Bringing RNNs Back to Efficient Open-Ended Video Understanding

Anonymous CVPR submission

Paper ID 9135

Abstract

The challenge of long video understanding lies in its high computational complexity and prohibitive memory cost, since the memory and computation required by transformer-based LLMs scale quadratically with input sequence length. We proposed AURORALONG *to address this challenge by replacing the LLM component in MLLMs with RWKV, an RNN-like language model that handles input sequence of arbitrary length with constant-size hidden states. To further increase throughput and efficiency, as well as to reduce the gap between RWKV's 4k context length and the long video token sequence length, we combine visual token merge with linear RNN models by reordering the visual tokens by their sizes in ascending order.* AURORALONG *shows superior performance on various video benchmarks, for example, obtaining an average accuracy of 87.0 on scene transition in MVBench, beating GPT-4V (83.5) and Gemini Pro (75.4), highliting the possibilities that efficient linear RNNs can democratize long video understanding. To our best knowledge, we are the first to use a non-transformer LLM backbone for video understanding.*

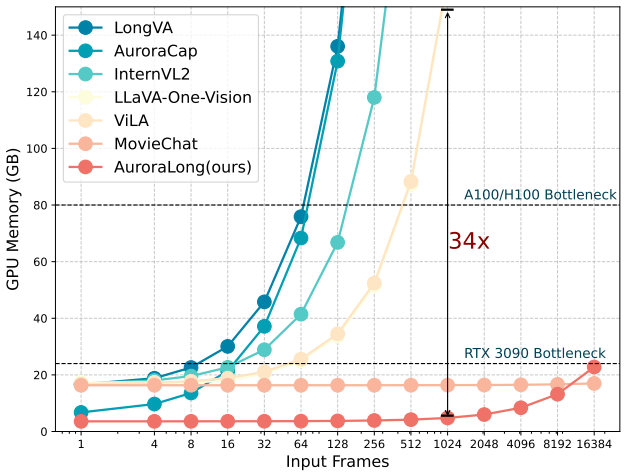


Figure 1. VRAM cost under gigabyte (GB) (y-axis) v.s. frame number (x-axis) comparison. We test the visual-only inference of all methods. While the previous method can only support around 100 frames of inference with A100 or H100, AURORALONG can handle videos with over 10 thousands frames on a 24GB GPU.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

1. Introduction

Through the integration of Transformer-based large language models (LLMs) [1, 77, 91] and visual extractors, large multimodal models (LMMs) [5, 11, 19, 39, 45, 68, 76, 109, 111] have demonstrated impressive abilities such as captioning and visual question-answering. Among these, image-based LMMs have shown strong performance in academic domains through effective modality alignment and visual instruction tuning. Expanding from image-based LMMs to video-based LMMs is a natural progression, as videos can be viewed as sequences of frames. While most LMMs [19, 65, 71, 72] start by loading pre-trained weights from image models and incorporate additional temporal modules on video-text data, [12] find that LLaVA-like models can be easily adapted to video without any additional parameters, relying solely on high-quality video-text instruction data for fine-tuning.

However, naively treating videos as a series of image frames can result in significant computational overhead. Although Transformers improve the modeling of long-range dependencies, their architecture is burdened by the inherent computational and memory complexity of the self-attention mechanism, leading to computational and memory requirements that increase quadratically with sequence length. Currently, linear RNN large language models [22, 33, 63, 64] utilize linear attention to replace the softmax attention in Transformer-based models to effectively reduce computational cost. RWKV [64] combines the parallelized training benefits of Transformers with the constant inference memory cost benefits of RNNs/LSTMs. Additionally, since their memory size is constant, although RNN-based language models can process infinitely long inputs, there is an upper bound to the amount of information the state can represent, and tokens beyond this upper bound

will be forgotten.

Reducing the number of visual tokens in language model inputs has been explored in various LMMs [2, 38, 60]. Token Merging (ToMe) [9] is first introduced based on token similarity, which has proven effective in image and video classification tasks. To develop more efficient LMMs, subsequent works propose various token merging strategies. FastV[14] reduces visual tokens based on attention ranks within LLM layers, MovieChat [71] utilizes a temporal memory bank, and Chat-UniVi [39] merges visual tokens from both spatial and temporal dimensions. Similarly, to accommodate more input frames within a fixed context length, we apply a token merging method to each layer of the vision transformer, reducing visual token while preserving visual information.

In this paper, we present AURORALONG, combining the simple yet efficient token merging strategy with linear RNN models by reordering the visual tokens to further increase throughput and efficiency, which is proved to be effective in various video understanding tasks. Since RWKV [64] is trained with a context length of only 4,096 tokens and remains challenging to extend with video-text data, we gradually combine similar visual tokens in each transformer layer with a bipartite soft matching algorithm to reduce the number of visual tokens. Previous works [36, 47, 58] improve visual sequence modeling in linear attention models by bidirectional scanning of visual tokens, increasing computational complexity. To better utilize the pretrained unidirectional textual sequence, we simply reorder the merged visual tokens within each layer by sorting them in descending order based on the number of tokens they combine. As shown in Figure 1, AURORALONG outperforms existing methods in terms of VRAM cost in the initial VRAM usage and remains approximately constant as the number of input frames grows. Under the RTX 3090 Bottleneck, AURORALONG can process up to 16K frames and has a 34 times advantage over other methods in VRAM cost when processing 1,024 frames. Following this pattern, our experiments show that we can use only 10% to 20% visual tokens compared to the original tokens generated by ViT with a marginal performance drop in various benchmarks.

Our main contributions are summarized as follows:

- We are the first to use a linear RNN model as the LLM backbone for video understanding, presenting a novel hybrid architecture that can handle video input of arbitrary length with lower memory requirement.
- We propose a training-free reordered visual token merge strategy to increase model throughput while retaining visual information for RNN-based large language models.
- Our architecture consistently performs favorably against several state-of-the-art larger LMMs across various video understanding tasks, while reducing computational complexity and memory consumption.

