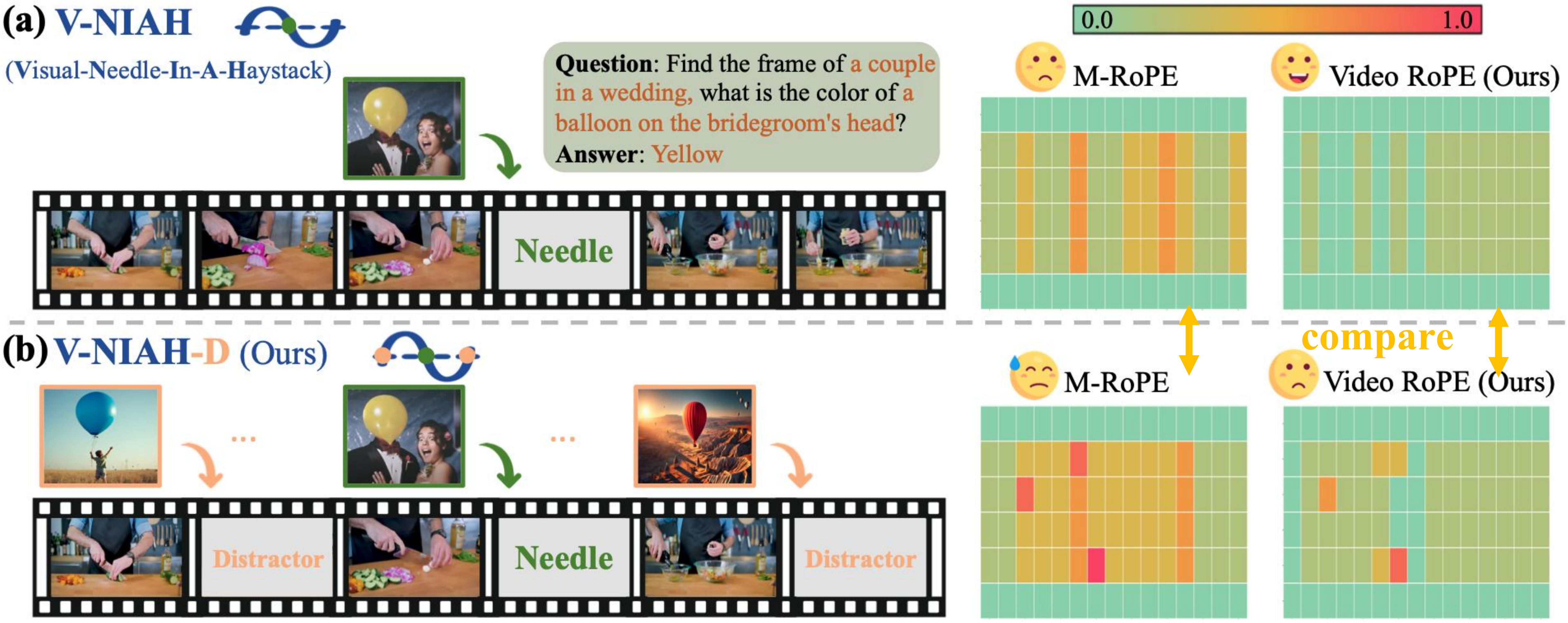
VideoRoPE

What Makes for Good Video Rotary Position Embedding?

Xilin Wei*, Xiaoran Liu*, Yuhang Zang, Xiaoyi Dong,
Pan Zhang, Yuhang Cao, Jian Tong, Haodong Duan,
Qipeng Guo, Jiaqi Wang, Xipeng Qiu, Dahua Lin

