# Bringing RNNs Back to Efficient Open-Ended Video Understanding

## Supplementary Material

The supplementary material is structured as follows:

- The details and visualization examples of token merging in Section A.
- The training data for LongVidRWKV in Section B.
- The detailed evaluation of LongVidRWKV in Section C.
- More ablation studies for LongVidRWKV in Section D.
- Case studies among several long videos in Section E.

## A. Token Merging

Since the RWKV [50] LLMs are trained on a context length of merely 4,096 tokens, we adopt Token Merge (ToMe) [4] to reduce the number of visual tokens passed to RWKV. By combining similar visual tokens in the Vision Transformer [19], 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* [5] is formulated by

$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{\mathbf{d}}} + \log \mathbf{s}\right) \tag{1}$$

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

The number of tokens $r$ merged in each layer of the Vision Transformer is calculated by the following formula:

$$r = (1 - \textit{ratio}) \cdot (W \cdot H / P^2) / L \tag{2}$$

where $W$, $H$ is the width and the height of the input image, $P$ is the patch size, $L$ is the number of layers in the transformer, $ratio$ is the proportion of tokens to be kept. The merging is evenly distributed across all layers of the SigLIP Vision Transformer [1]. We visualize the gradual token merging process within different Vision Transformer layers in Figure A1 using image samples from WIT [2], LAION Art [3], MSCOCO [4], and BLIP3-KALE [5] respectively.

| # Token | **Please describe the image in detail. (Figure A1 Row 3)** |
|---|---|
| 576 | The image depicts a man in the midst of a soccer game. He is dressed in a green shirt and black shorts, and he is wearing a pair of black socks. The man is holding a soccer ball in his right hand, which is positioned close to his body, suggesting he is either preparing to kick the ball or has just kicked it. He is also wearing a black wristband on his left wrist. The background shows a fence with a chain-link fence, and beyond the fence, there are trees and a clear sky. The man is standing on a grassy field, and there are other players visible in the background, indicating that this is a soccer practice or game. The image does not provide any specific context or background story. |
| 408 | The image depicts a man standing on a grassy field, focused on a soccer ball in his hands. He is dressed in a green shirt and black shorts, and his attire suggests he is a soccer player. The man is wearing a black wristband on his left wrist, and his right arm is extended, holding the ball. He is wearing black socks and black shoes, and his socks are visible. The background is a lush green field, and there are several other players in the distance, suggesting this is a soccer practice or training session. The sky is clear, indicating it might be a sunny day. There are no other people visible in the image, and the setting appears to be a private or semi-private field. |
| 288 | The image depicts a man in a green shirt and black shorts running on a grassy field. He is wearing a green soccer jersey with the number "27" visible, and he is holding a soccer ball in his right hand. The man is also wearing black shorts and socks, and his shoes are black with white laces. He is wearing black knee pads and black socks. In the background, there are several other people, including a man in a green shirt and black shorts, and another man in a green shirt and black shorts, both of whom are running. The field is surrounded by a chain-link fence, and there are trees in the background. The sky is clear, suggesting it is a sunny day. The image captures a moment of leisure and physical activity. |
| 72 | The image depicts a man in a green jersey with the num- |

[1] HuggingFace Model: google/siglip-large-patch16-384
[2] HuggingFace Dataset: google/wit
[3] HuggingFace Dataset: laion/laion-art
[4] 5K test set split from: http://cs.stanford.edu/people/karpathy/deepimagesent/caption_datasets.zip
[5] HuggingFace Dataset: Salesforce/blip3-kale


Figure A1. Visualization of token merging within a 24-layer Vision Transformer. From left to right, we show the preprocessed image, the merged image patches after the 6-th layer, the 12-th layer, the 18-th layer and the 24-th layer, which are represented by 576, 451, 325, 199, and 73 visual tokens respectively. From top to bottom, the image IDs are WIT:COCO-train2014-000000247906, LAION:2331508, COCO:sa-393200, KALE:71789111.

ber 19 on it, standing on a grassy field. He is wearing a pair of green socks and is holding a soccer ball in his right hand. The background is a blurred field with a chain-link fence in the background, and there are several other players in the background, some of whom are wearing jerseys with numbers. The sky is clear, suggesting it might be a sunny day. The man is looking towards the camera with a focused expression.

48 The image depicts a young man in a green sports uniform, running on a grassy field. He is wearing a black headband, black shorts, and black socks. The man is holding a soccer ball in his right hand, and his left hand is raised in a gesture that suggests he is either preparing to catch the ball or is in the process of throwing it. The background is a blurred green field with a few trees and a fence in the distance. The sky is clear, indicating it might be a sunny day. There are no other people or objects in the immediate vicinity of the man, giving the impression of a solitary moment.

**# Token** **What is on the cover of the book? (Figure A1 Row 4)**

576 The cover of the book features a cartoon-style illustration of a young boy sitting on the grass. He is dressed in a red shirt and a red and white striped sweater, and he is holding a green pencil in his right hand. The boy is smiling and appears to be in a relaxed and happy mood. The background of the cover is a light green color, and there is a white border around the image. The title of the book is prominently displayed in the upper left corner, and the subtitle "Spring Math Review" is written below

Table B1. Training hyper-parameters for LongVidRWKV

| 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% |

it in a smaller font. The overall design of the cover suggests a playful and educational theme, likely aimed at engaging young learners in the subject of mathematics.

