RoPE Scaling Methods

We compare VRPR with other RoPE scaling techniques e.g., LI (Linear Interpolation), NTK-aware scaling and YaRN in Table 3. Results confirm that standard RoPE scaling methods degrade visual details by diluting local positional frequencies, proving VRPR's hierarchical preservation of local precision is essential for high-quality generation.



(a) Impact of D_1 (b) Impact of D_2 (c) Impact of α

Figure 5. **Impact of D_1, D_2 and α on Subject Consistency and Imaging Quality**.

Table 3. **VRPR vs. other RoPE scaling methods.**

Method	SC (↑)	BC (↑)	MS (↑)	IQ (↑)	AQ (↑)	DD (↑)
LI	97.93	97.62	98.64	51.11	50.91	16.32
YARN	98.16	97.64	98.89	62.33	57.98	30.87
NTK-aware	97.89	97.59	98.82	56.47	53.54	24.21
Our VRPR	**98.44**	**97.78**	**98.97**	**68.84**	**61.21**	**36.27**

5.4. Different Granularity of VRPR

We validated the necessity of the 3-stage VRPR design (Fine → Medium → Coarse) against a 2-stage baseline which (Fine → Coarse) and a 4-stage baseline (Fine → Medium 1 → Medium 2 → Coarse). All these designs re-encode frame-level relative position within pretrained range. The result is shown in Table 4. The obtained results demonstrate that employing a multi-granularity (three granularities) approach, as opposed to a strategy with only two granularities, leads to improved video consistency and quality. However, diminishing returns are observed when the granularity continues to increase (e.g., four levels of granularity).

Table 4. **Ablation study of different granularity of VRPR.**

Method	SC (↑)	BC (↑)	MS (↑)	IQ (↑)	AQ (↑)	DD (↑)
2-stage	98.15	97.69	98.90	65.33	58.91	33.42
4-stage	**98.49**	97.93	98.95	**67.91**	60.90	35.97
Our 3-stage	98.44	**97.78**	**98.97**	67.44	**61.21**	**36.27**

5.5. More Layer-wise Strategy

To further verify the correctness and effectiveness of our layer-wise adaptive strategy, we conducted more experiments on different layer-wise strategy. (1) "Reverse" layer-wise strategy, we inverted the assignment logic: applying TSA only to layers identified as *not* sensitive to context length, and applying VRPR+TSA only to layers *not* sensitive to position. (2) Half Implementation 1: applying VRPR+TSA to the first half of the layers and VRPR to the second half.(3) Half Part Implementation 2: applying VRPR to the first half of the layers and VRPR+TSA to the second half. (4) Interleaved: alternately applying VRPR and VRPR+TSA strategies layer by layer.

As shown in Table 5, the Reverse configuration, Half Implementation and Interleaved configuration lead to a significant drop in both Subject Consistency (SC) and Imaging Quality (IQ), validating that our probing mechanism accurately identifies the specific needs of each layer.

Table 5. **Ablation study of more different layer-wise strategy.**

Method	SC (↑)	BC (↑)	MS (↑)	IQ (↑)	AQ (↑)	DD (↑)
Reverse	98.02	97.28	98.87	62.12	57.12	31.23
Half Implementation 1	98.18	97.41	98.91	64.21	59.12	32.19
Half Implementation 2	98.22	97.48	98.94	62.39	60.91	32.92
Interleaved	98.12	97.32	98.89	63.20	59.92	33.59
FreeLOC	**98.44**	**97.78**	**98.97**	**67.44**	**61.21**	**36.27**

6. More Experiment Results

6.1. Inference Efficiency Analysis

In this section, we provide a comprehensive comparison of efficiency, focusing on both inference time (measured as the time required for each denoising step) and peak GPU memory usage. We compare our method against other training-free methods, along with the *Direct Sampling* and *Sliding Window* baselines. For this comparison, all methods were applied to the Wan2.1-T2V-1.3B [5] model to generate 321-frame videos. All measurements were conducted on an NVIDIA A100 GPU. As presented in Table 6, while dramatically improving the quality of long videos generated by short video models, our method introduces no significant increase in inference time or peak memory consumption. This efficiency is primarily attributed to the optimized computations resulting from the Tiered Sparse Attention (TSA) mechanism, which effectively manages the computational cost and memory footprint associated with long-context attention.

