

Bringing RNNs Back to Efficient Open-Ended Video Understanding

Supplementary Material

The supplementary material is structured as follows:

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

A. Token Merging Visualization

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

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

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 Transformer¹. We visualize the gradual token merging process within different Vision Transformer layers in Figure A1 using image samples from WIT², LAION Art³, MSCOCO⁴, and BLIP3-KALE⁵ 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 number 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 it in a smaller font. The overall design of the cover sug-

¹HuggingFace Model: [google/siglip-large-patch16-384](https://huggingface.co/google/siglip-large-patch16-384)

²HuggingFace Dataset: [google/wit](https://huggingface.co/google/wit)

³HuggingFace Dataset: [laion/laion-art](https://huggingface.co/laion/laion-art)

⁴5K test set split from: http://cs.stanford.edu/people/karpathy/deepimagesent/caption_datasets.zip

⁵HuggingFace Dataset: [Salesforce/blip3-kale](https://huggingface.co/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.

gests 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 illustra-

tion 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 Data

AURORALONG 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 B, Table B2, and Table B3

C. Evaluation Results

In this section we report the detailed evaluation results of AURORALONG on MLVU and MVBench in Table C4 and Table C5. We find that although AURORALONG 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 AURORALONG’s strong capability in handling long video inputs. Note that we do not conduct LLM context extension as in [62]. Although only trained on videos less than 60 frames, AURORALONG generalizes well on long video tasks, aligning with the loss curve its LLM backbone shows when validated on extended textual context length up to 4X its pretrained context length, as showed in [43].

D. Ablation Studies

As a core strategy of AURORALONG, 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 Figure D2 and Figure D3. 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 AURORALONG focuses on spatial visual token merging, while the temporal features introduce additional complexity to explore the token merging laws. Appendix A shows more calculation details and the visualization results of token merging.

E. Case Study

We conduct extensive case studies of AURORALONG 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 AURORALONG with those produced by other leading models. As shown in the followings, AURORALONG 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 E4 [Question] In this video, how did the protagonist stop the train?

VILA [33] The protagonist stops the train by shooting webs at it .

InternVL2 [10] 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 web between the train cars and the tracks , effectively stopping the train from moving.

LongVU [46] 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 .

AURORALONG 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 falling into the cliff .

Method Answer

Figure E5 [Question] In this video, how did the man ride on the giant monster?

VILA [33] The man in the video rode on the giant monster by using a rope to climb onto its back .

InternVL2 [10] 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 [46] The man in the video rode on the giant monster by climbing up a hill .

AURORALONG 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 E6 [Question] Describe the events in the video shown by these frames in at least three sentences.

VILA [33] 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 [10] 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 [46] Two men are sitting at a table in a fast-food restaurant. One man is smoking a cigarette

Table B1. Summary of datasets used for training AURORALONG in Pretraining stage.

Task	# Sample	Dataset
Image Captioning	1.3M	LAION-CC-SBU-595K [37], ShareGPT4V [6], ALLaVA-Caption-LAION-4V [5], ALLaVA-Caption-VFLAN-4V [5], DenseFusion [30]

Table B2. Summary of datasets used for training AURORALONG in Vision stage. For classification, Reasoning, VQA, and Generation tasks, we adopt the dataset processed by M³IT [27] to fit the training objective of language models.

Task	# Sample	Dataset
Captioning	1,925K	ShareGPT4V-PT [6], TextCaps [47], Image-Paragraph-Captioning [23]
Object-centric	438K	COST [19], ChatterBox [53], V* [58]
Classification	238K	COCO-GOI [34], COCO-Text [55], ImageNet [44], COCO-ITM [34], e-SNLI-VE [22], Mocheg [59], IQA [14]
Reasoning	100K	CLEVR [20], NLVR [50], VCR [61], VisualMRC [51], Winoground [52]
VQA	3,518K	VQA v2 [17], Shapes VQA [1], DocVQA [41], OK-VQA [40], Text-VQA [48], OCR-VQA [42], A-OK-VQA [45], ScienceQA [38]
Generation	145K	ST-VQA [2], ViQuAE [24], LLaVA-OneVision [25]
Chinese	193K	Visual Storytelling [18], Visual Dialog [13], Multi30k [15]
		COCO-Caption CN [29], Flickr-8k-Caption CN [28], multimodal Chat [65], FM-IQA [16], ChineseFoodNet [8]
Total	6.6M	For all datasets, we uniformly sample without duplication.

