Bringing RNNs Back to Efficient Open-Ended Video Understanding

Weili Xu[1,2] Enxin Song[1] Wenhao Chai[3†] Xuexiang Wen[1] Tian Ye[4] Gaoang Wang[1✉]

[1] Zhejiang University [2] University of Illinois Urbana-Champaign
[3] University of Washington [4] Hong Kong University of Science and Technology (GZ)

Abstract

Transformer-based long video understanding models are hindered by their high computational complexity and prohibitive memory cost, since the memory and computation scale quadratically with input sequence length. We propose LongVidRWKV *to address this challenge by replacing the LLM component in MLLMs with RWKV, an RNN-like language model that handles input sequences of arbitrary length with constant-size hidden states. To reduce the gap between RWKV's 4k context length and the extended token sequences typical of long videos, we combine visual token merge with linear RNN models and reorder merged visual tokens. Despite having only 2B parameters and being trained exclusively on public data,* LongVidRWKV *achieves performance comparable to Transformer-based models of similar size trained on private datasets across multiple video benchmarks. This demonstrates the potential of efficient, linear RNNs to lower the computation entry barrier for long video understanding. To our knowledge, we are the first to use an RWKV LLM backbone in a LLaVA-like model for open-ended video understanding.*

1. Introduction

Large multimodal models (LMMs) [2, 35, 41, 49–54] have demonstrated strong abilities such as captioning and visual question answering. Among them, video-based LMMs often follow an architecture similar to LLaVA [26, 28], which faces challenges when processing longer videos with complex temporal dynamics. As more frames are sampled, computation in LLaVA's visual extractor scales linearly, with tokens attending only within the same frame. However, the computation in LLaVA's LLM backbone scales quadratically with the number of input frames due to causal self-attention mechanism, where each token attends to all previous tokens. To develop more efficient LMMs, prior works propose various token reduction strategies [1, 15, 17, 30, 38]. For instance, Token Merge [3] (ToMe) is a training-free method that gradually combines visual tokens based on token similarity, and has proven effective in image and video understanding tasks [5, 36].

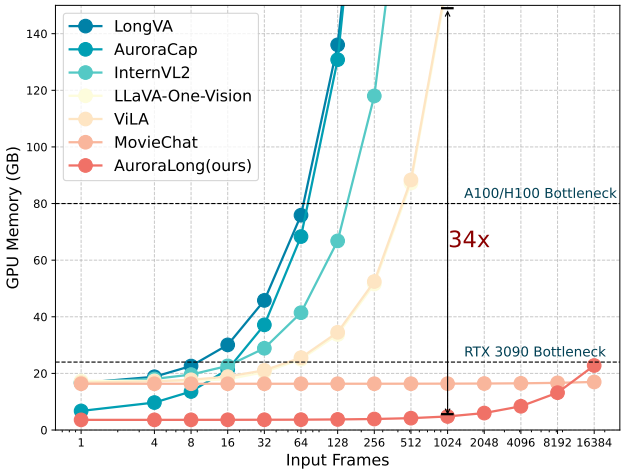

Figure 1. VRAM cost under gigabyte (GB) (y-axis) v.s. frame number (x-axis) comparison. While the previous method can only support around 100 frames of inference with A100 or H100, LongVidRWKV can handle videos with over 10 thousands frames on a 24GB GPU. LongVidRWKV has a 34× advantage over other methods in terms of VRAM cost when process 1,024 frames.

Currently, linear RNN models such as Mamba and RWKV [12, 14, 31, 32] utilize linear attention variants [18, 44] whose training memory cost scales linearly with input sequence length, making them ideal for the LLM backbone in LMMs. Previous works [6, 13, 23, 24, 29, 56] propose linear attention based visual encoders for tasks like image classification and action recognition. VisualRWKV [16] and VL-Mamba [33] use RWKV and Mamba respectively as the LLM backbone for image-based LMMs. Vamba [34] distills Qwen2-VL [41] into a hybrid video LMM. However, no previous works explore using RWKV as an efficient LLM backbone for open-ended video QA.