6.2. One-Time Cost and Negligible Compute

Our probing is a one-time architectural characterization: the layer-wise sensitivity profile is consistent across random

Table 6. **Comparison of inference time and peak GPU memory usage** for generating 321-frame videos on the Wan2.1-T2V-1.3B model. All measurements were conducted on an NVIDIA A100 GPU.

Method	Inference Time (s/step)	Mem (GB)
Direct Sampling	33.93	29.34
Sliding Window	16.72	30.33
FreeNoise [4]	17.21	30.34
Freelong [3]	48.15	40.71
RIFLEX [6]	33.94	29.34
FreeLOC	24.35	29.87

seeds, prompts and aspect ratios, and is fixed for that model once obtained. Although the main paper reports generating $M \times N$ videos for completeness, we empirically find that the profile is highly stable and does not require the full probing set. Specifically on Wan2.1-T2V-1.3B [5], using a single prompt to derive the profile yields high Spearman rank correlation (evaluated over 10 randomly sampled single prompts) of ($\rho = 0.984 \pm 0.009$) compared to the full profile. Thus, a single sample set is therefore sufficient in practice, rendering the probing cost negligible. For Wan2.1-T2V-1.3B, probing requires only 4 hours on an RTX 4090. Moreover, *we will release pre-computed sensitivity profiles for popular models*, eliminating the need for users *to rerun probing*.

6.3. User Study

To further assess our results based on human subjective judgment, we carried out a user study. In this study, participants were shown long videos generated using Wan2.1-T2V-1.3B as the base short video model. Videos from all compared methods were included, totaling 50 videos. The examples were presented to participants in a random order to eliminate potential bias.

Participants were asked to score the generated videos on a scale of 1 to 5 according to three evaluation criteria: (1) **Content Consistency**, (2) **Video Quality**, and (3) **Video-Text Alignment**. The average scores for each method are reported in Table 7. As shown, our method received the highest ratings across all three metrics.

7. More Qualitative Results

We provide extensive qualitative comparisons to visually demonstrate the superiority of FreeLOC. Figures 6, 7, 8 and 9 show additional qualitative results on Wan2.1-T2V-1.3B and Hunyuan Video. Our method consistently produces videos with higher temporal fidelity, sharper details, and more stable object identity compared to all baselines.

Table 7. **User study results**. Participants scored videos on a scale of 1 (worst) to 5 (best) across three criteria. Our method received the highest ratings across all three metrics.

Method	Content Consistency	Video Quality	Video-Text Alignment
Direct Sampling	3.27	1.02	1.92
Sliding Window	1.17	3.25	2.34
FreeNoise [4]	1.20	2.88	2.68
Freelong [3]	3.59	2.11	2.98
RIFLEX [6]	3.31	1.09	1.89
FreeLOC	**4.11**	**3.95**	**3.78**

References

[1] 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 Computer Vision and Pattern Recognition Conference*, pages 21807–21818, 2024. 4