Observation 1

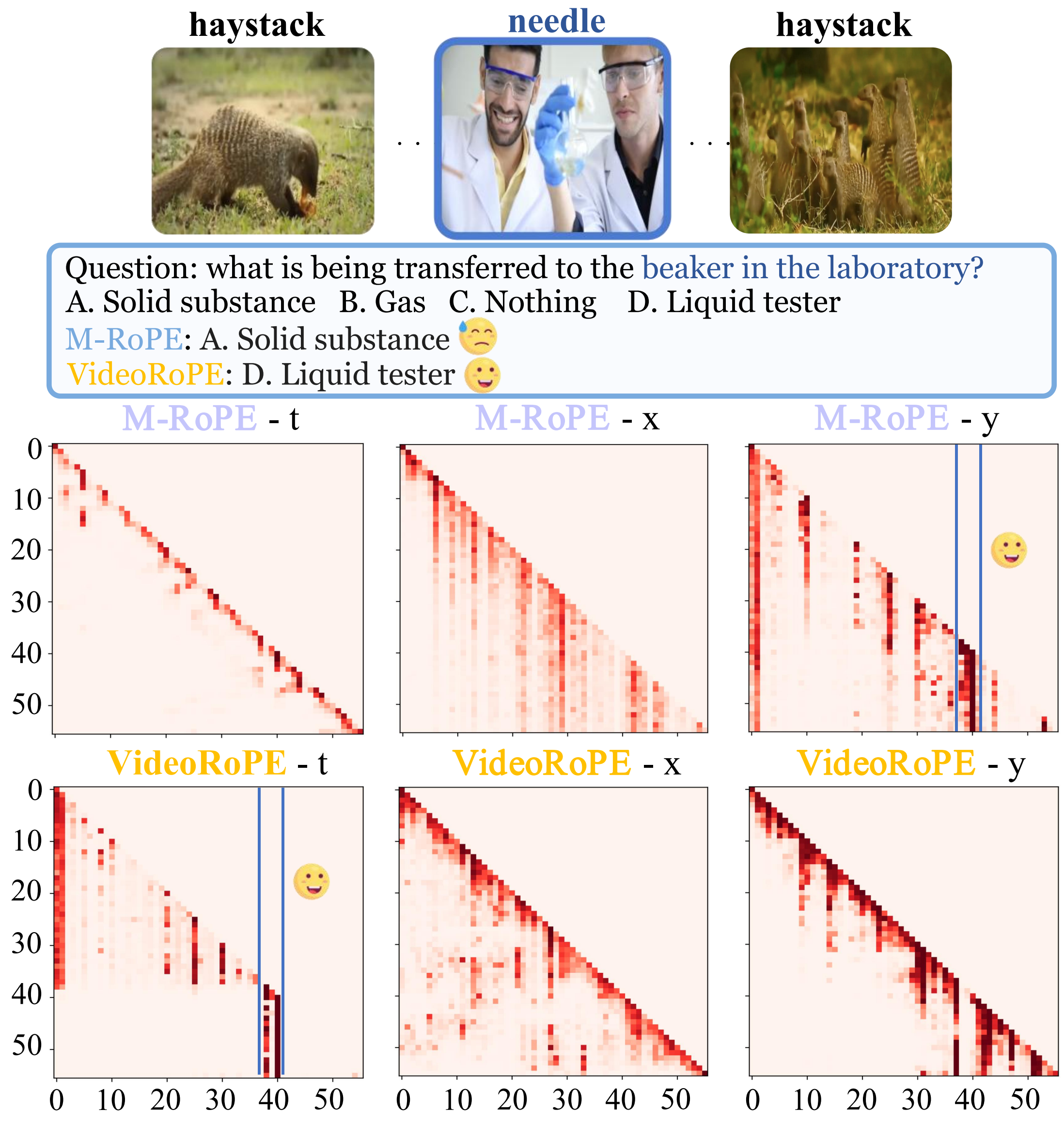
Interference Issues in Modern RoPE Variants



Current advanced position embedding methods, such as M-RoPE, are still susceptible to periodic distractor interference (V-NIAH-D).