In this paper, we present LongVidRWKV to incorporate more input frames within a limited 4k context length of the pretrained RWKV LLM, combining the simple yet efficient token merging strategy with linear RNN models by reordering the merged visual tokens, which is empirically proven to be beneficial to various video understanding tasks. Our main contributions are summarized as follows:

† Project lead, ✉ Corresponding author

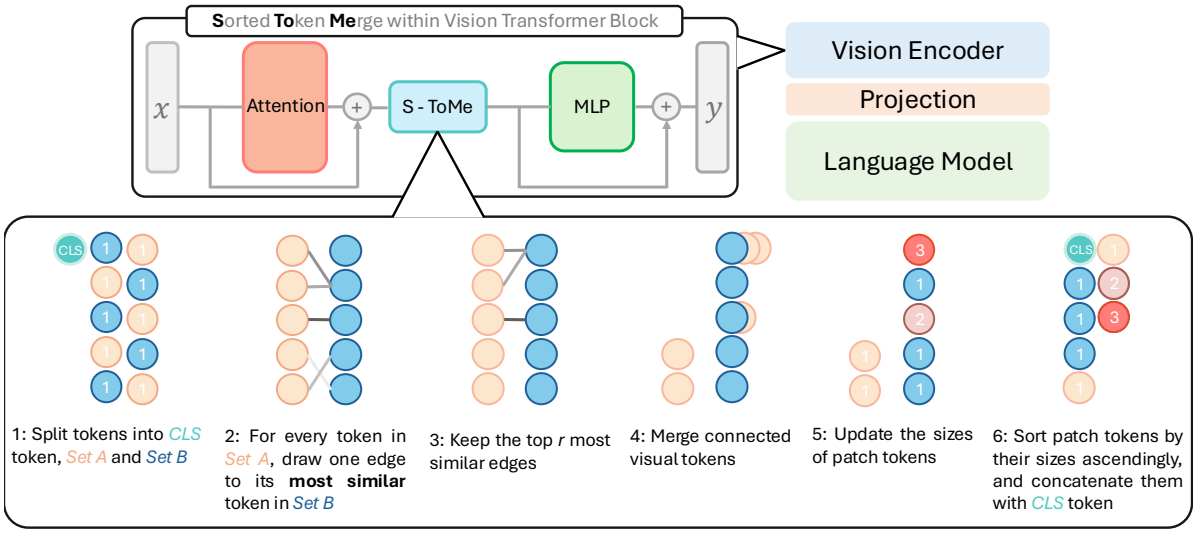

Figure 2. Visualization of the Sorted Token Merge (S-ToMe) algorithm used in LongVidRWKV's visual encoder.

- We employ a fully recurrent LLM backbone in a LLaVA-like model architecture for open-ended video QA, presenting a novel hybrid architecture that can handle long video inputs with lower memory requirements.
- We propose a training-free **S**orted visual **To**ken **Me**rge (S-ToMe) strategy to bridge the gap between long video token inputs and RWKV's limited pretrained 4k context length while retaining visual information.
- Despite only being trained on public data, our model performs favorably against several state-of-the-art larger LMMs across video understanding tasks, while reducing computational complexity and memory consumption.

2. Method

2.1. Network Architecture

We inherit the LMM architecture of LLaVA-1.5 [27], with SigLIP [46] as the vision encoder and a simple two-layer MLP as the cross-modal connector. We do not consider linear-attention based visual extractors since the computation overhead and memory requirement in vision transformers is already $O(N)$ with respect to number of sampled frames N. We use RWKV-v6 [32] as the LLM backbone. However, [8] and [48] show that linear RNNs fail to extrapolate beyond pretrained context length. RWKV lacks context extension techniques, necessitating the introduction of token merging [3] to reduce the number of visual tokens.

2.2. Sorted Visual Token Merge

Although RWKV's [32] is memory-efficient in handling arbitrary-length input, [8] shows that linear attention models overfit to their pretrained context. Since RWKV [32] is pretrained with a 4,096-token context, we introduce Token Merging [3] to merge similar visual tokens per frame,