2. Related Work

2.1. Long-form Video Understanding

With the develop of LLMs and LMMs [13, 28, 44, 49, 59, 61, 72, 88, 101, 106], many recent works have broadened their application to video understanding tasks, especially for long video understanding [42, 70, 74, 84, 89, 97, 99, 100, 103, 111]. For long videos, the computational complexity and memory costs associated with long-term temporal connections are significantly increased, posing additional challenges. MovieChat [71] introduces a training-free hierarchical memory bank that temporally consolidates visual inputs, enabling the sampling of thousands of frames. TimeChat [65] develop time-aware frame encoder and sliding video Q-Former to capture detailed video content. Long video understanding is evaluated using benchmarks [3, 6, 15, 27, 31, 34, 56, 82, 87, 96, 110] typically classified as open-ended or multiple-choice questions. For open-ended questions, benchmarks like MovieChat-1K [71] focus on 8-minute-long movie clips. Regarding multiple-choice questions, EgoSchema [62] focuses on 3-minute-long egocentric videos. Video-MME [31] features a diverse dataset of 900 videos across six primary visual domains and varying durations from 11 seconds to 1 hour. LongVideoBench [86] encompasses human-annotated multiple-choice questions across 17 categories with varying lengths of up to 1 hour.

2.2. Linear RNN Large Language Model

Current advances in large language models (LLMs) [1, 77, 91] mostly focus on Transformer-based architectures, showcasing remarkable achievements across various natural language processing tasks. However, they suffer from quadratic complexity issues in both computation and memory. Consequently, recent interest has arisen in RNN-based language models [30, 48, 66, 69]. Compared to Transformer-based models, RNN-based language models inherently handle temporal sequential data, and their per-token inference cost does not increase with sequence length. However, classical RNN-based models [25, 32, 67] pose challenges in parallelization across time dimensions during training. Linear attention [40] replaces the softmax attention in Transformer-based models with kernel-based approximations to reduce computational cost, achieving an inference complexity of $\mathcal{O}(N)$. Some linear RNN-based approaches [22, 24, 33, 63, 64, 92, 108] have demonstrated notable capabilities in many language processing tasks. Among linear attention variants, RWKV enjoys both the benefit of transformer and RNN/LSTM, which are parallelizable training and constant inference memory cost respectively. However, [16] indicates that some of these language models may fail to extrapolate beyond their training length.

CVPR #9135

3. Method

3.1. Preliminaries

RWKV backbone. RWKV [64] combines the parallelizable training efficiency of Transformers with the sequential inference capabilities of RNNs. Its recurrent mechanism examines only the immediate previous token, enabling unbounded sequence lengths during inference without increased computational power or memory requirements. Additionally, since RWKV does not utilize explicit positional encoding, it can handle contexts of arbitrary length without modification. RWKV's core architecture computes a weighted sum of past values, modulated by a receptance vector, to efficiently facilitate information flow across time steps, which can be expressed as:

$$\begin{aligned} \alpha_i &= e^{-w}\alpha_{i-1} + e^{k_i}v_i, \\ \beta_i &= e^{-w}\beta_{i-1} + e^{k_i}, \\ \text{wkv}_i &= \frac{e^{u+k_i}v_i + \alpha_{i-1}}{e^{u+k_i} + \beta_{i-1}}, \end{aligned} \tag{1}$$

where α_i and β_i are recursive state variables; k_i and v_i are the key and value vectors at time step i; w controls the decay rate; and u is an additional learned parameter.

Token merging. Since the RWKV [64] is trained on a context length of merely 4,096 tokens, we adopt Token Merge (ToMe) [8] to reduce the number of visual tokens passed to RWKV. By combining similar visual tokens in the Vision Transformer [26], ToMe increases the throughput of vision encoders and has been proven effective across various tasks. Token Merging is applied between the attention and MLP within each transformer block as:

1. Alternatively partition the tokens into two sets $\mathcal{A}$ and $\mathcal{B}$ of roughly equal size.
2. For each token in set $\mathcal{A}$, calculate the token similarity with each token in set $\mathcal{B}$ based on cosine similarity of the *Key* features in attention block.
3. Use bipartite soft matching and then select the most similar r pairs.
4. Merge the tokens using weighted average, and record the token size.
5. Concatenate the two sets $\mathcal{A}$ and $\mathcal{B}$ back together again.

Once the tokens have been merged, they actually carry features of more than one input patch. Therefore, the *proportional attention* [9] is formulated by

$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{\mathbf{d}}} + \log \mathbf{s}\right) \tag{2}$$

where $\mathbf{s}$ represents the number of patches each token represents after token merging in previous layers. In AURORALONGwe conduct frame-wise token merging, of which more visualization can be found in the Appendix.

Algorithm 1 Sorted 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 $\mathcal{V}[:\mathcal{N}-2]$ **do**
  # $\mathcal{X} \in [batch, tokens, channels]$
  $\mathcal{X} \leftarrow \text{Attention}_n(\mathcal{X})$
  # Assign tokens into $Set\ \mathcal{A}, Set\ \mathcal{B}$
  $\mathcal{A}, \mathcal{B} \leftarrow \mathcal{X}[..., :: 2, :], \mathcal{X}[..., 1 :: 2, :]$
  $Scores \leftarrow \text{similarity}(\mathcal{A}, \mathcal{B})$
  # Ignore CLS tokens
  $Scores[..., 0, :] \leftarrow$ `-math.inf`
  # 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

3.2. Method

3.2.1. Network Architecture

We inherit the architecture of LLaVA-1.5 [53] with different choices for the vision encoder and the language model. Specifically, we use SigLIP [98] (large-patch16-384) as the vision encoder to encode video frames and remove the final vision transformer layer following [81], with a simple two-layer MLP serving as the cross-modal connector. We use RWKV-v6-Finch [64] as the LLM backbone for its ability to handle sequences of arbitrary length with constant memory cost. However, since RNN models like RWKV lack context extension techniques like rotary position embeddings (RoPE) or multimodal positional embedding (M-RoPE) [73, 80], necessitating the introduction of visual token merge [9] to reduce the number of visual tokens.