Observation 2

Misfocused Attention Sink



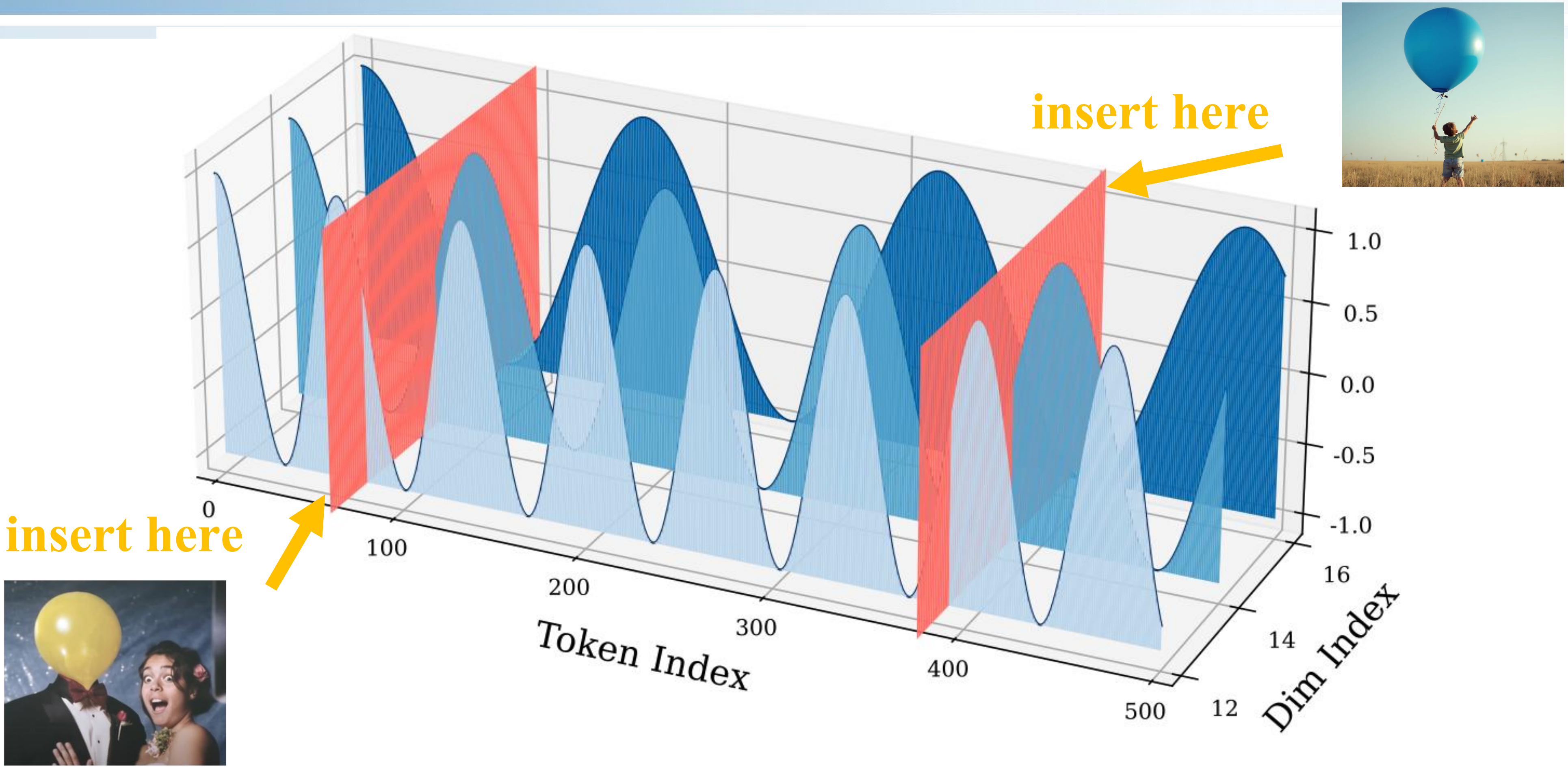
- Both M-RoPE and VideoRoPE **successfully locate the needle information** required to answer the question.
- Due to suboptimal frequency allocation, M-RoPE focuses on vertical cues at the expense of temporal semantics, leading to **poor long-range modeling and wrong answers.** VideoRoPE, by leveraging **temporal localization, answers correctly**.

Four Key Properties

- most previous methods only cover **part of** the table
- **VideoRoPE achieves a full-stack design,** addressing **all four core dimensions**: structural modeling, frequency allocation, spatial symmetry, and temporal scaling—surpassing prior RoPE variants.

	2D/3D Structure	Frequency Allocation	Spatial Symmetry	Temporal Index Scaling
Vanilla RoPE (Su et al., 2024)	✗	✗	✗	✗
TAD-RoPE (Gao et al., 2024)	✗	✗	✗	✓
RoPE-Tie (Su, 2024a)	✓	✗	✓	✗
M-RoPE (Wang et al., 2024a)	✓	✗	✗	✗
VideoRoPE (Ours)	✓	✓	✓	✓