Algorithm 1 Sorted Visual Token Merge

Require: Input visual tokens per frame $\mathcal{X}$
Require: Vision Transformer $\mathcal{V}$ with $\mathcal{N}$ layers
Require: Token Merging threshold r
for n in $\mathcal{V}[:\mathcal{N}-1]$ **do**
  # $\mathcal{X} \in [batch, tokens, channels]$
  $\mathcal{X} \leftarrow \text{Attention}_n(\mathcal{X})$
  # Split CLS tokens and patch tokens
  $CLS, \mathcal{X} \leftarrow \mathcal{X}[:,0,:], \mathcal{X}[:,1:,:]$
  # Assign patch tokens to $Set\ \mathcal{A}, Set\ \mathcal{B}$
  $\mathcal{A}, \mathcal{B} \leftarrow \mathcal{X}[:,::2,:], \mathcal{X}[:,1::2,:]$
  $Scores \leftarrow \text{similarity}(\mathcal{A}, \mathcal{B})$
  # Get merged tokens and unmerged tokens
  $src, unm \leftarrow \text{top}(\mathcal{X}, Scores, r)$
  $dst \leftarrow \text{merge}(src)$
  # Update patch count s for each token
  $\text{update}(dst.s)$
  # Sort tokens by s
  $\mathcal{X} \leftarrow \text{sort}(dst, unm)$
  $\mathcal{X} \leftarrow \text{concat}(CLS, \mathcal{X})$
  $\mathcal{X} \leftarrow \text{MLP}(CLS, \mathcal{X})$
end for

bridging the gap between the pretrained context and lengthy visual sequences. To model visual input sequence order, Transformers utilize explicit positional embedding [19, 37, 39, 40], while RNNs model sequence order implicitly due to their recurrent nature. Prior works [16, 23, 29] enhance detailed visual modeling in linear attention models by bidirectionally scanning tokens, increasing computational cost. Instead, we propose a simpler, training-free visual token reordering strategy to better utilize the pretrained unidirectional textual modeling capabilities while retaining as much

Table 1. Results on short video understanding benchmarks [5, 42, 45]. The best result is highlighted in bold, and the second best is underlined. Results with * are evaluated in-house, while others are sourced from official leaderboards of each benchmark.

Models	Size	#Frame	VDC [5]						ANet [4]		VATEX [42]
			Avg.	Short	Camera	Background	Main Object	Detailed	Acc.	Score	BLEU@1
LongVA [47]	7B	64	34.50	31.94	35.32	36.39	40.95	27.91	-	2.8	65.2*
ShareGPT4Video [7]	8B	16	36.17	39.08	33.28	35.77	37.12	35.62	-	-	56.6*
LLAVA-OneVision [21]	7B	32	37.45	32.58	37.82	37.43	38.21	41.20	56.6	-	54.2*
AuroraCap [5]	7B	16	38.21	32.07	43.50	35.92	39.02	41.30	**61.8**	3.8	57.1
InternVL-2 [9]	8B	16	37.72	33.02	39.08	37.47	44.16	34.89	-	-	-
LongVidRWKV (ours)	2B	1fps	**42.54**	**38.89**	**43.70**	40.26	46.32	**43.54**	60.0	**4.2**	**68.5**

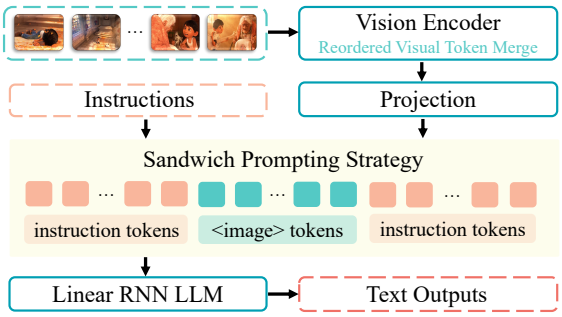

Figure 3. LongVidRWKV prompting strategy overview.

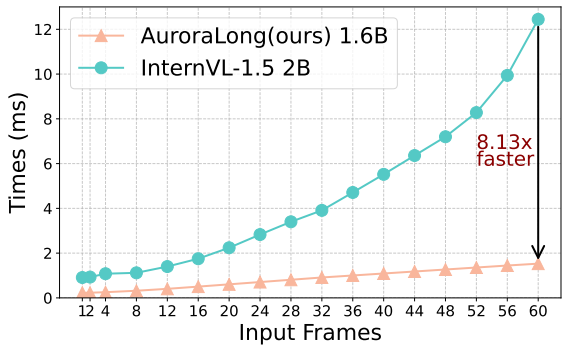

Figure 4. LongVidRWKV requires less computation and provides lower latency compared to transformer models of similar size.

spatial information as possible. As shown in Figure 2 and Algorithm 1, we reorder merged tokens by the number of visual patches in ascending order.