3.2.2. Reordered Visual Token Merge

Despite RWKV's [64] efficiency in handling inputs of arbitrary length, [16] indicates that linear attention models tend to overfit to their pretrained context length. The scarcity of high-quality vision-language data compared to the vast amount of unidirectional language-only data for training RWKV makes it challenging to fine-tune the model to accommodate longer multimodal sequences. Given that RWKV [64] is pretrained with a context length of only 4,096 tokens, we introduce Token Merging [9] to merge similar visual tokens, narrowing the length gap between pretrained context and the long sequence of visual tokens.

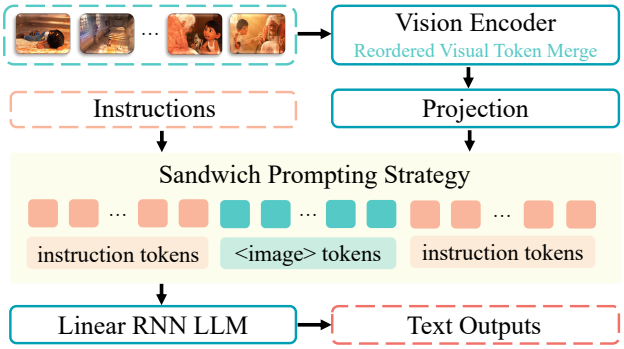


Figure 2. AURORALONG prompting strategy overview. Following VisualRWKV [36], we adopt sandwich prompting strategy, which places image tokens in the middle of instruction tokens.

Unlike [85] which merges visual tokens within a video segment, we conduct token merge at a spatial level within each frame, given the consideration that when sampled at 1 FPS or lower, frame-to-frame similarity is already quite low except in static scenes, thereby obviating the need to combine visual tokens temporally.

To model visual input sequence order, Transformers utilize explicit positional embedding[41, 73, 77, 79], while RNNs model sequence order implicitly due to their recurrent nature. Previous works [36, 47, 58] attempt to enhance the visual modeling capabilities of linear attention models like RWKV [63, 64] and Mamba [22, 33] by bidirectionally scanning visual tokens, leading to additional computation overhead. Therefore, we propose a simpler, training-free visual token reordering strategy to better utilize the pretrained unidirectional textual modeling capabilities while retaining as much spatial information as possible. Specifically, as illustrated in Algorithm 1, within each ViT layer, after merging similar visual tokens, we reorder the tokens by sorting them according to the number of visual patches they represent. We experiment with several sorting orders, and select the ascending order for its superior performance.

3.2.3. Prompting Strategy

Since RWKV [63] and other linear RNN language models are prompt-sensitive, it is crucial to employ an appropriate prompting strategy to enhance AURORALONG's instruction following ability. Following VisualRWKV [36], we utilize the sandwich prompting strategy, and insert the reordered merged visual tokens between the instruction tokens as illustrated in Figure 2.

3.3. Training Recipe

Following AURORACAP [12], we further adopt a three-stage training strategy, which can be noted as Pretraining stage, Vision stage and Language stage. The training data used in each stage are shown in Appendix.

Table 1. Training hyper-parameters for AURORALONG

Hyper-parameters	Pretrain	Vision	Language
ViT	×	✓	✓
MLP	✓	✓	✓
LLM	×	×	✓
epoch	1	1	2
peak learning rate	1e-4	4e-5	1e-5 / 8e-6
batch size	128	2,048	512
visual token kept ratio	100%	100%	10%

Pretraining stage. Similar to LLaVA [55], we first learn the alignment between visual features from the vision encoder and the word embedding space of RWKV [64]. To achieve this, we freeze the pretrained ViT and LLM, training solely the multimodal connector on image-caption pairs.

Vision stage. To achieve better vision generalization, we next unfreeze the pretrained ViT while freezing the LLM during the vision stage. Note that the data we use for this stage are public datasets from various image-based computer vision tasks, which may involve labels consisting of only a few words or short phrases. Therefore, we freeze the LLM to avoid degradation in its performance as in [12].

Language stage. Finally, we conduct end-to-end training using high-quality public data. To maintain context length similarity among samples and improve training efficiency, we distinguish the single-image data from the multiple-image samples (mainly from videos). Additionally, we set the visual token retention ratio to 0.1 to further enhance training efficiency. We start by training with high-quality single-image data and then transit to video datasets with a lower learning rate. To improve video understanding performance, we train on video captioning samples and video question answering samples for two epochs.

4. Experiments

In this section, we present the implementation details of AURORALONG and conduct quantitative and qualitative evaluations comparing AURORALONG with previous methods on various video understanding tasks. We also conduct ablation studies to evaluate model performance.

4.1. Implementation Details

We only compute cross-entropy loss for auto-regressive text generation. For all training stages, we use the AdamW [57] optimizer with a cosine decay schedule, setting the optimizer hyperparameters β_1 and β_2 to 0.9 and 0.999, respectively. Each stage employs a linear warmup schedule with a start factor of 1e-5 and a warmup ratio of 0.03. The differences in training hyperparameters across all stages are

Table 2. Results on comprehensive short video understanding benchmarks. Detailed results are provided in the Appendix. The best result is highlighted in bold, and the second best is underlined. We find that AURORALONG outperforms existing methods across various short video understanding tasks with a similar size and is competitive with models that have much larger parameters.