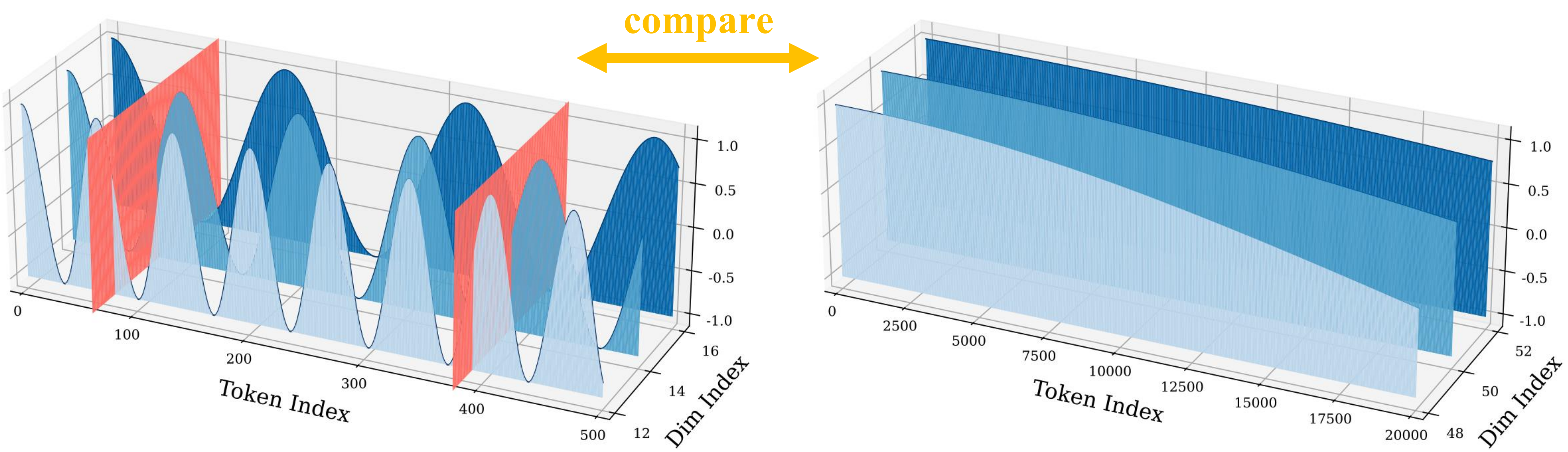
Low-frequency Temporal Allocation (LTA)



(a) Temporal Frequency Allocation in M-RoPE

As shown in the red planes, positions that are far apart in time can end up with **similar** positional encodings due to these **oscillations**.

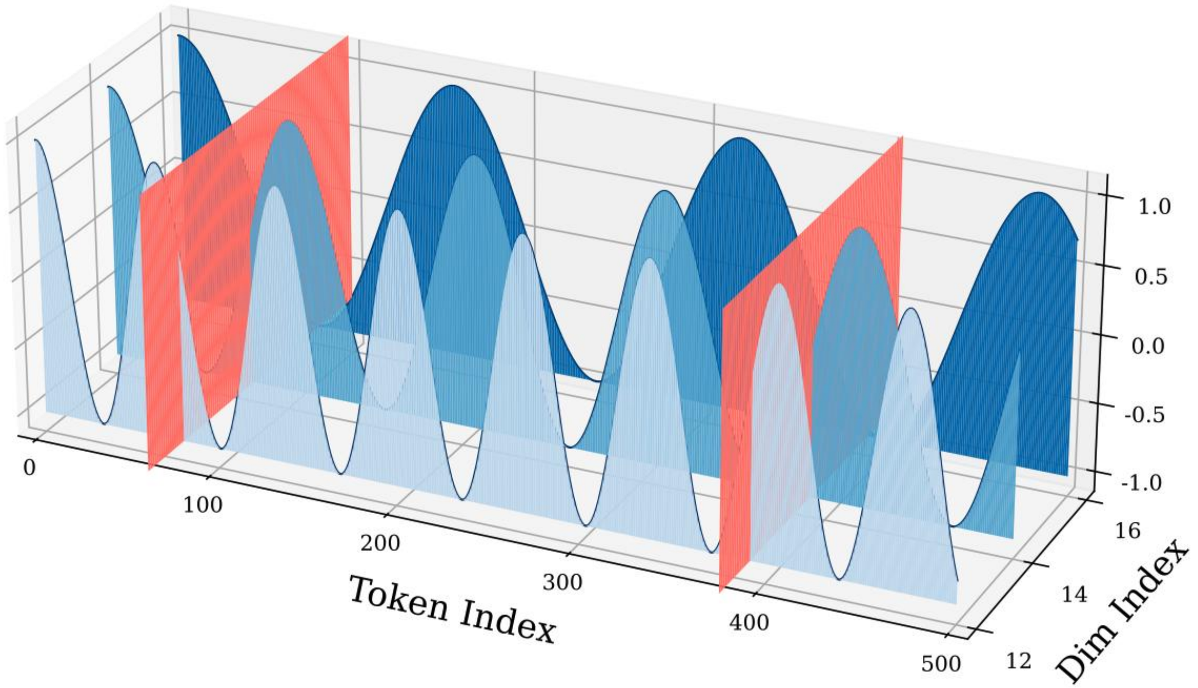
Low-frequency Temporal Allocation (LTA)



