Long-Context Vision Large Language Models: Empirical Insights and A Baseline

Yongshuo Zong [1] **Ismail Elezi** **Yongxin Yang** [1] **Jiankang Deng** [2] **Timothy Hospedales** [1]

Abstract

The development of long-context large language models (LLMs) has attracted significant interest. However, progress in advancing long-context vision large language models (VLLMs) falls behind, despite their vast potential in applications like high-resolution input, multimodal in-context learning, multi-image understanding, and video understanding. In this paper, we present an empirical study to identify major challenges in developing long-context VLLMs and present a simple yet effective baseline for long-context tasks. By captioning the images separately and aggregating the captions as inputs, we directly alleviate the input length issue and also show that it outperforms other context extension or token reduction strategies effectively.

1. Introduction

Vision Large Language Models (VLLMs) (Yin et al., 2023; OpenAI, 2023; Team et al., 2023; Liu et al., 2023b; Bai et al., 2023a; Zhang et al., 2023b), have attracted widespread attention for their remarkable multi-modal capabilities such as visual perception and reasoning. Beyond general single image understanding tasks, researchers have been probing other important capabilities of VLLMs, such as using high-resolution images as inputs (Dong et al., 2024b; McKinzie et al., 2024), multimodal in-context learning (Zong et al., 2024; Sun et al., 2023), multi-image understanding (Lin et al., 2024; Jiang et al., 2024), and video understanding (Lin et al., 2023; Jin et al., 2024b). These applications necessitate a long context window to effectively accommodate the long input sequence, especially the visual tokens. While researchers have been actively advancing long-context large language models (LLMs) (Liu et al., 2023c; Ding et al., 2024; Chen et al., 2024a; Munkhdalai et al., 2024), long-context research in VLLMs remains underexplored.

[1]University of Edinburgh [2]Imperial College London. Correspondence to: Yongshuo Zong <yongshuo.zong@ed.ac.uk>.

Proceedings of the 1st Workshop on Long-Context Foundation Models, Vienna, Austria. 2024.

Considering the various advancements in long-context LLMs, we first ask: does the long-context ability of LLMs directly transfer to VLLMs? To answer this, we first compare the long-context ability of VLLMs and their base LLMs on text benchmarks (Bai et al., 2023b). The results show that VLLMs consistently perform worse than their LLM counterparts on long-context tasks, while maintaining similar performance on short-context tasks (Hendrycks et al., 2021; Dua et al., 2019). This indicates that LLMs' long-context capability is harmed during VLLM fine-tuning.

We further investigate VLLMs on two practical long-context vision-language tasks: multimodal in-context learning and video understanding. A straightforward explanation for VLLMs' struggle with these tasks is their limited context length. However, even if we extend their context window by position encoding extrapolation (Jin et al., 2024a) or reduce the input length by pruning tokens (Shang et al., 2024), the improvement is not immediate. This is mainly because these methods do not address the perception problem introduced by interleaved image-text multi-modal understanding.

In light of this, we propose a simple training-free baseline to enable VLLMs with limited context length to handle long-context tasks. We adopt a divide-and-conquer strategy: first, the VLLMs caption the input images separately, and then we aggregate these captions as input for the target task. This approach transforms the long-context multimodal interleaved input into a short-context single-modality input, greatly simplifying the problem. Our experiments demonstrate that this straightforward strategy outperforms other methods in many cases.

2. Do LLMs Long-Context Capabilities Transfer to VLLMs?

Preliminary Current VLLMs typically comprise three main components: a vision encoder, a large language model, and a connection module that aligns vision and language tasks (Yin et al., 2023; Zong et al., 2023). During the training of VLLMs, these components are fine-tuned according to specific design choices. The LLM fine-tuning is particularly crucial to enable language models to integrate different modalities and interpret visual instructions effectively (Lin

et al., 2024; Bai et al., 2023a; Liu et al., 2023a).

Considering the extensive pre-trained long-context LLMs, we ask the natural question: *do VLLMs inherit the long-context ability from their LLMs?* In other words, will VLLM fine-tuning harm the existing long-context ability of the LLM? To answer this question, we compare the performance of VLLMs and their LLM counterparts on a long-context benchmark LongBench (Bai et al., 2023b).

Setup We evaluate state-of-the-art VLLMs including LLaVA-v1.5 (Liu et al., 2023a), VILA (Lin et al., 2024), InternLM-XComposer2 (Dong et al., 2024a), Qwen-VL (Bai et al., 2023a) and Phi-3-Vision (Abdin et al., 2024), and compare the performance with their base LLMs. To evaluate their capability with long contexts, we employ LongBench (Bai et al., 2023b), a comprehensive text benchmark consisting of various tasks including single-document QA, multi-document QA, summarization, few-shot learning, synthetic tasks, and code completion. These tasks have an average length of 6,711 English words. When input sequences exceed the models' maximum length, we follow LongBench's implementation to truncate the middle portion of the inputs and concatenate the beginning and end parts. Additionally, we use two widely used language benchmarks MMLU (Hendrycks et al., 2021) and DROP (Dua et al., 2019) to evaluate the text-only performance in short-context for ablation.