2.3. Prompting Strategy

RWKV [31] and other linear RNN language models have constant hidden states, limiting their instruction-following ability without careful prompting. To address this, we adopt the sandwich prompting strategy from VisualRWKV [16], inserting reordered merged visual tokens between duplicated instructional text tokens, as shown in Figure 3.

2.4. Training Recipe

Following [5], we further adopt a three-stage training strategy, which can be noted as Pretraining, Vision and Language stages. All data we use to train LongVidRWKV are publicly available. Details regarding the training recipe and its ablation are shown in supplementary materials.

3. Experiments

3.1. Efficiency Analysis

As shown in Figure 1 and 4, LongVidRWKV consumes significantly less GPU memory than its transformer-based counterparts. Also, LongVidRWKVachieves faster inference speed. Compared with InternVL-1.5 2B [11], LongVidRWKV takes less GPU memory when processing videos with 1,024 sampled frames and is 8X faster in inference speed when taking 60 input frames.

3.2. Quantitative Evaluation

3.2.1. Short Video Understanding

We conduct experiments to evaluate short video perception on multiple public datasets that provide various annotations with average video durations under 120 seconds. This includes open-ended question-answering, short captioning, and detailed captioning. For open-ended video question answering and dense captioning, we use LLM-assisted evaluation with default model choices and hyperparameter settings in LMMs-Eval [20]. Although the RWKV[32] LLM backbone is pretrained only on publicly available data, LongVidRWKV exceeds Gemini-1.5-Pro in VDC [5], a video detailed captioning benchmark.

3.2.2. Long Video Understanding

Since the RWKV LLM consumes much less memory than its Transformer counterparts when processing long input sequences, we are able to train LongVidRWKV on up to 60 input frames, which significantly enhances performance on common tasks like action counting (AC), action localization (AL) and needle QA (NQA), as is also observed in [43]. It is interesting that LongVidRWKV achieves comparable accuracy while consuming only about 60 tokens per frame, justifying our motivation for introducing token merge due to the spatial redundancy in long video understanding.

Table 2. Comparison with other methods on MLVU [55] and MovieChat-1k [36]. Both datasets have an average video length of about 12 minutes. Results with * are evaluated in-house, while others are sourced from official leaderboards. The best result is highlighted in bold, and the second best is underlined. CTX denotes LLM pretrained context length and maximum context length for proprietary models.

Models	Input	CTX	Size	MLVU								MovieChat-1K	
				AVG	AR	ER	AO	AC	TR	NQA	PQA	Global	Break
GPT4-o	0.5fps	128k	-	**54.5**	**68.8**	47.8	46.2	35.0	**83.7**	42.9	57.1	-	-
LLaVA-OneVision* [21]	32 frm	132k	0.5B	50.3	58.5	52.4	28.6	30.9	67.0	33.3	42.8	-	-
Qwen2-VL* [41]	32 frm	132k	2B	48.7	54.7	47.6	30.9	28.6	73.8	40.4	60.5	-	-
InternVL2* [10]	32 frm	200k	2B	48.2	57.4	**57.1**	35.7	33.4	66.7	28.5	50.0	-	-
LongVA [47]	256 frm	224k	7B	42.1	41.0	39.6	17.1	23.3	81.3	46.7	46.0	55.9	56.5
ShareGPT4Video [7]	16 frm	8k	8B	34.2	25.6	45.3	17.1	8.3	73.6	31.7	38.0	69.0	60.9
InternVL-1.5 [11]	16 frm	8k	26B	37.9	51.3	24.5	14.3	13.3	80.2	40.0	42.0	57.7	61.1
VILA-1.5 [25]	14 frm	276k	40B	46.2	56.4	35.8	34.3	11.7	84.7	38.3	**62.0**	57.2	60.1
LongVidRWKV **(ours)**	48 frm	4k	2B	52.7	59.5	**57.1**	33.2	**42.9**	69.0	45.2	61.9	**84.0**	**64.0**

Table 3. Results on MVBench [22] whose videos primarily range from 5s to 35s. Results with * are evaluated in-house, while others are sourced from official leaderboards. The best result is highlighted in bold, and the second best is underlined. We find that despite only being trained on public datasets, LongVidRWKV is competitive with models of similar size trained on large-scale high-quality proprietary data.