(a) Temporal Frequency Allocation in M-RoPE (b) Temporal Frequency Allocation in VideoRoPE (ours)

VideoRoPE adopts **low-frequency modeling for the temporal dimension**, achieving better long-range monotonicity and **avoiding oscillations**, which effectively reduces distractor interference in V-NIAH-D.

Low-frequency Temporal Allocation (LTA)



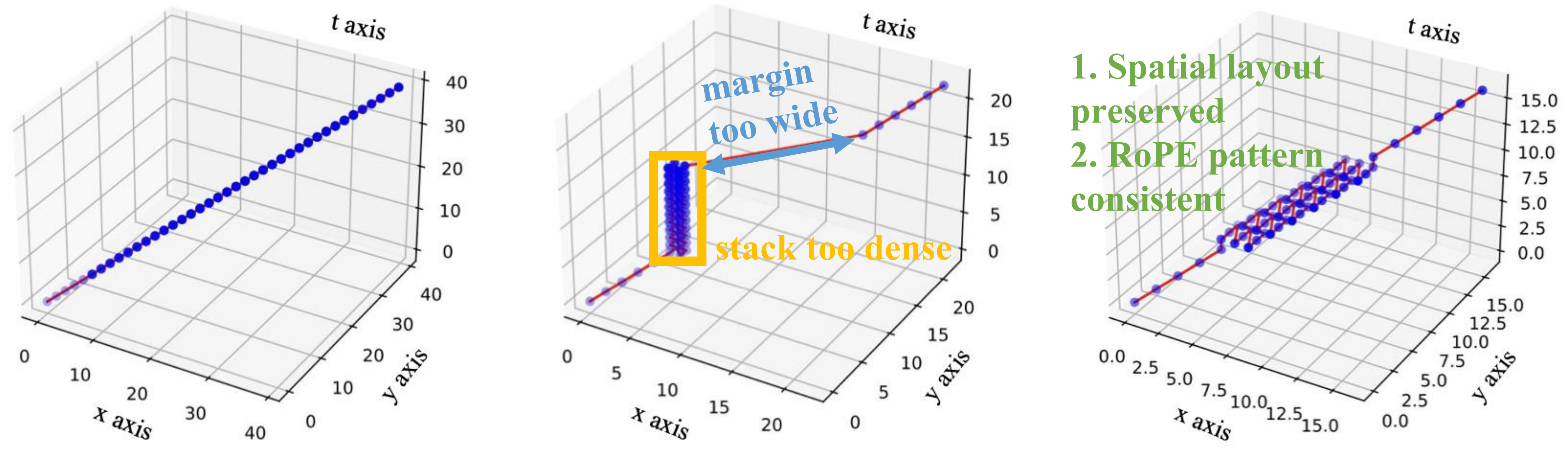
(a) Temporal Frequency Allocation in M-RoPE

$$
\underbrace{\begin{pmatrix} q^{(96)} \\ q^{(97)} \\ q^{(98)} \\ q^{(99)} \\ \vdots \\ q^{(126)} \\ q^{(127)} \end{pmatrix}^{\top}
\begin{pmatrix}
\cos\theta_{48}\Delta t & -\sin\theta_{48}\Delta t & 0 & 0 & \cdots & 0 & 0 \\
\sin\theta_{48}\Delta t & \cos\theta_{48}\Delta t & 0 & 0 & \cdots & 0 & 0 \\
0 & 0 & \cos\theta_{49}\Delta t & -\sin\theta_{49}\Delta t & \cdots & 0 & 0 \\
0 & 0 & \sin\theta_{49}\Delta t & \cos\theta_{49}\Delta t & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & 0 & \cdots & \cos\theta_{63}\Delta t & -\sin\theta_{63}\Delta t \\
0 & 0 & 0 & 0 & \cdots & \sin\theta_{63}\Delta t & \cos\theta_{63}\Delta t
\end{pmatrix}
\begin{pmatrix} k^{(96)} \\ k^{(97)} \\ k^{(98)} \\ k^{(99)} \\ \vdots \\ k^{(126)} \\ k^{(127)} \end{pmatrix}}_{\text{modeling temporal dependency with lower frequency}}
$$