Table B3. Summary of datasets used for training AURORALONG in Language stage.

Task	# Sample	Dataset
Image Captioning	1,779K	ShareGPT4V [6], ALLaVA-Caption-LAION-4V [5], ALLaVA-Caption-VFLAN-4V [5], DenseFusion [30], FaceCaption [12]
Video Captioning	1,659K	MiraData [21], LLaVA-Hound [63], ShareGPT4Video [7]
Image Instruction	9,742K	LVIS-Instruct4V [56], ALLaVA-Instruct-LAION-4V [5], ALLaVA-Instruct-VFLAN-4V [5], Cambrian [54], LLaVA-Mix-665K [35], M4-Instruct [36]
Video Instruction	446K	LLaVA-Hound [63], ShareGPT4Video [7], LLaVA-Video-178k [64]
Language-only	143K	Evol-Instruct-GPT4-Turbo-143K [5]
Total	15.4M	We duplicate video captioning and instruction datasets in training.

220
221
222
223
224
225
226
227
228

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

AURORALONG

is reading a letter and then puts it down. 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

229
230
231
232
233
234
235
236
237
238
239
240

Table C4. Comparison of AURORALONG with SOTA methods on MLVU [66] 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, AURORALONG 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
<i>Proprietary Models</i>											
GPT4-o	0.5fps	128k	-	54.5	68.8	<u>47.8</u>	<u>46.2</u>	<u>35.0</u>	83.7	42.9	57.1
GPT4-Turbo	16frm	128k	-	43.8	<u>61.5</u>	41.5	22.9	6.7	<u>85.7</u>	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	200k+	-	21.8	30.8	17.0	10.0	6.7	53.8	14.0	20.0
<i>Open-Source Video LMMs</i>											
LLAMA-VID [31]	1 fps	4k	7B	18.1	23.1	11.3	18.6	15.0	20.9	21.7	16.0
mPLUG-Owl-V [60]	16 frm	4k	7B	16.7	15.4	13.2	14.3	20.0	25.3	6.7	22.0
Video-ChatGPT [39]	16 frm	2k	7B	21.2	17.9	32.1	17.1	13.3	17.6	28.3	22.0
MovieChat [49]	2048 frm	4k	7B	16.5	10.3	15.1	17.1	15.0	18.7	23.3	16.0
Video-LLAVA [39]	8 frm	4k	7B	30.1	38.5	26.4	20.0	21.7	70.3	13.3	26.0
LLaVA-1.6 [36]	16 frm	8k	7B	27.1	17.9	26.4	21.4	16.7	63.7	13.3	30.0
LongVA [62]	256 frm	224k	7B	42.1	41.0	39.6	17.1	23.3	81.3	46.7	46.0
VideoChat2 [26]	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 [7]	16 frm	8k	8B	34.2	25.6	45.3	17.1	8.3	73.6	31.7	38.0
Video-LLAMA-2 [11]	16 frm	131k	13B	18.8	12.8	17.0	15.7	8.3	52.7	13.3	12.0
InternVL-1.5 [9]	16 frm	4k	26B	37.9	51.3	24.5	14.3	13.3	80.2	40.0	42.0
VILA-1.5 [33]	14 frm	4k	40B	46.2	56.4	35.8	34.3	11.7	84.7	38.3	62.0
InternVL-2	16 frm	8k	76B	47.5	51.3	47.2	32.9	15.0	<u>85.7</u>	<u>48.3</u>	52.0
AURORALONG(ours)	48 frm	4k	2B	<u>51.1</u>	59.5	54.8	29.4	42.9	59.5	45.2	<u>61.9</u>

Table C5. Results on MVBench [26] whose videos primarily range from 5s to 35s. We find that despite only trained on publicly available short videos, AURORALONG outperforms proprietary models like GPT4-V and Gemini Pro. The best result is highlighted in bold, and the second best is underlined. CTX denotes LLM pretrained context length and maximum context length for proprietary models.