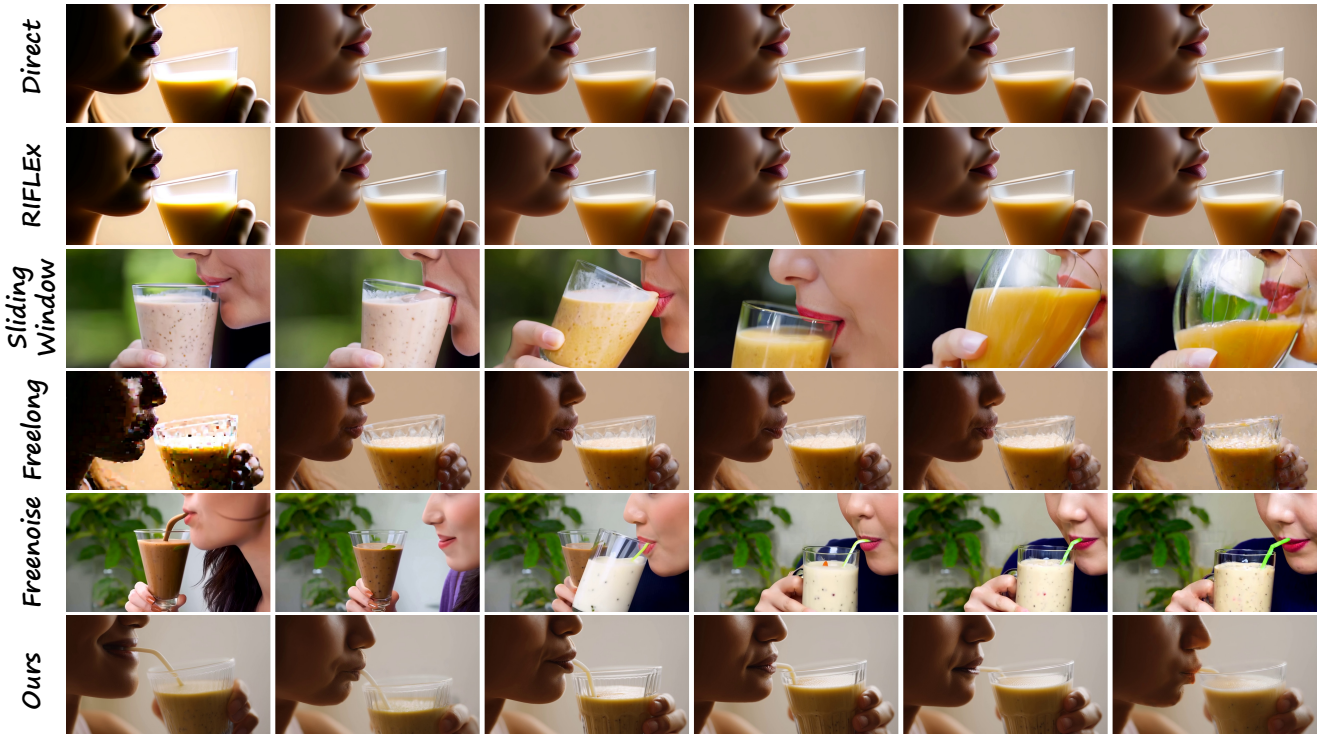
[2] Hongye Jin, Xiaotian Han, Jingfeng Yang, Zhimeng Jiang, Zirui Liu, Chia-Yuan Chang, Huiyuan Chen, and Xia Hu. Llm maybe longlm: Self-extend llm context window without tuning. *arXiv preprint arXiv:2401.01325*, 2024. 2

[3] Yu Lu, Yuanzhi Liang, Linchao Zhu, and Yi Yang. Freelong: Training-free long video generation with spectralblend temporal attention. *NeurIPS*, 37:131434–131455, 2024. 4, 7

[4] 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. 4, 7

[5] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 4, 6, 7

[6] Min Zhao, Guande He, Yixiao Chen, Hongzhou Zhu, Chongxuan Li, and Jun Zhu. Riflex: A free lunch for length extrapolation in video diffusion transformers. *arXiv preprint arXiv:2502.15894*, 2025. 4, 7