408 The cover of the book features a cartoon-styled illustration of a young boy sitting on a green blanket. He is smiling and looking towards the camera, giving a cheerful and inviting impression. The boy is wearing a red shirt and has blonde hair. The background of the cover is a light blue color, which contrasts with the vibrant colors of the boy's attire. The title of the book is prominently displayed in the upper left corner, and the subtitle "Spring Math Number 1-20" is located at the bottom of the cover. The overall design of the cover suggests a playful and engaging approach to teaching basic number concepts to young learners.

288 The cover of the book features a cartoon-style illustration of a young boy sitting on a green blanket. He is smiling and appears to be in a relaxed and happy mood. The boy is wearing a red shirt and has a green backpack on his lap. The background of the cover is a light blue color, and there are some white dots scattered around the image. The overall impression is that of a cheerful and engaging educational activity for young learners.

72 The cover of the book features a cartoon character.

48 The cover of the book features a cartoon character.

## B. Training Recipe

LongVidRWKV is trained on a diverse collection of datasets to enhance its performance in multimodal tasks, particularly in video captioning and question-answering. The utilized datasets during training is shown in Table B2, Table B3 and Table B4.

During training, we only compute cross-entropy loss for auto-regressive text generation. For all training stages, we use the AdamW [44] optimizer with a cosine decay schedule, setting the optimizer hyperparameters $\beta_1$ and $\beta_2$ 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 detailed in Table B1. 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 [19, 70] layer. Our model is trained on 8 NVIDIA A800 GPUs.

**Pretraining stage.** Similar to LLaVA [43], we first learn the alignment between visual features from the vision encoder and the word embedding space of RWKV [50]. 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 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 [7] and [2].

**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 so that we can feed as much input frames to LongVidRWKV as possible while further enhancing the 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.

Table B2. Summary of datasets used for training LongVidRWKV in Pretraining stage.

| Task | # Sample | Dataset |
|---|---|---|
| Image Captioning | 1.3M | LAION-CC-SBU-595K [43], ShareGPT4V [9], ALLaVA-Caption-LAION-4V [8], ALLaVA-Caption-VFLAN-4V [8], DenseFusion [37] |

Table B3. Summary of datasets used for training LongVidRWKV in Vision stage. For classification, Reasoning, VQA, and Generation tasks, we adopt the dataset processed by $M^3$IT [34] to fit the training objective of language models.

| Task | # Sample | Dataset |
|---|---|---|
| Captioning | 1,925K | ShareGPT4V-PT [9], TextCaps [54], Image-Paragraph-Captioning [30] |
| Object-centric | 438K | COST [26], ChatterBox [60], V* [66] |
| Classification | 238K | COCO-GOI [40], COCO-Text [62], ImageNet [51], COCO-ITM [40], e-SNLI-VE [29], Mocheg [67], IQA [20] |
| Reasoning | 100K | CLEVR [27], NLVR [57], VCR [69], VisualMRC [58], Winoground [59] |
| VQA | 3,518K | VQA v2 [23], Shapes VQA [1], DocVQA [48], OK-VQA [47], Text-VQA [55], OCR-VQA [49], A-OK-VQA [52], ScienceQA [45] ST-VQA [3], ViQuAE [31], LLaVA-OneVision [32] |
| Generation | 145K | Visual Storytelling [25], Visual Dialog [18], Multi30k [21] |
| Chinese | 193K | COCO-Caption CN [36], Flickr-8k-Caption CN [35], multimodal Chat [75], FM-IQA [22], ChineseFoodNet [11] |
| Total | 6.6M | For all datasets, we uniformly sample without duplication. |

Table B4. Summary of datasets used for training LongVidRWKV in Language stage.

| Task | # Sample | Dataset |
|---|---|---|
| Image Captioning | 1,779K | ShareGPT4V [9], ALLaVA-Caption-LAION-4V [8], ALLaVA-Caption-VFLAN-4V [8], DenseFusion [37], FaceCaption [17] |
| Video Captionin | 1,659K | MiraData [28], LLaVA-Hound [73], ShareGPT4Video [10] |
| Image Instruction | 9,742K | LVIS-Instruct4V [63], ALLaVA-Instruct-LAION-4V [8], ALLaVA-Instruct-VFLAN-4V [8], Cambrian [61], LLaVA-Mix-665K [41], M4-Instruct [42] |
| Video Instruction | 446K | LLaVA-Hound [73], ShareGPT4Video [10], LLaVA-Video-178k [74] |
| Language-only | 143K | Evol-Intruct-GPT4-Turbo-143K [8] |
| Total | 15.4M | We duplicate video captioning and instruction datasets in training. |

## C. Evaluation Results

In this section we report the detailed evaluation results of LongVidRWKV on MLVU [76] and MVBench [33] in Table B5 and Table B6. We find that although LongVidRWKV has only 2B parameters, it achieves comparable or even better performance when compared with industry leading proprietary models like GPT4-o and open-weight SOTA video LLMs like VILA-1.5 and InternVL-2. These results demonstrate LongVidRWKV's strong capability in handling long video inputs. Note that we do not conduct LLM context extension as in [72]. Although only trained on videos less than 60 frames, LongVidRWKV generalizes well on long video tasks.