Models	Input	Size	MVBench																	
			AVG.	AC	AL	AP	AS	CO	CI	EN	FGA	MA	MC	MD	OE	OI	OS	ST	SC	UA
<i>Proprietary Models</i>																				
GPT4-V	16 frm	-	<u>43.7</u>	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	<u>83.5</u>	45.0	<u>73.5</u>
Gemini Pro	16 frm	-	37.7	37.7	<u>40.0</u>	41.8	35.4	38.7	33.7	<u>36.4</u>	36.2	41.5	18.0	16.5	43.5	37.5	<u>39.8</u>	75.4	42.3	67.1
<i>Open-Source Video LMMs</i>																				
Video-LLaMA [32]	16 frm	7B	34.1	34.0	22.5	25.5	27.5	40.0	<u>37.0</u>	30.0	29.0	32.5	22.5	22.5	48.0	40.5	38.0	43.0	<u>45.5</u>	39.0
LLAMA-VID [31]	16 frm	7B	41.4	<u>42.0</u>	26.5	43.0	42.0	39.0	34.5	36.5	35.5	<u>44.5</u>	<u>28.5</u>	19.0	55.6	37.5	34.0	84.5	40.5	56.5
mPLUG-Owl-V [60]	16 frm	7B	29.4	34.5	24.0	20.0	25.0	37.0	37.0	25.5	27.0	31.5	22.0	<u>23.0</u>	36.0	24.0	34.0	34.5	40.0	23.5
Video-ChatGPT [39]	100 frm	7B	32.7	30.5	20.0	26.0	23.5	33.0	35.5	29.5	22.5	39.5	25.5	<u>23.0</u>	<u>54.0</u>	28.0	40.0	31.0	48.5	26.5
AURORALONG(ours)	48 frm	2B	46.4	44.5	27.0	<u>47.5</u>	<u>46.3</u>	<u>47.0</u>	38.5	31.25	<u>37.5</u>	59.0	45.6	40.6	62.5	<u>45.0</u>	31.5	88.8	44.4	74.4

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 E7 [Question] Describe the events in the video shown by these frames in at least three sentences.

VILA [33] 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 [10] 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 hero-