Models	Size	MVBench																	
		Avg.	UA	AC	MA	OE	ST	AL	AP	AS	CO	CI	EN	FGA	MC	MD	OI	OS	SC
LLaVA-OneVision [21]	0.5B	45.5	72.5	43.5	49.5	50.0	85.5	12.5	41.0	54.0	49.0	35.5	21.5	42.0	33.0	17.5	61.0	32.5	45.5
InternVL2* [10]	2B	52.9	60.5	30.5	**78.0**	79.0	83.5	31.0	**67.0**	**72.0**	36.0	**55.0**	32.0	38.0	**65.5**	32.0	64.0	30.0	44.5
Qwen2-VL* [41]	2B	**53.5**	73.0	43.5	75.5	**82.0**	82.0	12.5	41.0	54.0	49.0	35.5	21.5	**48.0**	55.0	45.0	55.0	29.5	43.0
LongVA* [47]	7B	50.8	68.5	47.0	56.5	49.5	89.0	45.0	58.0	55.6	61.5	41.0	**39.0**	43.5	28.0	36.5	**65.5**	30.5	49.0
ShareGPT4Video* [7]	8B	47.2	56.5	34.0	74.5	81.8	84.5	34.5	48.0	45.2	46.0	51.0	25.0	35.0	60.5	**54.0**	56.5	33.0	50.0
InternVL-1.5* [11]	26B	50.6	73.5	27.5	62.5	44.0	**89.5**	39.3	61.0	62.0	**64.0**	40.5	34.5	46.5	33.0	36.0	**65.5**	28.5	53.0
VILA-1.5* [25]	40B	42.7	60.0	41.5	34.5	50.0	69.5	36.5	39.5	40.5	44.0	40.0	27.0	33.0	37.0	27.5	59.5	38.0	47.5
LongVidRWKV **(ours)**	2B	53.2	**75.0**	**52.0**	65.5	62.5	87.0	**48.0**	47.5	49.5	47.0	52.0	35.0	46.5	48.5	44.0	54.0	37.5	**53.5**

3.3. Ablation Study on Input Token Order

Previous Linear Attention-based image-LMMs [16] and video encoders [23] enhance visual modeling by scanning visual tokens bidirectionally, increasing computational complexity. In contrast, LongVidRWKV simply reorders merged visual tokens in ascending order by size to better leverage pretrained unidirectional textual data. RWKV's [32] recurrent mechanism acts as an implicit position encoding but is disrupted by visual token merging. However, the SigLIP encoder preserves positional information via explicit embeddings, ensuring spatial coherence. Since merging occurs within the same frame, temporal information remains intact. As is shown in Table 4, we examine three sorting strategies—random, ascending, and descending—before feeding merged tokens into RWKV, finding that ascending order performs best. This likely helps RWKV6 prioritize critical information in tasks like visual question answering by leveraging its data-dependent token-shifting mechanism [32] to retain key frame details.

Table 4. Ablation on input order for merged visual tokens within a frame, where ascending order suggests tokens that are never merged come first among tokens of the same frame. We found that sorting merged tokens in an ascending manner brings the best performance. The best result is highlighted in bold.

Token Order	ANet [45]	VATEX [42]	VDC [5]	MovieChat-1K [36]
Random	53.1	67.6	40.9	76.5
Descending	55.0	67.0	41.1	76.0
Ascending	**56.3**	**68.5**	**41.3**	**78.5**

4. Conclusion

In this paper, we introduce LongVidRWKV, the first video-based LMM with a fully recurrent RWKV [31] LLM backbone. By incorporating a token merging strategy, LongVidRWKVreduces computational overhead without sacrificing performance and mitigates overfitting to training context lengths—an issue common in linear attention variants. We conduct extensive experiments on multiple video understanding benchmarks, achieving improved performance with more input frames compared to larger LMMs. The ablation studies validate the effectiveness of the token merging ratio and the token reordering strategy we propose. We hope this work provides a strong baseline for hybrid LMMs and inspires further research in efficient and scalable architectures for long video understanding.

Acknowledgments

We thank the RWKV team for their innovative work in sequence modeling architecture, which inspired this research.