## D. Ablation Studies

As a core strategy of LongVidRWKV, token merging plays a significant role in reducing the number of visual tokens. We conduct extensive ablation studies to explore the impact of the token kept ratio and the merged tokens order in terms of performance across multiple tasks including video captioning, and video question answering as shown in Fig-

Table B5. Comparison of LongVidRWKV with SOTA methods on MLVU [76] whose average video length is about 12 minutes. The best result is highlighted in bold, and the second best is underlined. We find that even with only 2B parameters, LongVidRWKV outperforms models up to 38X larger across various long video understanding tasks.

| Models | Input | CTX | Size | MLVU | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | AVG | AR | ER | AO | AC | TR | NQA | PQA |
| *Proprietary Models* | | | | | | | | | | | |
| GPT4-o | 0.5fps | 128k | - | **54.5** | **68.8** | 47.8 | 46.2 | 35.0 | 83.7 | 42.9 | 57.1 |
| GPT4-Turbo | 16frm | 128k | - | 43.8 | 61.5 | 41.5 | 22.9 | 6.7 | **85.7** | 40.0 | 48.0 |
| Qwen-VL-Max | 10frm | 32k | - | 34.4 | 53.8 | 26.4 | 20.0 | 11.7 | 75.8 | 15.0 | 38.0 |
| Claude-3-Opus | 16frm | 128k | - | 21.8 | 30.8 | 17.0 | 10.0 | 6.7 | 53.8 | 14.0 | 20.0 |
| *Open-Source Video LMMs* | | | | | | | | | | | |
| LLAMA-VID [38] | 1 fps | 4k | 7B | 18.1 | 23.1 | 11.3 | 18.6 | 15.0 | 20.9 | 21.7 | 16.0 |
| mPLUG-Owl-V [68] | 16 frm | 4k | 7B | 16.7 | 15.4 | 13.2 | 14.3 | 20.0 | 25.3 | 6.7 | 22.0 |
| Video-ChatGPT [46] | 16 frm | 2k | 7B | 21.2 | 17.9 | 32.1 | 17.1 | 13.3 | 17.6 | 28.3 | 22.0 |
| MovieChat [56] | 2048 frm | 4k | 7B | 16.5 | 10.3 | 15.1 | 17.1 | 15.0 | 18.7 | 23.3 | 16.0 |
| Video-LLAVA [46] | 8 frm | 4k | 7B | 30.1 | 38.5 | 26.4 | 20.0 | 21.7 | 70.3 | 13.3 | 26.0 |
| LLaVA-1.6 [42] | 16 frm | 8k | 7B | 27.1 | 17.9 | 26.4 | 21.4 | 16.7 | 63.7 | 13.3 | 30.0 |
| LongVA [72] | 256 frm | 224k | 7B | 42.1 | 41.0 | 39.6 | 17.1 | 23.3 | 81.3 | 46.7 | 46.0 |
| VideoChat2 [33] | 16 frm | 8k | 7B | 30.9 | 30.8 | 28.3 | 17.1 | 23.3 | 72.5 | 18.3 | 26.0 |
| Video-XL | 256 frm | 131k | 7B | 46.3 | 28.2 | 41.5 | **48.6** | 31.7 | 78.0 | **50.0** | 46.0 |
| ShareGPT4Video [10] | 16 frm | 8k | 8B | 34.2 | 25.6 | 45.3 | 17.1 | 8.3 | 73.6 | 31.7 | 38.0 |
| Video-LLAMA-2 [16] | 16 frm | 131k | 13B | 18.8 | 12.8 | 17.0 | 15.7 | 8.3 | 52.7 | 13.3 | 12.0 |
| InternVL-1.5 [13] | 16 frm | 4k | 26B | 37.9 | 51.3 | 24.5 | 14.3 | 13.3 | 80.2 | 40.0 | 42.0 |
| VILA-1.5 [39] | 14 frm | 4k | 40B | 46.2 | 56.4 | 35.8 | 34.3 | 11.7 | 84.7 | 38.3 | **62.0** |
| LongVidRWKV **(ours)** | 48 frm | 4k | 2B | 52.7 | 59.5 | **57.1** | 33.2 | **42.9** | 69.0 | 45.2 | 61.9 |