Prompt: A person is drinking a smoothie from a glass

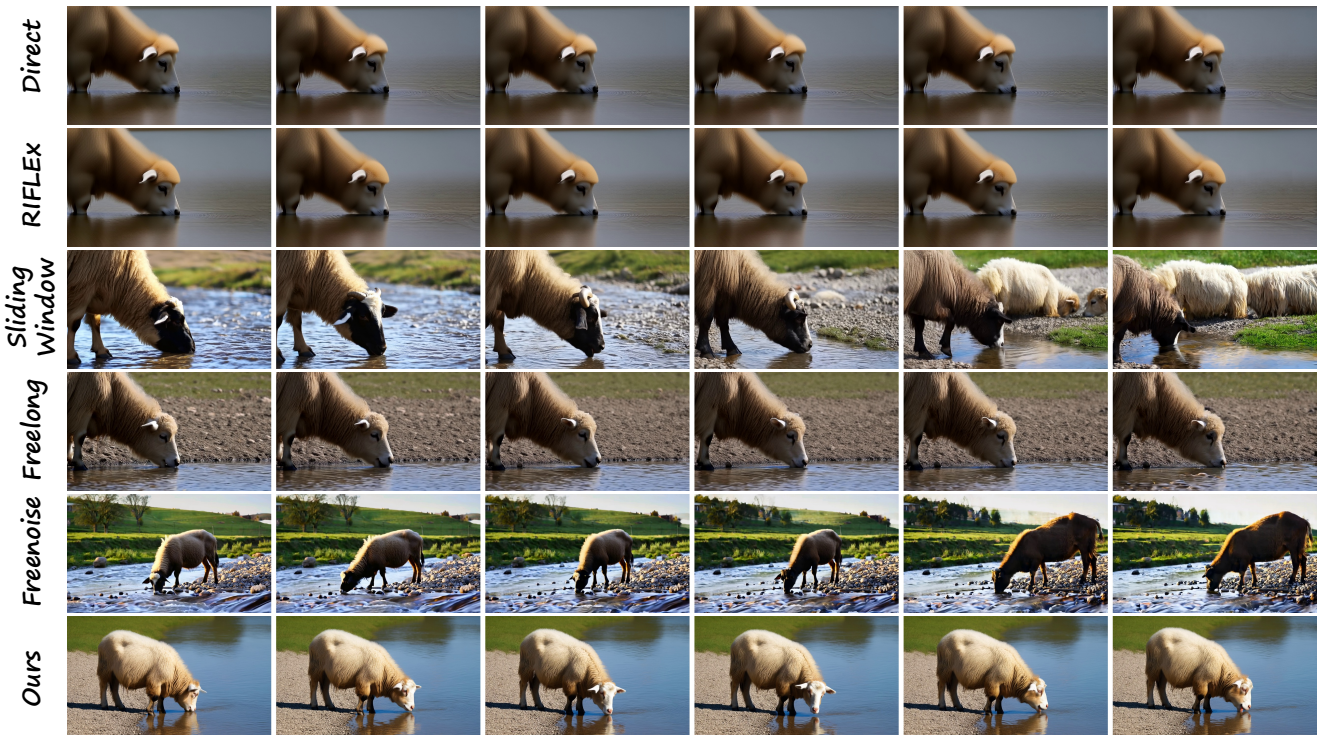
Prompt: In a dynamic scene, the camera gracefully orbits around a vibrant basketball court, capturing the essence of the game. The court is alive with energy, featuring players in motion, their jerseys a blur of color against the polished wooden floor. As the camera circles, it reveals the intricate details of the game: the focused expressions of the players, the swift movement of the ball, and the rhythmic bounce echoing in the air. The camera's journey offers a 360-degree view, showcasing the audience's anticipation, the scoreboard's glow, and the towering hoop standing as the centerpiece of this thrilling spectacle.

Figure 6. **Qualitative Results of Wan2.1-1.3B-T2V with $4\times$ extension (321-frame)**