$$
+\underbrace{\begin{pmatrix} q^{(0)} \\ q^{(1)} \\ q^{(4)} \\ q^{(5)} \\ \vdots \\ q^{(92)} \\ q^{(93)} \end{pmatrix}^{\top}
\begin{pmatrix}
\cos\theta_{0}\Delta x & -\sin\theta_{0}\Delta x & 0 & 0 & \cdots & 0 & 0 \\
\sin\theta_{0}\Delta x & \cos\theta_{0}\Delta x & 0 & 0 & \cdots & 0 & 0 \\
0 & 0 & \cos\theta_{2}\Delta x & -\sin\theta_{2}\Delta x & \cdots & 0 & 0 \\
0 & 0 & \sin\theta_{2}\Delta x & \cos\theta_{2}\Delta x & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & 0 & \cdots & \cos\theta_{46}\Delta x & -\sin\theta_{46}\Delta x \\
0 & 0 & 0 & 0 & \cdots & \sin\theta_{46}\Delta x & \cos\theta_{46}\Delta x
\end{pmatrix}
\begin{pmatrix} k^{(0)} \\ k^{(1)} \\ k^{(4)} \\ k^{(5)} \\ \vdots \\ k^{(92)} \\ k^{(93)} \end{pmatrix}}_{\text{modeling horizontal dependency with interleaved high frequency}}
$$

$$
+\underbrace{\begin{pmatrix} q^{(2)} \\ q^{(3)} \\ q^{(6)} \\ q^{(7)} \\ \vdots \\ q^{(94)} \\ q^{(95)} \end{pmatrix}^{\top}
\begin{pmatrix}
\cos\theta_{1}\Delta y & -\sin\theta_{1}\Delta y & 0 & 0 & \cdots & 0 & 0 \\
\sin\theta_{1}\Delta y & \cos\theta_{1}\Delta y & 0 & 0 & \cdots & 0 & 0 \\
0 & 0 & \cos\theta_{3}\Delta y & -\sin\theta_{3}\Delta y & \cdots & 0 & 0 \\
0 & 0 & \sin\theta_{3}\Delta y & \cos\theta_{3}\Delta y & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & 0 & \cdots & \cos\theta_{47}\Delta y & -\sin\theta_{47}\Delta y \\
0 & 0 & 0 & 0 & \cdots & \sin\theta_{47}\Delta y & \cos\theta_{47}\Delta y
\end{pmatrix}
\begin{pmatrix} k^{(2)} \\ k^{(3)} \\ k^{(6)} \\ k^{(7)} \\ \vdots \\ k^{(94)} \\ k^{(95)} \end{pmatrix}}_{\text{modeling vertical dependency with interleaved high frequency}}
$$

Diagonal Layout (DL)



(a) 3D visualization for Vanilla RoPE. (b) 3D visualization for M-RoPE. (c) 3D visualization for VideoRoPE.

3D visualization of position embeddings: (a) Vanilla RoPE **lacks spatial modeling**. (b) M-RoPE introduces **inconsistent index growth** across frames. (c) VideoRoPE **balances spatial modeling with consistent indexing**, preserving RoPE’s desirable structure.

Adjustable Temporal Spacing (ATS)

$$
(t, x, y) = \begin{cases} (\tau, \tau, \tau) & \text{if } 0 \leq \tau < T_s \\ \begin{pmatrix} T_s + \delta(\tau - T_s), \\ T_s + \delta(\tau - T_s) + w - \frac{W}{2}, \\ T_s + \delta(\tau - T_s) + h - \frac{H}{2} \end{pmatrix} & \text{if } T_s \leq \tau < T_s + T_v \\ \begin{pmatrix} \tau + (\delta - 1)T_v, \\ \tau + (\delta - 1)T_v, \\ \tau + (\delta - 1)T_v \end{pmatrix} & \text{if } T_s + T_v \leq \tau < T_s + T_v + T_e \end{cases}
$$

Adjustable Temporal Spacing(ATS). To scale the temporal index, we introduce a scaling factor δ to better align temporal information between visual and textual tokens.