Models	Size	#Frame	VDC w. VDCscore						ANet		VATEX
			Avg.	Short	Camera	Background	Main Object	Detailed	Acc.	Score	BLEU@1
Proprietary Models											
Gemini-1.5-Pro	-	1fps	41.73	35.71	38.68	43.84	47.32	43.11	-	-	-
Open-Source LMMs											
ShareGPT4Video [13]	3B	16	36.17	39.08	33.28	35.77	37.12	35.62	-	-	-
BLIP-3-Video [90]	4B	-	-	-	-	-	-	-	56.9	3.6	-
Video-LLAVA [51]	7B	8	32.80	30.67	37.48	32.50	36.01	27.36	45.3	3.3	-
LLAMA-VID [50]	7B	1fps	30.86	29.92	39.47	28.01	31.24	25.67	47.4	3.3	-
Video-ChatGPT [61]	7B	100	31.12	29.36	37.46	33.68	30.47	24.61	35.2	2.8	-
Chat-UniVi [39]	7B	64	-	-	-	-	-	-	46.1	3.3	-
LLAVA-NeXT [107]	7B	32	35.46	30.63	39.73	36.54	36.54	33.84	53.5	3.2	-
Video-LLAMA2 [19]	7B	16	-	-	-	-	-	-	50.2	3.3	-
LongVA [105]	7B	64	34.50	31.94	35.32	36.39	40.95	27.91	-	2.8	-
VideoChat2 [46]	7B	16	-	-	-	-	-	-	49.1	3.3	-
LLAVA-OneVision [43]	7B	32	37.45	32.58	37.82	37.43	38.21	41.20	56.6	-	
AuroraCap [12]	7B	16	38.21	32.07	43.50	35.92	39.02	41.30	**61.8**	3.8	57.1
Video-CCAM [29]	9B	96	-	-	-	-	-	-	59.7	3.8	-
AURORALONG (ours)	2B	1fps	**42.54**	**38.89**	**43.70**	**40.26**	**46.32**	**43.54**	60.0	**4.2**	**68.5**

detailed in Table 1. For visual data preprocessing, we resize each visual input so that its short side is 384 pixels while maintaining the original aspect ratio. For token merging, we keep the number of visual tokens being merged the same among each Vision Transformer [26, 98] layer. Our model was trained on 8 NVIDIA A800 GPUs.

4.2. Quantitative Evaluation

4.2.1. Short Video Understanding

We primarily conduct three tasks to assess the short video understanding capability of AURORALONG: video question answering, video captioning, and video detailed captioning.

We conducted 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 tasks like ActivityNet [10], sparse captioning tasks like VATEX [83], and dense captioning like VDC [12]. For open-ended video question answering and dense captioning, we use LLM-assisted evaluation with default model choices and hyperparameter settings in LMMs-Eval [7, 102]. Following the standard practice in VideoLLM evaluation, we report a percentage accuracy and an average score on a scale from 0 to 5. For video sparse captioning, we assess AURORALONG using the CIDEr (C), BLEU-4 (B@4), BLEU-1 (B@1), METEOR (M), and ROUGE-L (R) metrics on MSR-VTT and VATEX, presenting CIDEr scores. Additional results are provided in the Appendix. As illustrarted in 2, when trained on the same dataset consisting of high quality recaptioned visual instruction samples, AURORALONG achieves even better performance than AuroraCap[12], whose Transformer-based LLM backbone [20] is finetuned on LLaMA-2 [78], a strong foundation model pretrained on large-scale proprietary data that are carfully curated. Although the RWKV[64] LLM backbone are pretrained only on publicly available data, AURORALONG exceeds Gemini-1.5-Pro on average in VDC[12], a dense captioning benchmark for short videos.

4.2.2. Long Video Understanding

Although AURORALONGwas only trained on short video datasets mostly consisting of videos with 8 to 12 frames, we evaluate it on multiple long video quesetion-answering benchmarks[46, 71, 110] to assess its zero-shot long video understanding capability. To provide a fair comparison, we follow the standard and default settings in each benchmark. To validate AURORALONG's capability in long video understanding, we compare with industry leading proprietary models [4, 37, 75, 93] and open weight models [43, 51, 71, 104] that are up to 13 times larger than AURORALONG in terms of model parameter size.

Since most long video questions answering tasks requires understanding of multiple frames, many prior models are trained on more visual frames and use much more tokens per frame than AURORALONG. It is interesting that AURORALONG achieves comparable accuracy while consuming only 58 tokens per frame, justfying our motivation of introducing token merge due to the spatial redundancy nature of long video understanding. Note that although AURORALONG was only trained on short videos within one-minute its RWKV [64] backbone was only

Table 3. Results on comprehensive long video understanding benchmarks. The best result is highlighted in bold, and the second best is underlined. We find that despite only trained on short videos, AURORALONG outperforms models with up to 20X larger parameters across various long video understanding tasks. More detailed results are provided in the Appendix.

Models	Size	MovieChat-1K		MVBench				MLVU			
		Global	Breakpoint	AC	MA	OE	ST	TR	ER	AO	AC
Proprietary Models											
GPT4-V	-	-	-	39.0	22.5	18.5	83.5	-	-	-	-
GPT4-o	-	-	-	-	-	-	-	**68.8**	47.8	**46.2**	35.0
GPT4-Turbo	-	-	-	-	-	-	-	61.5	41.5	22.9	6.7
Gemini Pro	-	-	-	3.9	41.5	43.5	75.4	-	-	-	-
Qwen-VL-Max	-	-	-	-	-	-	-	53.8	26.4	20.0	11.7
Claude-3-Opus	-	-	-	-	-	-	-	30.8	17.0	10.0	6.7
Open-Source Video LMMs											
Video-LLaMA [51]	7B	51.7	39.1	-	-	-	-	-	-	-	-
LLAMA-VID [50]	7B	-	-	42.0	44.5	55.6	84.5	23.1	11.3	18.6	15.0
mPLUG-Owl-V [94]	7B	-	-	34.5	31.5	36.0	34.5	15.4	13.2	14.3	20.0
Video-ChatGPT [61]	7B	47.6	48.0	30.5	39.5	54.0	31.0	17.9	32.1	17.1	13.3
MovieChat [71]	7B	62.3	48.3	-	-	-	-	10.3	15.1	17.1	15.0
Video-LLAVA [61]	7B	-	-	34.0	32.5	48.0	43.0	38.5	26.4	20.0	21.7
LLaVA-1.6 [54]	7B	-	-	-	-	-	-	17.9	26.4	21.4	16.7
LongVA [105]	7B	-	-	-	-	-	-	41.0	39.6	17.1	23.3
VideoChat2 [46]	7B	-	-	-	-	-	-	30.8	28.3	17.1	23.3
ShareGPT4Video [13]	8B	-	-	-	-	-	-	25.6	45.3	17.1	8.3
VideoLLAMA2 [19]	13B	-	-	-	-	-	-	12.8	17.0	15.7	8.3
InternVL-1.5 [17]	26B	-	-	-	-	-	-	51.3	24.5	14.3	13.3
VILA-1.5 [52]	40B	-	-	-	-	-	-	56.4	35.8	34.3	11.7
AURORALONG(ours)	2B	**84.0**	**64.0**	**44.5**	**59.0**	**62.5**	**87.0**	59.5	**54.8**	29.4	**42.9**

trained on a context length of 4096, it still outperforms several long context Transformer-based video understanding models on long video tasks without any modifications such as adjusting r in RoPE [73] as is usually practiced on Transformer-based video understanding models. This generalizability aligns with the loss curve its LLM backbone shows when validated on extended textual context length that up to 4X its pretrained context length, as is illustrated in Appendix. We provide more in-depth analysis of AURORALONG's context length generalizability in ablation study.