Table B6. Results on MVBench [33] 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 |
| *Proprietary Models* | | | | | | | | | | | | | | | | | | | |
| GPT4-V | - | 43.7 | 39.0 | 40.5 | 63.5 | 55.5 | 52.0 | 11.0 | 31.0 | 46.5 | 22.5 | 12.0 | 12.0 | 18.5 | 59.0 | 29.5 | 83.5 | 45.0 | 73.5 |
| Gemini Pro | - | 37.7 | 37.7 | 40.0 | 41.8 | 35.4 | 38.7 | 33.7 | 36.4 | 36.2 | 41.5 | 18.0 | 16.5 | 43.5 | 37.5 | 39.8 | 75.4 | 42.3 | 67.1 |
| *Open Weight Models* | | | | | | | | | | | | | | | | | | | |
| Video-LLAMA [71] | 7B | 34.1 | 39.0 | 34.0 | 32.5 | 48.0 | 43.0 | 22.5 | 25.5 | 27.5 | 40.0 | 37.0 | 30.0 | 29.0 | 22.5 | 22.5 | 40.5 | 38.0 | 45.4 |
| mPLUG-Owl-V [68] | 7B | 29.4 | 23.5 | 34.5 | 31.5 | 36.0 | 34.5 | 24.0 | 20.0 | 25.0 | 37.0 | 37.0 | 25.5 | 27.0 | 22.0 | 23.0 | 24.0 | 34.0 | 40.0 |
| Video-ChatGPT [46] | 7B | 32.7 | 26.5 | 30.5 | 39.5 | 54.0 | 31.0 | 20.0 | 26.0 | 23.5 | 33.0 | 35.5 | 29.5 | 22.5 | 25.5 | 23.0 | 28.0 | **40.0** | 48.5 |
| MovieChat [56] | 7B | 33.7 | 28.0 | 42.5 | 42.5 | 39.5 | 36.0 | 26.5 | 29.0 | 33.0 | 32.5 | 32.5 | 28.5 | 31.0 | 37.5 | 27.5 | 32.0 | 35.5 | 39.5 |
| LLAVA-NeXT [42] | 7B | 32.8 | 35.0 | 35.5 | 42.5 | 34.6 | 58.0 | 20.5 | 31.0 | 33.4 | 34.5 | 17.0 | 31.5 | 38.0 | 26.5 | 25.0 | 42.0 | 13.8 | 38.5 |
| LLAMA-VID [38] | 7B | 42.0 | 56.5 | 44.5 | 41.4 | 55.6 | 84.5 | 26.5 | 43.0 | 42.0 | 39.0 | 34.5 | 36.5 | 35.5 | 28.5 | 19.0 | 37.5 | 34.0 | 40.5 |
| VILA-1.5* [39] | 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 |
| LLaVA-OneVision [32] | 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 |
| ShareGPT4Video* [10] | 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 |
| LongVA* [72] | 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 |
| InternVL-1.5* [15] | 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 |
| InternVL2* [14] | 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* [64] | 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 |
| 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** |

ure D2, Figure D3 and Figure D4. We define the performance percentage as the proportion between the highest and lowest values on the entire performance curve. We identify the minimum retention thresholds for achieving 90% and 80% performance. Note that LongVidRWKV focuses on spatial visual token merging, while the temporal features

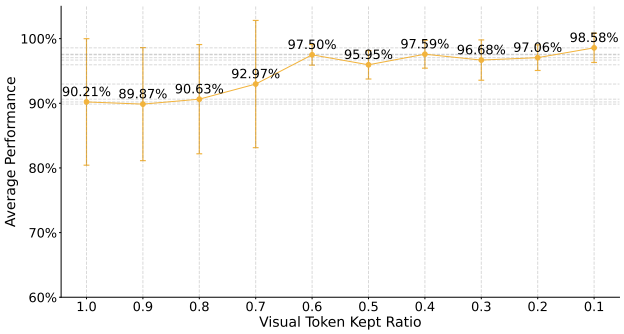



Figure D2. Visualization of token merging ratio on various video understanding tasks. The solid points indicate the average performance and the bounding bars the performance variability across various tasks. All metrics considered here are of percentage scale.

introduce additional complexity to explore the token merging laws. Appendix A shows more calculation details and the visualization results of token merging.

### D.1. Ablation on Training Strategy

In this section, we explore the alternative training strategies for the language stage of LongVidRWKV. 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 [24] , the motivation of this training scheme is to enchance LongVidRWKV'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 D5, 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 our final recipe.