Adjustable Temporal Spacing (ATS)

$$(t,x,y)=\begin{cases}(\tau,\tau,\tau) & \text{if } 0\le\tau<T_s\\ \begin{pmatrix}T_s+\delta(\tau-T_s),\\ T_s+\delta(\tau-T_s)+w-\frac{W}{2},\\ T_s+\delta(\tau-T_s)+h-\frac{H}{2}\end{pmatrix} & \text{if } T_s\le\tau<T_s+T_v\\ \begin{pmatrix}\tau+(\delta-1)T_v,\\ \tau+(\delta-1)T_v,\\ \tau+(\delta-1)T_v\end{pmatrix} & \text{if } T_s+T_v\le\tau<T_s+T_v+T_e\end{cases}$$

Adjustable Temporal Spacing(ATS). To scale the temporal index, we introduce a scaling factor δ to better align temporal information between visual and textual tokens.

Adjustable Temporal Spacing (ATS)

$$
(t, x, y) = \begin{cases} (\tau, \tau, \tau) & \text{if } 0 \leq \tau < T_s \\ \begin{pmatrix} T_s + \delta(\tau - T_s), \\ T_s + \delta(\tau - T_s) + w - \frac{W}{2}, \\ T_s + \delta(\tau - T_s) + h - \frac{H}{2} \end{pmatrix} & \text{if } T_s \leq \tau < T_s + T_v \\ \begin{pmatrix} \tau + (\delta - 1)T_v, \\ \tau + (\delta - 1)T_v, \\ \tau + (\delta - 1)T_v \end{pmatrix} & \text{if } T_s + T_v \leq \tau < T_s + T_v + T_e \end{cases}
$$

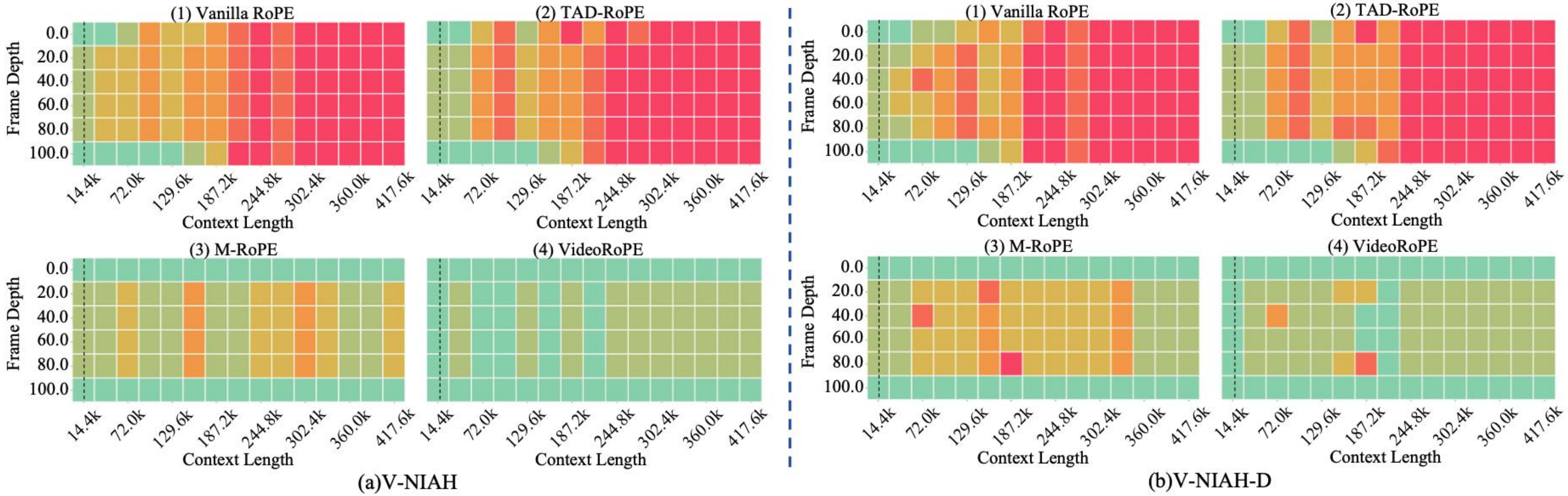
Adjustable Temporal Spacing(ATS). To scale the temporal index, we introduce a scaling factor δ to better align temporal information between visual and textual tokens.