4.2.3. Efficiency Analysis

As shown in Figure 1 and Figure 3, we compare the GPU memory consumption and inference speed directly with existing leading methods. While the memory consumption of other transformer-based models increases rapidly in a quadratic manner, AURORALONG consumes significantly less GPU memory, which grows linearly with respect to number of input frames. Despite the fact that when processing videos exceeding 10,000 frames, AURORALONG requires slightly more GPU memory than MovieChat [71] which adopts a constant sliding window for short term feature extraction, AURORALONGdoes not require additional memory mechanisms and consumes substantially less GPU memory when processing fewer frames. On the other hand, AURORALONG also achieves a significantly faster inference speed. In practice, when compared with InternVL-1.5 2B [18], AURORALONG has a 34X advantage in GPU memory consumption when processing videos with 1,024 sampled frames and achieves an 8X improvement in inference speed when processing one-minute long videos at 1fps. Note that for the purpose of a fair comparison, we do not adopt kernel-level optimization techniques such as FlashAttention [21, 23] when deploying InternVL-1.5 2B [18].

4.3. Ablation Study

4.3.1. Token Merging Ratio

As a core strategy of AURORALONG, token merging plays a significant role in reducing the number of visual tokens. In this section, we further study how video understanding capability is influenced by token merging ratio across multiple tasks. Following AuroraCap [12], we report the performance percentage between the highest and lowest values on the entire performance curve and identify the minimum retention thresholds for achieving 90% and 80% of the peak performance. As shown in Figure 4, for most tasks, AURO-

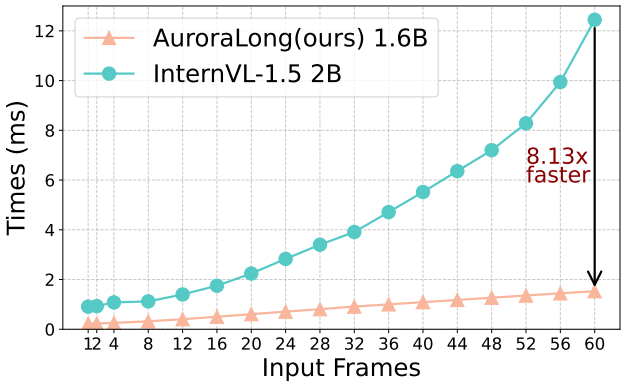


Figure 3. AURORALONG outperforms the SoTA transfomer models while requires less computation and provides lower latency.

RALONG maintain satisfactory performance even with only 0.2 of visual token kept ratio. We further gather the visualization of token merging ratio on all tested video understanding tasks in Figure 5.

Interestingly, as illustrated in Figure 5, the performance of AURORALONG remains relatively stable even at low token retention levels on question answering tasks like ANet [10] and MovieChat-1k [71] while reaching its peak at a 0.1 token kept ratio on captioning tasks such as VATEX [83] and VDC [12], contrasting with [12] whose performance generally declines with fewer visual tokens across most benchmarks and reaches a peak performance when token kept ratio is higher than or equal to 0.5. Refering to [16], we attribute this phenomenon to overfitting as the the RWKV model's recurrent state being overparameterized for the relatively short visual context length per frame in training, which is less than 60 tokens when token merge ratio is set to 0.1. Despite the overfitting tendency in spatial dimension, AURORALONG generalizes well in temporal dimension, handling well long videos up to 10 minutes long at zero-shot scenarios. More calculation details and the visualization results can be found in the Appendix.

4.3.2. Input Token Order

The recurrent mechanism of RWKV [64] omits positional encoding, naturally retaining sequential order information. However, the token merging process disrupts this original sequence order. Therefore, we investigate how organizing the order of merged tokens impacts performance in video understanding. In each merging operation, we merge the two most similar tokens and record the size of the merged token, i.e. total number of original tokens contained in each merged token. Before feeding the merged visual tokens into the RWKV LLM backbone, we consider three sorting strategies: no sorting (random order), sorting tokens in ascending order by size, and sorting tokens in descending order by size. Table 4 indicates that sorting the merged tokens in an ascending manner brings the best performance.

Table 4. Ablation on input order for merged visual tokens within a frame, where descending order suggests tokens merged by most original tokens comes first and 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 [95]	VATEX [83]	VDC [12]	MovieChat-1K [71]
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**

4.3.3. Training Strategy

In this section, we explore the alternative training strategies for the language stage of AURORALONG. For a fair comparison, we use the same training datasets across all settings and maintain consistent hyper-parameters. The following training settings are explored:

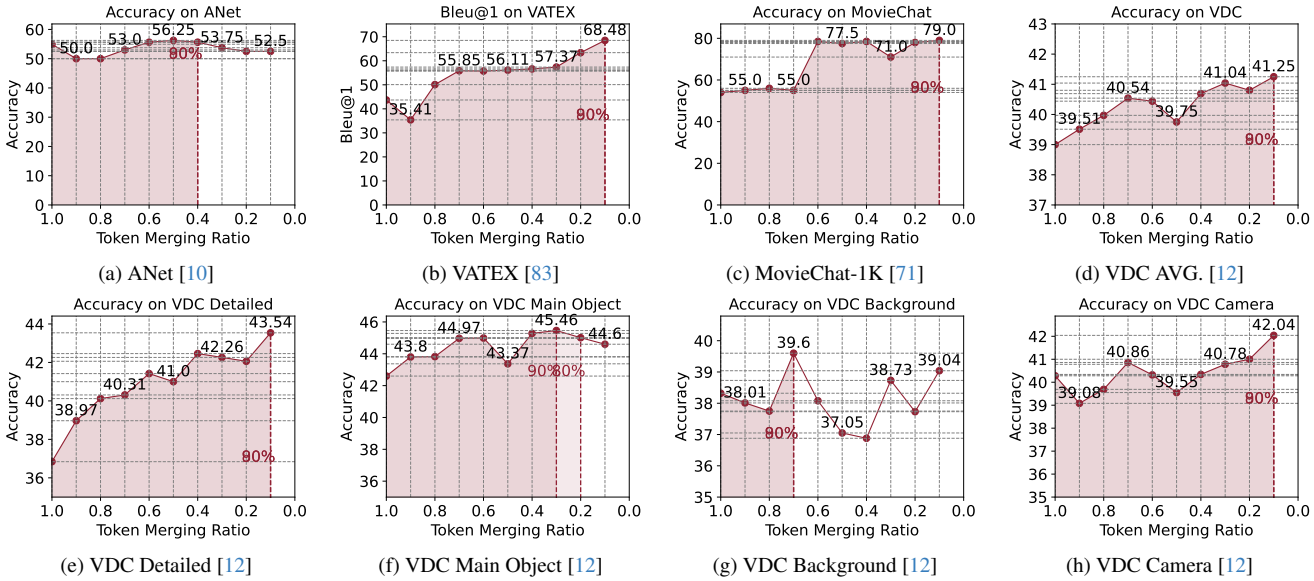
- **Setting** $\mathbb{A}$: Do not apply token merge to single image samples. For video and multi-image samples mostly ranging from 8 to 12 images, apply tome merge with a token kept ratio of 0.1. The purpose of this setting is to keep number of visual tokens passed to LLM backbone roughly the same, providing a smooth transition to multi-frame training in the temperal dimension.
- **Setting** $\mathbb{B}$: Throughout the entire language stage training, always apply token merge with a token kept ratio of 0.1. Inspired by the high masking ratio in Masked Autoencoders [35] , the motivation of this training scheme is to enchance AURORALONG's visual modelling by forcing it to capture fine-grained visual details from few visual tokens per single-image training sample, then transit to multi-frame training by utilizing the temporal generalization capability of the RWKV LLM backbone.

We implement these two training strategies, track the training costs in A800 hours, and evaluate on various video understanding tasks. As shown in Figure 6, training with setting $\mathbb{A}$ brings an extra 50% training time overhead and leads to performance degradation across benchmarks. Therefore, we choose Setting $\mathbb{B}$ as the final training strategy.

5. Limitation

Although AURORALONG has demonstrated impressive abilities in video understanding, it is still an early stage prototype and has some limitations, including: 1) Limited multiple-choice question answering: AURORALONG's performance is hindered by the small size of the pretrained RWKV [63, 64] model, which affects its understanding of complex multiple-choice questions. 2) Challenges in specific domains: Despite showing competitive performance

CVPR
#9135



(a) ANet [10] (b) VATEX [83] (c) MovieChat-1K [71] (d) VDC AVG. [12]

(e) VDC Detailed [12] (f) VDC Main Object [12] (g) VDC Background [12] (h) VDC Camera [12]

Figure 4. Ablation study of token merging in short video question answering on ANet [10], short video sparse captioning on VATEX [83], short video dense captioning on VDC [12], and long video question answering on MovieChat-1K [71]. We find that token merging significantly reduces the number of tokens while maintaining minimal performance drop, and even shows improvement in some tasks. We highlight the token merging ratio when achieving 90% and 80% performance with the dash line and filled area.

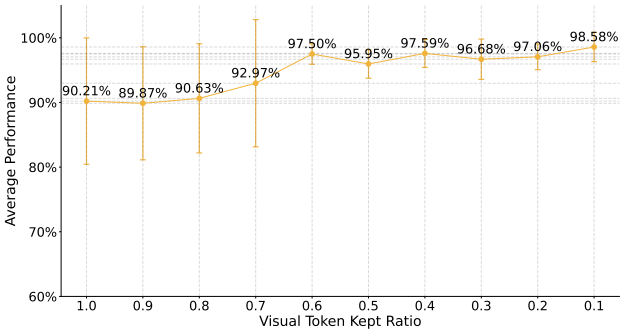


Figure 5. Visualization of token merging ratio on various video understanding tasks. The solid points indicate the average performance and the bounding bars the performance variability across various tasks. All metrics considered here are of percentage scale.

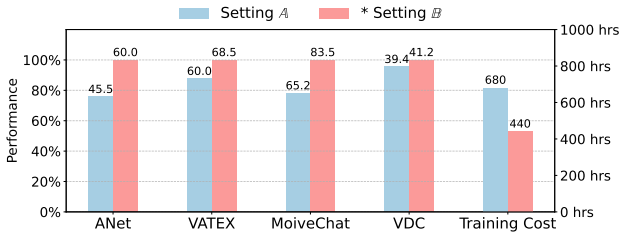


Figure 6. Comparison between different training strategy in Language stage. We take Accuracy for Question-Answering tasks and CIDEr for captioning tasks as the evaluation metric and present the performance percentage. We choose Setting $\mathbb{B}$ as the final training strategy as shown with *. The number shows the maximum value for each benchmark.

on academic datasets, AURORALONG has limited capacity to address problems in certain areas. In the future, we will study how to utilize higher-quality training data to further improve the performance of AURORALONG.