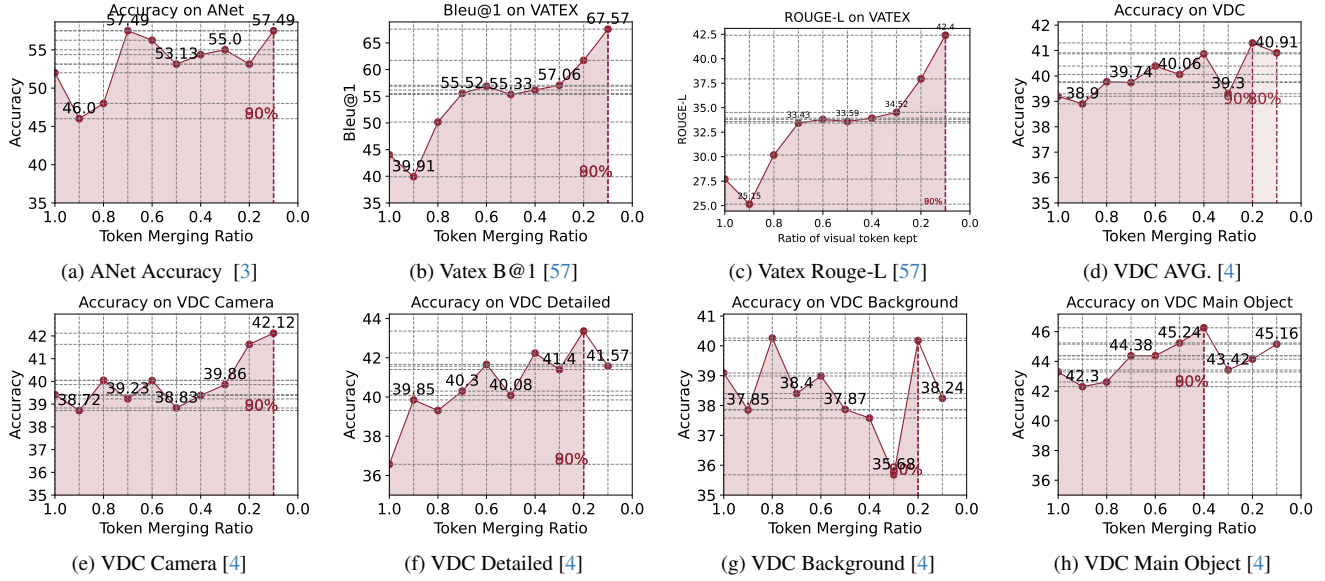


Figure D2. Ablation study of token merging with no sort in short video question answering on ANet [3], short video sparse captioning on VATEX [57], short video dense captioning on VDC [4], and long video question answering on MovieChat-1K [49]. We find that token merging significantly reduces the number 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.

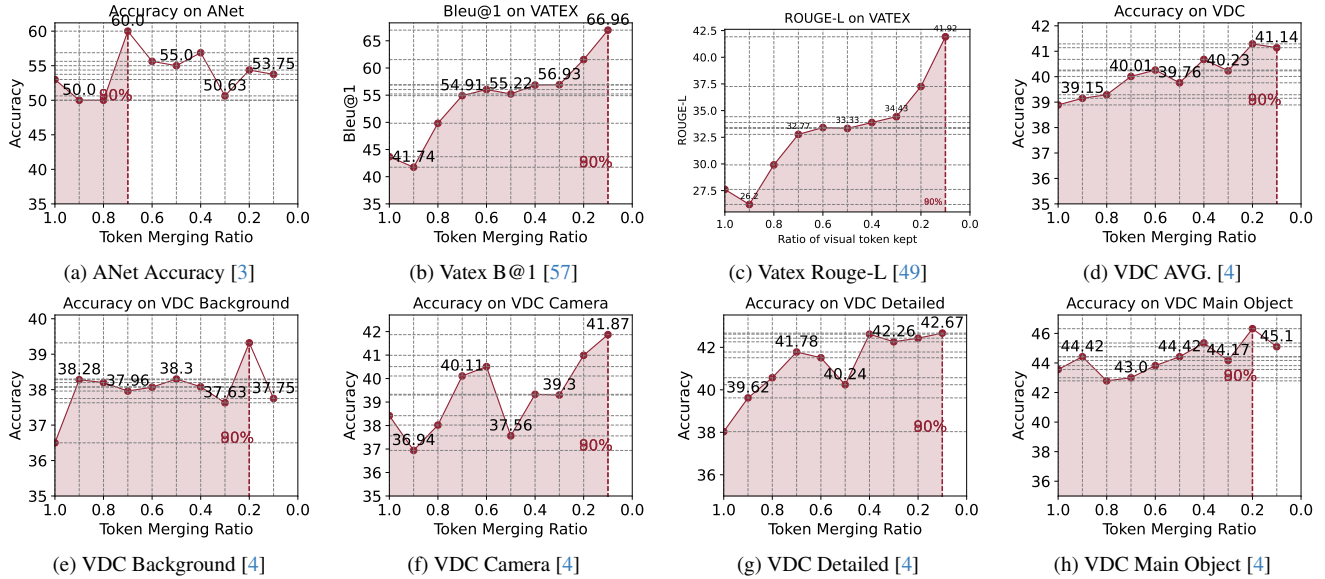


Figure D3. Ablation study of token merging with descending order in short video question answering on ANet [3], short video sparse captioning on VATEX [57], short video dense captioning on VDC [4], and long video question answering on MovieChat-1K [49]. We find that token merging significantly reduces 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.

LongVU [46] 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,

273 with a focus on the character's hands and the
274 smartphone they're holding. The perspective shifts
275 to show the character's feet as they leap off a
276 building, with a view of the cityscape below and a
277 large, illuminated billboard in the background.