Experiments on Long Video Understanding

Method	LongVideoBench				MLVU				Video-MME			
	8k	16k	32k	64k	8k	16k	32k	64k	8k	16k	32k	64k
Vanilla RoPE (Su et al., 2024)	**54.97**	54.87	54.56	54.04	63.31	65.79	65.93	62.02	60.67	60.00	61.33	58.33
TAD-RoPE (Gao et al., 2024)	54.14	55.08	53.94	53.42	63.67	65.28	65.28	60.73	60.33	**61.33**	**62.00**	58.67
M-RoPE (Wang et al., 2024a)	53.42	52.80	53.11	54.35	60.41	60.68	61.56	61.10	60.67	59.67	61.00	59.67
VideoRoPE (Ours)	54.46	**55.29**	**57.15**	**57.26**	**65.19**	**66.29**	**66.02**	**65.56**	**61.33**	61.00	61.67	**61.33**

- Benchmarks: LongVideoBench, MLVU, VideoMME
- Consistent gains over M-RoPE: +2.91 / +4.46 / +1.66 @64k context
- Robust to long-range dependencies
- Strong adaptability across tasks

Experiments on Long Video Retrieval



Method	V-NIAH Acc.	V-NIAH-D Acc.
Vanilla RoPE (Su et al., 2024)	31.78	30.22
TAD-RoPE (Gao et al., 2024)	29.33	29.56
M-RoPE (Wang et al., 2024a)	78.67	74.67
VideoRoPE	**91.11**	**87.11**

- V-NIAH-D is more challenging than V-NIAH
- Vanilla RoPE / TAD-RoPE: limited extrapolation
- VideoRoPE > M-RoPE in long context extrapolation
- +12.44% over M-RoPE on Video Retrieval

Experiments on Video Hallucination

Method	OR	T	SD	F	NF	Avg.
Vanilla RoPE (Su et al., 2024)	51.5	30.0	48.0	8.0	43.0	36.1
TAD-RoPE (Gao et al., 2024)	51.0	37.0	48.0	11.5	47.5	39.0
M-RoPE (Wang et al., 2024a)	39.0	29.0	43.5	12.5	47.5	34.3
VideoRoPE	**57.0**	**58.5**	**50.5**	**15.0**	**50.0**	**46.2**

- +29.5% on Temporal Hallucination → better temporal reasoning
- +18.0% on Spatial/Object-Relation Hallucination → better spatial understanding
- Robust to complex video hallucinations

Experiments on Ablation Studues

Method	LongVideoBench				MLVU			
	8k	16k	32k	64k	8k	16k	32k	64k
Baseline	53.42	52.80	53.11	54.35	60.41	60.68	61.56	61.10
+ DL	52.17	52.07	53.31	53.63	62.06	63.03	62.52	62.75
+ DL & LTA	**54.46**	**55.49**	54.66	55.60	63.35	64.09	64.00	63.26
+ DL & LTA & ATS	**54.46**	55.29	**57.15**	**57.26**	**65.19**	**66.29**	**66.02**	**65.56**

- Ablation on LongVideoBench & MLVU (64k context)
- Baseline (M-RoPE): 54.35 / 61.10
- +DL → +LTA → +ATS → performance improves progressively
- Final: 57.26 / 65.56
- Effective use of spatio-temporal position encoding

* Further ablations on layout strategies, frequency allocation, ATS scaling, and DL are provided in the main paper.

Conclusion

- **Four key criteria** for effective positional encoding:
  - 2D/3D structure, frequency allocation, spatial symmetry, temporal index scaling

- Prior RoPE variants struggle with **temporal distractors due to improper allocation**
  - VideoRoPE **addresses** this with:
  - 3D spatiotemporal structure
  - Low-frequency temporal allocation (reduces oscillations)
  - Diagonal spatial layout (ensures symmetry)
  - Adjustable temporal spacing (ATS)

- **Superior performance** in:
  - Long video retrieval
  - Video understanding
  - Video hallucination tasks

| Thanks

Github

Paper

Homepage

wechat

X

Contact us: wiselnn570@gmail.com