6. Conclusion

In this paper, we introduce AURORALONG, an efficient video understanding model that leverages the linear RNN model RWKV [63] as the language component. By employing a token merging strategy, we significantly reduce computational overhead without compromising performance and overcome overfitting on the training context length in linear attention variant models. We conduct extensive experiments on both short and long video understanding benchmarks, achieving improved performance with more input frames compared to advanced vision language models (VLMs) with larger parameters. Additionally, we carry out ablation studies to evaluate the effectiveness of the token merging ratio and the token reordering strategy we propose. The results validate the effectiveness of our proposed model and demonstrate that there is still room for improvement in applying linear RNNs to VLMs. We hope this work can serve a strong baseline in hydrid architecture for video understanding and facilitate further research in the field of non-transformer long video VLMs.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 2

[2] 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. 2

[3] Kirolos Ataallah, Chenhui Gou, Eslam Abdelrahman, Khushbu Pahwa, Jian Ding, and Mohamed Elhoseiny. Infinibench: A comprehensive benchmark for large multimodal models in very long video understanding. *arXiv preprint arXiv:2406.19875*, 2024. 2

[4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 5

[5] 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

[6] Rohit Bharadwaj, Hanan Gani, Muzammal Naseer, Fahad Shahbaz Khan, and Salman Khan. Vane-bench: Video anomaly evaluation benchmark for conversational lmms. *arXiv preprint arXiv:2406.10326*, 2024. 2

[7] Kaichen Zhang* Fanyi Pu* Xinrun Du Yuhao Dong Haotian Liu Yuanhan Zhang Ge Zhang Chunyuan Li Bo Li*, Peiyuan Zhang* and Ziwei Liu. Lmms-eval: Accelerating the development of large multimoal models, 2024. 5

[8] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. In *The Eleventh International Conference on Learning Representations*, 2022. 3

[9] 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. 2, 3

[10] 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. 5, 7, 8

[11] Mu Cai, Jianwei Yang, Jianfeng Gao, and Yong Jae Lee. Matryoshka multimodal models. *arXiv preprint arXiv:2405.17430*, 2024. 1

[12] 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, 4, 5, 6, 7, 8

[13] 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. 2, 5, 6

[14] Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. *arXiv preprint arXiv:2403.06764*, 2024. 2

[15] Ling-Hao Chen, Shunlin Lu, Ailing Zeng, Hao Zhang, Benyou Wang, Ruimao Zhang, and Lei Zhang. Motionllm: Understanding human behaviors from human motions and videos. *arXiv preprint arXiv:2405.20340*, 2024. 2

[16] 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, 3, 7

[17] 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. 6

[18] 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. 6

[19] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 1, 5, 6

[20] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023. 5

[21] Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*, 2024. 6

[22] 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, 2, 4

[23] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 6

[24] Soham De, Samuel L Smith, Anushan Fernando, Aleksandar Botev, George Cristian-Muraru, Albert Gu, Ruba Haroun, Leonard Berrada, Yutian Chen, Srivatsan Srinivasan, et al. Griffin: Mixing gated linear recurrences with local attention for efficient language models. *arXiv preprint arXiv:2402.19427*, 2024. 2

[25] Rahul Dey and Fathi M Salem. Gate-variants of gated recurrent unit (gru) neural networks. In *2017 IEEE 60th*

international midwest symposium on circuits and systems (MWSCAS), pages 1597–1600. IEEE, 2017. 2

[26] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 3, 5

[27] Yifan Du, Kun Zhou, Yuqi Huo, Yifan Li, Wayne Xin Zhao, Haoyu Lu, Zijia Zhao, Bingning Wang, Weipeng Chen, and Ji-Rong Wen. Towards event-oriented long video understanding. *arXiv preprint arXiv:2406.14129*, 2024. 2

[28] Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A memory-augmented multimodal agent for video understanding. In *European Conference on Computer Vision*, pages 75–92. Springer, 2025. 2

[29] Jiajun Fei, Dian Li, Zhidong Deng, Zekun Wang, Gang Liu, and Hui Wang. Video-ccam: Enhancing video-language understanding with causal cross-attention masks for short and long videos. *arXiv preprint arXiv:2408.14023*, 2024. 5

[30] Leo Feng, Frederick Tung, Mohamed Osama Ahmed, Yoshua Bengio, and Hossein Hajimirsadegh. Were rnns all we needed? *arXiv preprint arXiv:2410.01201*, 2024. 2

[31] Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024. 2

[32] Alex Graves and Alex Graves. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pages 37–45, 2012. 2

[33] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 1, 2, 4

[34] Mingfei Han, Linjie Yang, Xiaojun Chang, and Heng Wang. Shot2story20k: A new benchmark for comprehensive understanding of multi-shot videos. *arXiv preprint arXiv:2312.10300*, 2023. 2

[35] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 7

[36] 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. 2, 4

[37] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 5

[38] 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. 2

[39] Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13700–13710, 2024. 1, 2, 5

[40] 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. 2

[41] Guolin Ke, Di He, and Tie-Yan Liu. Rethinking positional encoding in language pre-training. *arXiv preprint arXiv:2006.15595*, 2020. 4

[42] Seon Ho Lee, Jue Wang, Zhikang Zhang, David Fan, and Xinyu Arthur Li. Video token merging for long-form video understanding. 2024. 2

[43] 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. 5

[44] Dongxu Li, Yudong Liu, Haoning Wu, Yue Wang, Zhiqi Shen, Bowen Qu, Xinyao Niu, Guoyin Wang, Bei Chen, and Junnan Li. Aria: An open multimodal native mixture-of-experts model. *arXiv preprint arXiv:2410.05993*, 2024. 2

[45] Junyan Li, Delin Chen, Tianle Cai, Peihao Chen, Yining Hong, Zhenfang Chen, Yikang Shen, and Chuang Gan. Flexattention for efficient high-resolution vision-language models. *arXiv preprint arXiv:2407.20228*, 2024. 1

[46] 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. 5, 6

[47] 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. 2, 4

[48] Shuai Li, Wanqing Li, Chris Cook, Ce Zhu, and Yanbo Gao. Independently recurrent neural network (indrnn): Building a longer and deeper rnn. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5457–5466, 2018. 2

[49] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. *arXiv preprint arXiv:2311.17043*, 2023. 2

[50] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pages 323–340. Springer, 2025. 5, 6

[51] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 5, 6

[52] 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. 6

[53] 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. 3

[54] 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. 6

[55] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 4

[56] Ye Liu, Zongyang Ma, Zhongang Qi, Yang Wu, Ying Shan, and Chang Wen Chen. Et bench: Towards open-ended event-level video-language understanding. *arXiv preprint arXiv:2409.18111*, 2024. 2