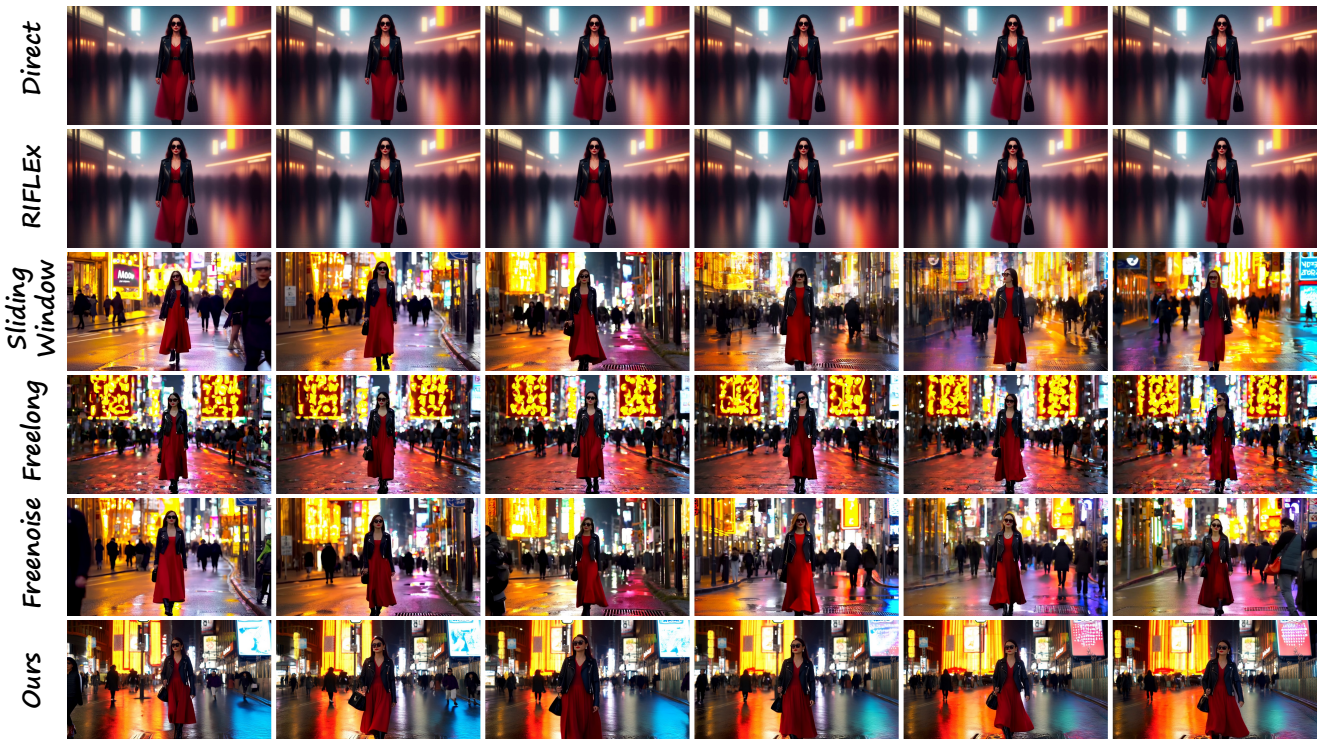
Prompt: Origami dancers in white paper, 3D render, on white background, studio shot, dancing modern dance.

Prompt: In a serene, sunlit meadow, the camera gracefully orbits around a solitary, ancient oak tree, its gnarled branches reaching skyward, casting intricate shadows on the lush grass below. As the camera circles, the scene reveals a vibrant tapestry of wildflowers in full bloom, their colors vivid against the verdant backdrop. The gentle rustling of leaves accompanies the camera's smooth motion, capturing the tranquil ambiance of this secluded haven. As the camera continues its orbit, the distant mountains come into view, their majestic peaks bathed in the warm glow of the setting sun, completing the picturesque panorama.

Figure 7. **Qualitative Results of Wan2.1-1.3B-T2V with $2\times$ extension (161-frame)**



Prompt: a sheep bending down to drink water from a river



Prompt: A stylish woman walks down a Tokyo street filled with warm glowing neon and animated city signage. She wears a black leather jacket, a long red dress, and black boots, and carries a black purse. She wears sunglasses and red lipstick. She walks confidently and casually. The street is damp and reflective, creating a mirror effect of the colorful lights. Many pedestrians walk about.

Figure 8. **Qualitative Results of HunyuanVideo with $4\times$ extension (509-frame)**

Figure 9. **Qualitative Results of HunyuanVideo with $2\times$ extension (253-frame)**