### D.2. Ablation on Token Merging Ratio

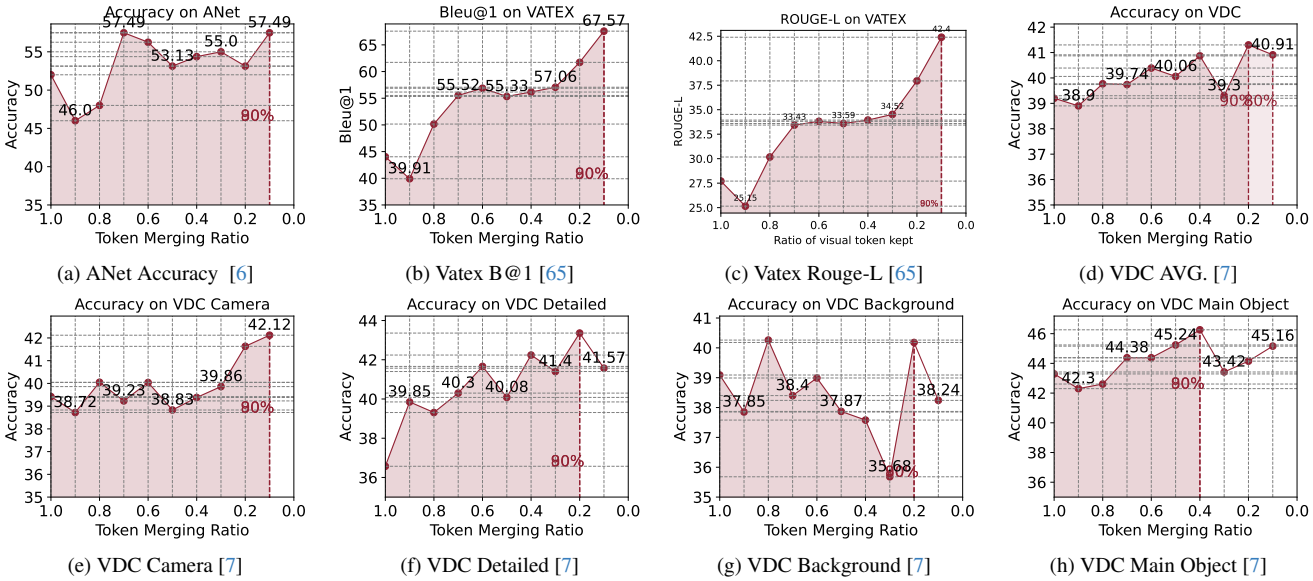
As a core strategy of LongVidRWKV, token merging plays a significant role in reducing the number of visual tokens, bridging the gap between the large number of video tokens and the pretrained 4k context length of RWKV LLMs. In this section, we further study how video understanding capability is influenced by token merging ratio across multiple tasks. 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 D3, for most tasks, LongVidRWKV reaches performance peak even with a visual token kept ratio of only 0.1. We further gather the visualization of token merging ratio on all tested video understanding tasks in the Appendix.

Interestingly, as illustrated in Figure D3, for most video captioning tasks such as VATEX and [65] VDC [7], LongVidRWKV 's seem to perform better at lower token retention levels. This contrasts with most Transformer-based token-reduction methods as in [7] and [77], where performance generally declines when fewer visual tokens are retained per frame and reaches a peak performance when token kept ratio is higher than or equal to 0.5. Referring to [12], 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, LongVidRWKV 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.

## E. Case Study

We conduct extensive case studies of LongVidRWKV on a diverse set of videos to evaluate its performance in detailed video captioning and video question-answering tasks. We compare the captions and answers generated by LongVidRWKV with those produced by other leading models. As shown in the followings, LongVidRWKV is capable of providing excellent detailed captions that accurately capture the content of the videos, and answering questions with significantly fewer hallucinations compared to existing models. Green highlights the correct answer, red indicates that the model is wrong or hallucinating, and yellow represents the repeated parts.

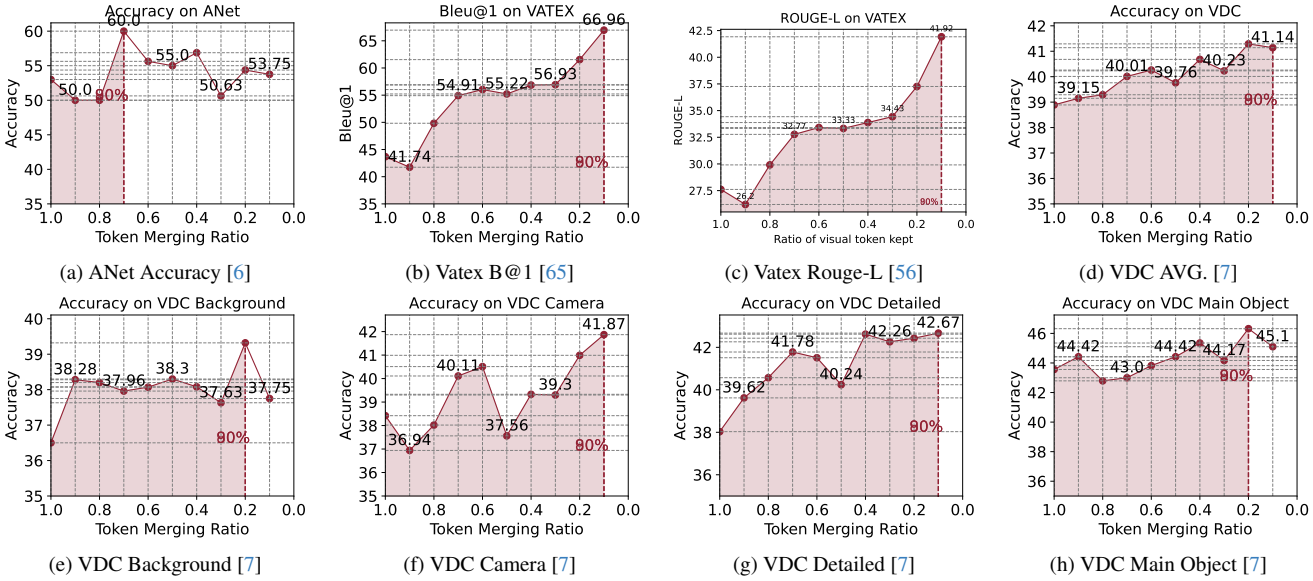
| Method | Answer |
|---|---|
| Figure E6 | *[Question] In this video, how did the protagonist stop the train?* |
| VILA [39] | The protagonist stops the train by shooting webs at it . |
| InternVL2 [15] | The protagonist in this video is Spider-Man , played by actor Andrew Garfield. He stopped the train by using his web-slinging abilities to create |