References

[1] Kazi Hasan Ibn Arif, JinYi Yoon, Dimitrios S Nikolopoulos, Hans Vandierendonck, Deepu John, and Bo Ji. Hired: Attention-guided token dropping for efficient inference of high-resolution vision-language models in resource-constrained environments. *arXiv preprint arXiv:2408.10945*, 2024. 1

[2] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024. 1

[3] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. In *International Conference on Learning Representations*, 2023. 1, 2

[4] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 3

[5] Wenhao Chai, Enxin Song, Yilun Du, Chenlin Meng, Vashisht Madhavan, Omer Bar-Tal, Jeng-Neng Hwang, Saining Xie, and Christopher D Manning. Auroracap: Efficient, performant video detailed captioning and a new benchmark. *arXiv preprint arXiv:2410.03051*, 2024. 1, 3, 4

[6] Guo Chen, Yifei Huang, Jilan Xu, Baoqi Pei, Zhe Chen, Zhiqi Li, Jiahao Wang, Kunchang Li, Tong Lu, and Limin Wang. Video mamba suite: State space model as a versatile alternative for video understanding. *arXiv preprint arXiv:2403.09626*, 2024. 1

[7] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. Sharegpt4video: Improving video understanding and generation with better captions. *arXiv preprint arXiv:2406.04325*, 2024. 3, 4

[8] Yingfa Chen, Xinrong Zhang, Shengding Hu, Xu Han, Zhiyuan Liu, and Maosong Sun. Stuffed mamba: State collapse and state capacity of rnn-based long-context modeling. *arXiv preprint arXiv:2410.07145*, 2024. 2

[9] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Zhong Muyan, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023. 3

[10] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 4

[11] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 3, 4

[12] Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*, 2024. 1

[13] Yuchen Duan, Weiyun Wang, Zhe Chen, Xizhou Zhu, Lewei Lu, Tong Lu, Yu Qiao, Hongsheng Li, Jifeng Dai, and Wenhai Wang. Vision-rwkv: Efficient and scalable visual perception with rwkv-like architectures. *arXiv preprint arXiv:2403.02308*, 2024. 1

[14] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 1

[15] Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13504–13514, 2024. 1

[16] Haowen Hou, Peigen Zeng, Fei Ma, and Fei Richard Yu. Visualrwkv: Exploring recurrent neural networks for visual language models. *arXiv preprint arXiv:2406.13362*, 2024. 1, 2, 3, 4

[17] Shibo Jie, Yehui Tang, Jianyuan Guo, Zhi-Hong Deng, Kai Han, and Yunhe Wang. Token compensator: Altering inference cost of vision transformer without re-tuning. In *European Conference on Computer Vision*, pages 76–94. Springer, 2025. 1

[18] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pages 5156–5165. PMLR, 2020. 1

[19] Guolin Ke, Di He, and Tie-Yan Liu. Rethinking positional encoding in language pre-training. *arXiv preprint arXiv:2006.15595*, 2020. 2

[20] Bo Li, Peiyuan Zhang, Kaichen Zhang, Fanyi Pu, Xinrun Du, Yuhao Dong, Haotian Liu, Yuanhan Zhang, Ge Zhang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Accelerating the development of large multimoal models, 2024. 3

[21] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 3, 4

[22] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024. 4

[23] Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding. In *European Conference on Computer Vision*, pages 237–255. Springer, 2025. 1, 2, 4

[24] Shufan Li, Harkanwar Singh, and Aditya Grover. Mamba-nd: Selective state space modeling for multi-dimensional data. In *European Conference on Computer Vision*, pages 75–92. Springer, 2024. 1

[25] Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi,

and Song Han. Vila: On pre-training for visual language models. *arXiv preprint arXiv:2312.07533*, 2023. 4

[26] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 1

[27] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 2

[28] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 1

[29] Hui Lu, Albert Ali Salah, and Ronald Poppe. Videomambapro: A leap forward for mamba in video understanding. *arXiv preprint arXiv:2406.19006*, 2024. 1, 2

[30] Yunsheng Ma, Amr Abdelraouf, Rohit Gupta, Ziran Wang, and Kyungtae Han. Video token sparsification for efficient multimodal llms in autonomous driving. *arXiv preprint arXiv:2409.11182*, 2024. 1