[57] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 4

[58] Hui Lu, Albert Ali Salah, and Ronald Poppe. Videomambapro: A leap forward for mamba in video understanding. *arXiv preprint arXiv:2406.19006*, 2024. 2, 4

[59] Fan Ma, Xiaojie Jin, Heng Wang, Yuchen Xian, Jiashi Feng, and Yi Yang. Vista-llama: Reliable video narrator via equal distance to visual tokens. *arXiv preprint arXiv:2312.08870*, 2023. 2

[60] 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. 2

[61] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 2, 5, 6

[62] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023. 2

[63] 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, 2, 4, 7, 8

[64] Bo Peng, Daniel Goldstein, Quentin Anthony, Alon Albalak, Eric Alcaide, Stella Biderman, Eugene Cheah, Teddy Ferdinan, Haowen Hou, Przemysław Kazienko, et al. Eagle and finch: Rwkv with matrix-valued states and dynamic recurrence. *arXiv preprint arXiv:2404.05892*, 2024. 1, 2, 3, 4, 5, 7

[65] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14313–14323, 2024. 1, 2

[66] Melissa Roemmele and Andrew S Gordon. Automated assistance for creative writing with an rnn language model. In *Companion Proceedings of the 23rd International Conference on Intelligent User Interfaces*, pages 1–2, 2018. 2

[67] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997. 2

[68] 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

[69] Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Shirui Pan, and Chengqi Zhang. Disan: Directional self-attention network for rnn/cnn-free language understanding. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 2

[70] Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*, 2024. 2

[71] 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, 2, 5, 6, 7, 8

[72] Enxin Song, Wenhao Chai, Tian Ye, Jenq-Neng Hwang, Xi Li, and Gaoang Wang. Moviechat+: Question-aware sparse memory for long video question answering. *arXiv preprint arXiv:2404.17176*, 2024. 1, 2

[73] Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2021. 3, 4, 6

[74] Reuben Tan, Ximeng Sun, Ping Hu, Jui-hsien Wang, Hanieh Deilamsalehy, Bryan A Plummer, Bryan Russell, and Kate Saenko. Koala: Key frame-conditioned long video-llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13581–13591, 2024. 2

[75] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 5

[76] Qwen team. Qwen2-vl. 2024. 1

[77] 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. 1, 2, 4

[78] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 5

[79] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 4

[80] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 3

[81] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023. 3

[82] Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Shiyu Huang, Bin Xu, Yuxiao Dong, Ming Ding, et al. Lvbench: An extreme long video understanding benchmark. *arXiv preprint arXiv:2406.08035*, 2024. 2

[83] 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. 5, 7, 8

[84] Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. *arXiv preprint arXiv:2403.10517*, 2024. 2

[85] Yuetian Weng, Mingfei Han, Haoyu He, Xiaojun Chang, and Bohan Zhuang. Longvlm: Efficient long video understanding via large language models. In *European Conference on Computer Vision*, pages 453–470. Springer, 2025. 4

[86] 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. 2

[87] Liang Xu, Shaoyang Hua, Zili Lin, Yifan Liu, Feipeng Ma, Yichao Yan, Xin Jin, Xiaokang Yang, and Wenjun Zeng. Motionbank: A large-scale video motion benchmark with disentangled rule-based annotations. *arXiv preprint arXiv:2410.13790*, 2024. 2

[88] Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994*, 2024. 2

[89] Fuzhao Xue, Yukang Chen, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, et al. Longvila: Scaling long-context visual language models for long videos. *arXiv preprint arXiv:2408.10188*, 2024. 2

[90] Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, et al. xgen-mm (blip-3): A family of open large multimodal models. *arXiv preprint arXiv:2408.08872*, 2024. 5

[91] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. 1, 2

[92] Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention transformers with hardware-efficient training. *arXiv preprint arXiv:2312.06635*, 2023. 2

[93] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1, 2023. 5

[94] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 6

[95] 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. 7

[96] Zihao Yue, Yepeng Zhang, Ziheng Wang, and Qin Jin. Movie101v2: Improved movie narration benchmark. *arXiv preprint arXiv:2404.13370*, 2024. 2

[97] Xiangyu Zeng, Kunchang Li, Chenting Wang, Xinhao Li, Tianxiang Jiang, Ziang Yan, Songze Li, Yansong Shi, Zhengrong Yue, Yi Wang, et al. Timesuite: Improving mllms for long video understanding via grounded tuning. *arXiv preprint arXiv:2410.19702*, 2024. 2

[98] 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. 3, 5

[99] Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple llm framework for long-range video question-answering. *arXiv preprint arXiv:2312.17235*, 2023. 2

[100] Chaoyi Zhang, Kevin Lin, Zhengyuan Yang, Jianfeng Wang, Linjie Li, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. Mm-narrator: Narrating long-form videos with multimodal in-context learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13647–13657, 2024. 2

[101] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 2

[102] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Reality check on the evaluation of large multimodal models, 2024. 5

[103] Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, et al. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*, 2024. 2

[104] 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. 5

[105] 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. 5, 6

[106] Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023. 2

[107] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, 2024. 5

[108] Yu Zhang, Songlin Yang, Ruijie Zhu, Yue Zhang, Leyang Cui, Yiqiao Wang, Bolun Wang, Freda Shi, Bailin Wang, Wei Bi, et al. Gated slot attention for efficient linear-time sequence modeling. *arXiv preprint arXiv:2409.07146*, 2024. 2

[109] Yue Zhao, Long Zhao, Xingyi Zhou, Jialin Wu, Chun-Te Chu, Hui Miao, Florian Schroff, Hartwig Adam, Ting Liu, Boqing Gong, et al. Distilling vision-language models on millions of videos. *arXiv preprint arXiv:2401.06129*, 2024. 1

[110] 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. 2, 5

[111] Xingyi Zhou, Anurag Arnab, Shyamal Buch, Shen Yan, Austin Myers, Xuehan Xiong, Arsha Nagrani, and Cordelia Schmid. Streaming dense video captioning. *arXiv preprint arXiv:2404.01297*, 2024. 1, 2