(a) ANet Accuracy [6] (b) Vatex B@1 [65] (c) Vatex Rouge-L [65] (d) VDC AVG. [7]

(e) VDC Camera [7] (f) VDC Detailed [7] (g) VDC Background [7] (h) VDC Main Object [7]

Figure D3. Ablation study of token merging with no sort in short video question answering on ANet [6], short video sparse captioning on VATEX [65], short video dense captioning on VDC [7], and long video question answering on MovieChat-1K [56]. 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.




(a) ANet Accuracy [6] (b) Vatex B@1 [65] (c) Vatex Rouge-L [56] (d) VDC AVG. [7]

(e) VDC Background [7] (f) VDC Camera [7] (g) VDC Detailed [7] (h) VDC Main Object [7]

Figure D4. Ablation study of token merging with descending order in short video question answering on ANet [6], short video sparse captioning on VATEX [65], short video dense captioning on VDC [7], and long video question answering on MovieChat-1K [56]. 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.

a web between the train cars and the tracks, effectively stopping the train from moving.

LongVU [53] The protagonist in this video is Spider-Man. He stopped the train by using his web-slinging abilities to grab onto the train and prevent it from crashing into the station.

LongVidRWKV The Spider-Man uses his web-shooters to swing onto the train and then onto the train's roof, using his web-slinging abilities to stop the train from

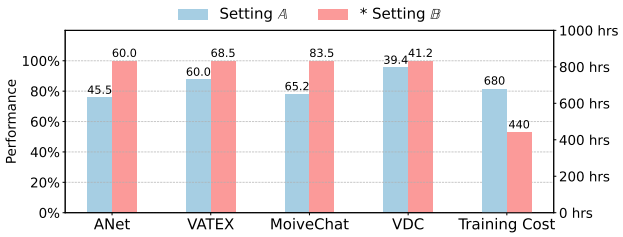



Figure D5. Comparison between different training strategy in Language stage. We take Accuracy for Question-Answering tasks and CIDEr for captioning tasks as the evaluation metric and present the performance percentage. We choose Setting $\mathbb{B}$ as the final training strategy as shown with *. The number shows the maximum value for each benchmark.

falling into the cliff .

| Method | Answer |
|---|---|
| Figure E7 | *[Question] In this video, how did the man ride on the giant monster?* |
| VILA [39] | The man in the video rode on the giant monster by using a rope to climb onto its back . |
| InternVL2 [15] | In the video, the man rode on the giant monster by using a rope to tie around its neck and then sliding down its back . |
| LongVU [53] | The man in the video rode on the giant monster by climbing up a hill . |
| LongVidRWKV | Before the giant monster appeared, the man was seen running and jumping in the desert hill . When riding the giant monster, the man controlled it with his long metal hook . |

| Method | Answer |
|---|---|
| Figure E8 | *[Question] Describe the events in the video shown by these frames in at least three sentences.* |
| VILA [39] | The video shows two men sitting at a table outside, one of them is eating and the other is talking. The man who is eating is wearing a tan shirt and the other man is wearing a yellow jacket. They are both sitting at a picnic table and there is a car parked behind them. |
| InternVL2 [15] | In the video, two men are seated at a white picnic table outside a restaurant, engaged in a conversation. The man dressed in a mustard yellow turtleneck and plaid blazer is holding a piece of paper, while the other man, dressed in a beige sweater, is smoking a cigarette. The table is cluttered with various items, including cups, a bag of chips, and a blue napkin . The conversation appears to be serious in nature, with both men looking intently at each other. |
| LongVU [53] | Two men are sitting at a table in a fast-food restaurant. One man is smoking a cigarette and holding a yellow cup, while the other man is reading a letter. The man with the cigarette is smoking and looking at the camera . The man with the letter is reading it and then puts it down. The man with the cigarette is smoking and looking at the man with the letter. The man with the letter is reading a letter and then puts it down. The man with the cigarette is smoking and looking at the man with the letter. The man with the letter is reading a letter and then puts it down. |
| LongVidRWKV | The video depicts a man sitting at a table in a casual outdoor setting, engaging in a conversation with another individual. The man in the yellow jacket is holding a piece of paper and appears to be explaining or discussing something with the other individual. The setting is outdoors, with a clear view of a suburban street and a parked car, suggesting a casual, relaxed atmosphere. The man in the yellow jacket is seen holding a cup, possibly a beverage, and appears to be in a relaxed state, indicating a casual and informal conversation. The scene is set in a residential area with a suburban house visible in the background, and the weather is clear, suggesting a pleasant day. The video captures the interaction between the two men, focusing on their conversation and the man in the yellow jacket holding the paper, which seems to be a piece of paper with text on it. |