[31] Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, et al. Rwkv: Reinventing rnns for the transformer era. *arXiv preprint arXiv:2305.13048*, 2023. 1, 3, 4

[32] Bo Peng, Daniel Goldstein, Quentin Anthony, Alon Albalak, Eric Alcaide, Stella Biderman, Eugene Cheah, Xingjian Du, Teddy Ferdinan, Haowen Hou, et al. Eagle and finch: Rwkv with matrix-valued states and dynamic recurrence. *arXiv preprint arXiv:2404.05892*, 2024. 1, 2, 3, 4

[33] Guo Longteng Chen Sihan Zhao Zijia Sun Mingzhen Wu Qi Qiao Yanyuan, Yu Zheng and Liu Jing. Vl-mamba: Exploring state space models for multimodal learning. *arXiv preprint arXiv:2403.13600*, 2024. 1

[34] Weiming Ren, Wentao Ma, Huan Yang, Cong Wei, Ge Zhang, and Wenhu Chen. Vamba: Understanding hour-long videos with hybrid mamba-transformers, 2025. 1

[35] Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. *arXiv preprint arXiv:2403.15388*, 2024. 1

[36] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Xun Guo, Tian Ye, Yan Lu, Jenq-Neng Hwang, et al. Moviechat: From dense token to sparse memory for long video understanding. *arXiv preprint arXiv:2307.16449*, 2023. 1, 4

[37] Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2021. 2

[38] Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. Dycoke: Dynamic compression of tokens for fast video large language models. *arXiv preprint arXiv:2411.15024*, 2024. 1

[39] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2

[40] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 2

[41] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1, 4

[42] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4581–4591, 2019. 3, 4

[43] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *arXiv preprint arXiv:2407.15754*, 2024. 3

[44] Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing linear transformers with the delta rule over sequence length. *Advances in Neural Information Processing Systems*, 37:115491–115522, 2025. 1

[45] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9127–9134, 2019. 3, 4

[46] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023. 2

[47] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024. 3, 4

[48] Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Khai Hao, Xu Han, Zhen Leng Thai, Shuo Wang, Zhiyuan Liu, et al. ınftybench: Extending long context evaluation beyond 100k tokens. In *ACL (1)*, 2024. 2

[49] Zhenyu Zhang, Benlu Wang, Weijie Liang, Yizhi Li, Xuechen Guo, Guanhong Wang, Shiyan Li, and Gaoang Wang. Sam-guided enhanced fine-grained encoding with mixed semantic learning for medical image captioning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1731–1735. IEEE, 2024. 1

[50] Zhonghan Zhao, Wenhao Chai, Xuan Wang, Boyi Li, Shengyu Hao, Shidong Cao, Tian Ye, and Gaoang Wang. See and think: Embodied agent in virtual environment. In *European Conference on Computer Vision*, pages 187–204. Springer, 2024.

[51] Zhonghan Zhao, Wenhao Chai, Xuan Wang, Ke Ma, Kewei Chen, Dongxu Guo, Tian Ye, Yanting Zhang, Hongwei Wang, and Gaoang Wang. Steve series: Step-by-step construction of agent systems in minecraft. *Computer Vision and Pattern Recognition Workshop*, 2024.

[52] Zhonghan Zhao, Kewei Chen, Dongxu Guo, Wenhao Chai, Tian Ye, Yanting Zhang, and Gaoang Wang. Hierarchical

auto-organizing system for open-ended multi-agent navigation. *arXiv preprint arXiv:2403.08282*, 2024.

[53] Zhonghan Zhao, Ke Ma, Wenhao Chai, Xuan Wang, Kewei Chen, Dongxu Guo, Yanting Zhang, Hongwei Wang, and Gaoang Wang. Do we really need a complex agent system? distill embodied agent into a single model. *arXiv preprint arXiv:2404.04619*, 2024.

[54] Zhonghan Zhao, Wenwei Zhang, Haian Huang, Kuikun Liu, Jianfei Gao, Gaoang Wang, and Kai Chen. Rig: Synergizing reasoning and imagination in end-to-end generalist policy. *arXiv preprint arXiv:2503.24388*, 2025. 1

[55] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024. 4

[56] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. In *International Conference on Machine Learning*, 2024. 1