278 AURORALONG The video showcases a dynamic and visually strik-
279 ing scene set in a bustling cityscape, where a super-
280 hero, clad in a striking blue and red costume, is seen
281 performing a series of acrobatic maneuvers against a
282 backdrop of a city skyline. Initially, the superhero is
283 seen in mid-air, executing a series of acrobatic flips
284 and spins, with a large, fiery explosion occurring in
285 the background, suggesting a dramatic or explosive
286 moment. As the video progresses, the superhero
287 continues to perform acrobatic stunts, including a
288 dramatic leap and a flip, while the cityscape remains
289 a constant backdrop, emphasizing the action. The
290 scene is filled with a sense of urgency and drama,
291 highlighted by the presence of a large, fiery explo-
292 sion and the superhero's intense focus on the task at
293 hand. The video captures the essence of a superhero
294 in action, with the cityscape serving as a backdrop
295 that enhances the dramatic and action-packed nature
296 of the scene.



Figure E4. 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.

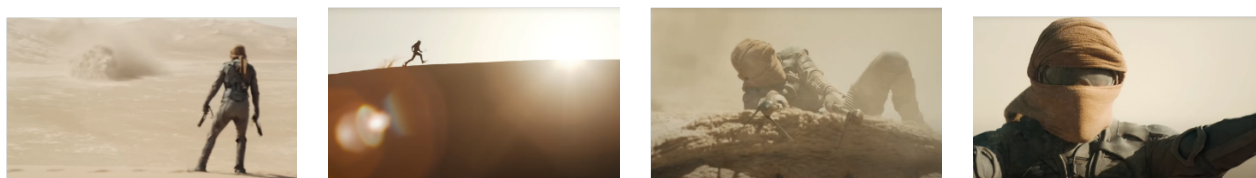


Figure E5. Video clips from Dune, a sci-fi epic, which tells the story of Paul Atreides, a young nobleman who must navigate political intrigue and warfare on the desert planet Arrakis while embracing his destiny to protect its precious resource and lead its people.



Figure E6. 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 E7. 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.

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] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluís 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
- [3] 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. 6
- [4] 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. 6
- [5] Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. *arXiv preprint arXiv:2402.11684*, 2024. 4
- [6] 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
- [7] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. Sharegpt4video: Improving video understanding and generation with better captions. *arXiv preprint arXiv:2406.04325*, 2024. 4, 5
- [8] 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
- [9] Zhe Chen, Jiannan Wu, Wenhao 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
- [10] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 3, 5
- [11] 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
- [12] Dawei Dai, YuTang Li, YingGe Liu, Mingming Jia, Zhang YuanHui, and Guoyin Wang. 15m multimodal facial image-text dataset, 2024. 4
- [13] 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
- [14] 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
- [15] Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30k: Multilingual english-german image descriptions. *arXiv preprint arXiv:1605.00459*, 2016. 4
- [16] 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
- [17] 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
- [18] 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
- [19] Jitesh Jain, Jianwei Yang, and Humphrey Shi. Vcoder: Versatile vision encoders for multimodal large language models. *arXiv preprint arXiv:2312.14233*, 2023. 4
- [20] 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
- [21] 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
- [22] 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
- [23] 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
- [24] 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
- [25] 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
- [26] 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
- [27] 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
- [28] 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
- [29] 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
- [30] 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
- [31] 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
- [32] 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
- [33] 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. 3, 5
- [34] 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
- [35] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 4
- [36] 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
- [37] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 4
- [38] 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
- [39] 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
- [40] 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
- [41] 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
- [42] 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
- [43] 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. 3
- [44] 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
- [45] 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
- [46] Xiaoqian Shen, Yunsang 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. 3, 6
- [47] 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
- [48] 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
- [49] 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, 6

- [50] 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
- [51] 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
- [52] 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
- [53] 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
- [54] 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
- [55] 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
- [56] 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
- [57] 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
- [58] Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. *arXiv preprint arXiv:2312.14135*, 2023. 4
- [59] 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
- [60] 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
- [61] 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
- [62] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024. 3, 5
- [63] Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander Hauptmann, Yonatan Bisk, et al. Direct preference optimization of video large multimodal models from language model reward. *arXiv preprint arXiv:2404.01258*, 2024. 4
- [64] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data, 2024. 4
- [65] Yinhe Zheng, Guanyi Chen, Xin Liu, and Jian Sun. Mmchat: Multi-modal chat dataset on social media. *arXiv preprint arXiv:2108.07154*, 2021. 4
- [66] 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. 5