| Method | Answer |
|---|---|
| Figure E9 | *[Question] Describe the events in the video shown by these frames in at least three sentences.* |
| VILA [39] | The video shows a man in a red and blue suit swinging through the city. He is seen fighting a bald man and a woman in a black dress. The man is also seen swinging through the city and fighting a man in a black suit. |
| InternVL2 [15] | The video captures Spider-Man engaging in a series of action-packed scenes. In the initial frame, he is seen flying above a city, with the iconic spider logo on his back. The subsequent images show him swinging between skyscrapers, battling a villain with mechanical tentacles, and rescuing a woman holding a baby amidst chaos. Throughout the video, Spider-Man displays his agility and heroism, navigating through intense action sequences that include explosions and high-stakes rescues, culminating in a dramatic leap from a building at night. |
| LongVU [53] | The video begins with a close-up of a person's hand holding a smartphone, capturing a scene of a cityscape at night with a large, illuminated billboard and a vehicle with flashing red and blue lights. The scene then transitions to a first-person perspective of a character in a red and blue suit, running through a cityscape at night, with a focus on the character's hands and the smartphone they're holding. The perspective shifts |

Figure E6. Video clips from Spider-Man, a superhero story, which tells the tale of a young man named Peter Parker who gains spider-like abilities and uses them to fight villains, protect his city, and navigate the challenges of responsibility and heroism.

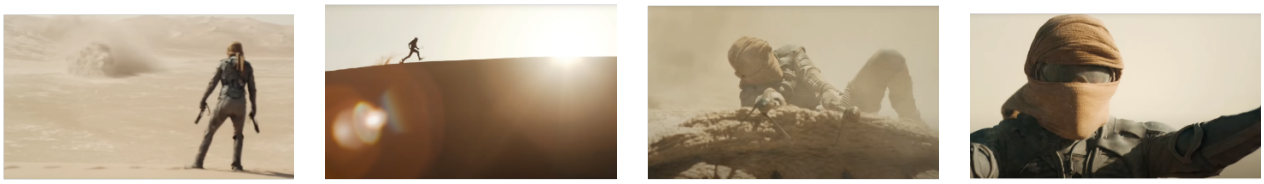


Figure E7. Video clips from Dune, a sci-fi epic, which tells the story of Paul Atreides, a young nobleman who must navigate political intrigue and warfare on the desert planet Arrakis while embracing his destiny to protect its precious resource and lead its people.

| | |
|---|---|
| | to show the character's feet as they leap off a building, with a view of the cityscape below and a large, illuminated billboard in the background. |
| LongVidRWKV | The video showcases a dynamic and visually striking scene set in a bustling cityscape, where a superhero, clad in a striking blue and red costume, is seen performing a series of acrobatic maneuvers against a backdrop of a city skyline. Initially, the superhero is seen in mid-air, executing a series of acrobatic flips and spins, with a large, fiery explosion occurring in the background, suggesting a dramatic or explosive moment. As the video progresses, the superhero continues to perform acrobatic stunts, including a dramatic leap and a flip, while the cityscape remains a constant backdrop, emphasizing the action. The scene is filled with a sense of urgency and drama, highlighted by the presence of a large, fiery explosion and the superhero's intense focus on the task at hand. The video captures the essence of a superhero in action, with the cityscape serving as a backdrop that enhances the dramatic and action-packed nature of the scene. |

## References


[1] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 39–48, 2016. 4

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 3

[3] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4291–4301, 2019. 4

[4] 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. 1

[5] 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

[6] 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. 7

[7] 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. 3, 6, 7

[8] Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harness-

Figure E8. Video clips from Green Book, a drama, which tells the story of an Italian-American bouncer named Tony who becomes the driver for an African-American pianist, Dr. Shirley, on a concert tour in the segregated South, leading to an unexpected bond and mutual understanding.

Figure E9. Video clips from Spider-Man, a superhero story, which tells the tale of a young man named Peter Parker who gains spider-like abilities and uses them to fight villains, protect his city, and navigate the challenges of responsibility and heroism.


ing gpt4v-synthesized data for a lite vision-language model. *arXiv preprint arXiv:2402.11684*, 2024. 4

[9] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023. 4

[10] 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. 4, 5

[11] Xin Chen, Yu Zhu, Hua Zhou, Liang Diao, and Dongyan Wang. Chinesefoodnet: A large-scale image dataset for chinese food recognition. *arXiv preprint arXiv:1705.02743*, 2017. 4

[12] 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. 6

[13] 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. 5

[14] 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. 5

[15] 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. 5, 6, 8

[16] 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. 5

[17] Dawei Dai, YuTang Li, YingGe Liu, Mingming Jia, Zhang YuanHui, and Guoyin Wang. 15m multimodal facial image-text dataset, 2024. 4

[18] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335, 2017. 4

[19] 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. 1, 3

[20] Zhengfang Duanmu, Wentao Liu, Zhongling Wang, and Zhou Wang. Quantifying visual image quality: A bayesian view. *Annual Review of Vision Science*, 7:437–464, 2021. 4

[21] Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30k: Multilingual english-german image descriptions. *arXiv preprint arXiv:1605.00459*, 2016. 4

[22] Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. Are you talking to a machine? dataset and methods for multilingual image question. *Advances in neural information processing systems*, 28, 2015. 4

[23] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 4

[24] 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. 6

[25] Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. Visual storytelling. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 1233–1239, 2016. 4

[26] Jitesh Jain, Jianwei Yang, and Humphrey Shi. Vcoder: Versatile vision encoders for multimodal large language models. *arXiv preprint arXiv:2312.14233*, 2023. 4

[27] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017. 4

[28] Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. Miradata: A large-scale video dataset with long durations and structured captions. *arXiv preprint arXiv:2407.06358*, 2024. 4

[29] Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, Cornelius Emde, Virginie Do, Zeynep Akata, and Thomas Lukasiewicz. e-vil: A dataset and benchmark for natural language explanations in vision-language tasks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1244–1254, 2021. 4

[30] Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. A hierarchical approach for generating descriptive image paragraphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 317–325, 2017. 4

[31] Paul Lerner, Olivier Ferret, Camille Guinaudeau, Hervé Le Borgne, Romaric Besançon, José G Moreno, and Jesús Lovón Melgarejo. Viquae, a dataset for knowledge-based visual question answering about named entities. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3108–3120, 2022. 4

[32] 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. 4, 5

[33] 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, 5

[34] Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, et al. M3it: A large-scale dataset towards multi-modal multilingual instruction tuning. *arXiv preprint arXiv:2306.04387*, 2023. 4

[35] Xirong Li, Weiyu Lan, Jianfeng Dong, and Hailong Liu. Adding chinese captions to images. In *Proceedings of the 2016 ACM on international conference on multimedia retrieval*, pages 271–275, 2016. 4

[36] Xirong Li, Chaoxi Xu, Xiaoxu Wang, Weiyu Lan, Zhengxiong Jia, Gang Yang, and Jieping Xu. Coco-cn for cross-lingual image tagging, captioning, and retrieval. *IEEE Transactions on Multimedia*, 21(9):2347–2360, 2019. 4

[37] Xiaotong Li, Fan Zhang, Haiwen Diao, Yueze Wang, Xinlong Wang, and Ling-Yu Duan. Densefusion-1m: Merging vision experts for comprehensive multimodal perception. *arXiv preprint arXiv:2407.08303*, 2024. 4

[38] 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

[39] 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. 5, 6, 8

[40] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 4

[41] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 4

[42] 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. 4, 5

[43] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 3, 4

[44] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 3

[45] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. 4

[46] 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. 5

[47] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019. 4

[48] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021. 4

[49] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference*

*on document analysis and recognition (ICDAR)*, pages 947–952. IEEE, 2019. 4

[50] 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, 3

[51] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 4

[52] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pages 146–162. Springer, 2022. 4

[53] 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. 7, 8

[54] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 742–758. Springer, 2020. 4

[55] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 4

[56] 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. 5, 7

[57] Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223, 2017. 4

[58] Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. Visualmrc: Machine reading comprehension on document images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13878–13888, 2021. 4

[59] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022. 4

[60] Yunjie Tian, Tianren Ma, Lingxi Xie, Jihao Qiu, Xi Tang, Yuan Zhang, Jianbin Jiao, Qi Tian, and Qixiang Ye. Chatterbox: Multi-round multimodal referring and grounding. *arXiv preprint arXiv:2401.13307*, 2024. 4

[61] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024. 4

[62] Andreas Veit, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge Belongie. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140*, 2016. 4

[63] Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. To see is to believe: Prompting gpt-4v for better visual instruction tuning. *arXiv preprint arXiv:2311.07574*, 2023. 4

[64] 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. 5

[65] 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. 6, 7

[66] Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. *arXiv preprint arXiv:2312.14135*, 2023. 4

[67] Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2733–2743, 2023. 4

[68] 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. 5

[69] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731, 2019. 4

[70] 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

[71] 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. 5

[72] 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. 4, 5

[73] Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander Hauptmann, Yonatan Bisk, et al. Direct preference optimiza-

tion of video large multimodal models from language model reward. *arXiv preprint arXiv:2404.01258*, 2024. 4

[74] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data, 2024. 4

[75] Yinhe Zheng, Guanyi Chen, Xin Liu, and Jian Sun. Mmchat: Multi-modal chat dataset on social media. *arXiv preprint arXiv:2108.07154*, 2021. 4

[76] 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, 5

[77] Orr Zohar, Xiaohan Wang, Yann Dubois, Nikhil Mehta, Tong Xiao, Philippe Hansen-Estruch, Licheng Yu, Xiaofang Wang, Felix Juefei-Xu, Ning Zhang, et al. Apollo: An exploration of video understanding in large multimodal models. *arXiv preprint arXiv:2412.10360*, 2024. 6