Results We compare the long-context and general (short-context) text capabilities of VLLMs (e.g., LLaVA-v1.5-7B) and their corresponding LLMs (e.g., Vicuna-7B) in Table 1 and the breakdown of different tasks in Appendix Table 3. Across various VLLMs of different sizes and model families, there is a general consistent decline in performance compared to their base LLMs across LongBench. Although one might conjecture that fine-tuning of VLLMs negatively affects their general text ability, our findings suggest otherwise. Standard evaluations using MMLU and DROP show that VLLMs generally outperform or are comparable to their LLM counterparts, demonstrating robust short-range text capabilities. This suggests that the decline in long-context performance is not due to a deterioration in general text ability. Instead, it likely stems from the nature of the vision-language fine-tuning data (e.g., single-image QA), which is typically much shorter than the text in the pretraining corpus, causing the VLLMs to lose their long-context capability during fine-tuning.

3. How do VLLMs Perform on Multimodal Long-Context Tasks?

In this section, we examine current VLLMs on long-context vision-language tasks. Specifically, we consider two scenarios where VLLMs require long input sequences: (1) multimodal in-context learning (ICL), and (2) video understanding.

Since most VLLMs are trained under a very limited input length (e.g., 4096 tokens for LLaVA), inputs that exceed the pre-trained context window result in a position encoding out-of-distribution (*O.O.D.*) problem, as noted in previous studies (Jin et al., 2024a; Chen et al., 2023). Therefore, in addition to experimenting with vanilla models, we explore two approaches to mitigate the position encoding *O.O.D.* problem to better understand the behaviours of VLLMs: (1) extrapolating position encoding and (2) reducing the number of visual tokens. The experimental details for these approaches are described in the following subsections.

3.1. Multimodal In-Context Learning

Vision-language ICL involves constructing a task demonstration that includes a few examples, each with an input-label pair. A query is then appended to these few-shot demonstrations as the input prompt for the VLLMs. The VLLMs are expected to learn the input-answer mapping from the few-shot examples and respond to the query without updating the model parameters. For LLMs, increasing the number of shots typically improves task performance in ICL. Naturally, we expect VLLMs to perform better with more shots. However, since each shot includes at least one image, resulting in a large number of visual tokens, vision-language ICL with many shots requires a longer context window.

Setup We conduct evaluation on VL-ICL Bench (Zong et al., 2024), which consists of various ICL tasks. We use VILA (Lin et al., 2024) 2.7B and 7B variants for experiments, as it is trained on interleaved image-text data and shown to be an effective in-context learner. Additionally, as the context window of VILA is 4096, which can only take less than 8 images, we consider two approaches to make it accept more images for comparison: SelfExtend (Jin et al., 2024a) for position encoding extrapolation and visual token reduction (Shang et al., 2024).

Results Figure 1 shows the results of VILA-7B on various tasks (we also provide results of VILA-2.7B in the appendix). It can be seen that the performance of the original VILA saturates quickly within 2-4 shots. Using the context extension strategy SelfExtend, we do observe the improvement of the performance with respect to more shots, e.g., on TextOCR. However, for many other subsets, including CLEVR, Open MiniImageNet, and Operator Induction, there is minimal or no improvement (as seen with Operator Induction). These tasks require more complex reasoning over interleaved images and texts, and simply extending the context via position encoding extrapolation or token reduction does not provide fundamental benefits as the model still does not effectively understand the input.

Table 1. Comparisons of language-based long-context (LongBench) and short-context (MMLU and DROP) tasks between VLLMs and their base LLMs. VLLMs maintain the short-context ability while consistently performing worse than their LLM counterparts on long-context tasks.

Model	Max. Length	LongBench					Standard Evaluation	
		0-4K	4-8K	8-16K	16K+	Average	MMLU	DROP
Vicuna-7B	4K	40.46	33.74	32.26	33.35	34.95	48.55	39.92
LLaVA-v1.5-7B	4K	38.35 (2.11 ↓)	33.33 (0.41 ↓)	32.17 (0.09 ↓)	33.23 (0.12 ↓)	34.33 (0.62 ↓)	40.52	43.44
Llama2-Chat-7B	4K	44.06	34.48	35.42	36.64	37.86	45.36	34.39
VILA-7B	4K	41.22 (2.84 ↓)	36.92 (2.44 ↑)	34.92 (0.50 ↓)	35.24 (1.40 ↓)	37.08 (0.78 ↓)	48.10	51.16
QwenLM-7B	8K	39.27	36.69	34.42	36.12	36.62	49.01	48.08
Qwen-VL-7B	8K	31.50 (7.77 ↓)	27.70 (8.99 ↓)	26.40 (8.02 ↓)	32.93 (3.19 ↓)	28.53 (8.09 ↓)	50.87	48.13
InternLM2-Chat-1.8B	32K	41.97	39.97	38.97	35.37	39.07	42.84	43.94
InternLM-XComposer2-1.8B	32K	41.14 (0.83 ↓)	37.35 (2.62 ↓)	35.88 (3.09 ↓)	30.52 (4.85 ↓)	38.25 (0.82 ↓)	44.87	44.17
Phi-3	128K	50.80	47.94	47.03	39.80	46.39	63.27	53.93
Phi-3-Vision	128K	47.76 (3.04 ↓)	41.78 (6.16 ↓)	36.97 (10.06 ↓)	24.93 (14.87 ↓)	37.86 (8.53 ↓)	62.85	53.55

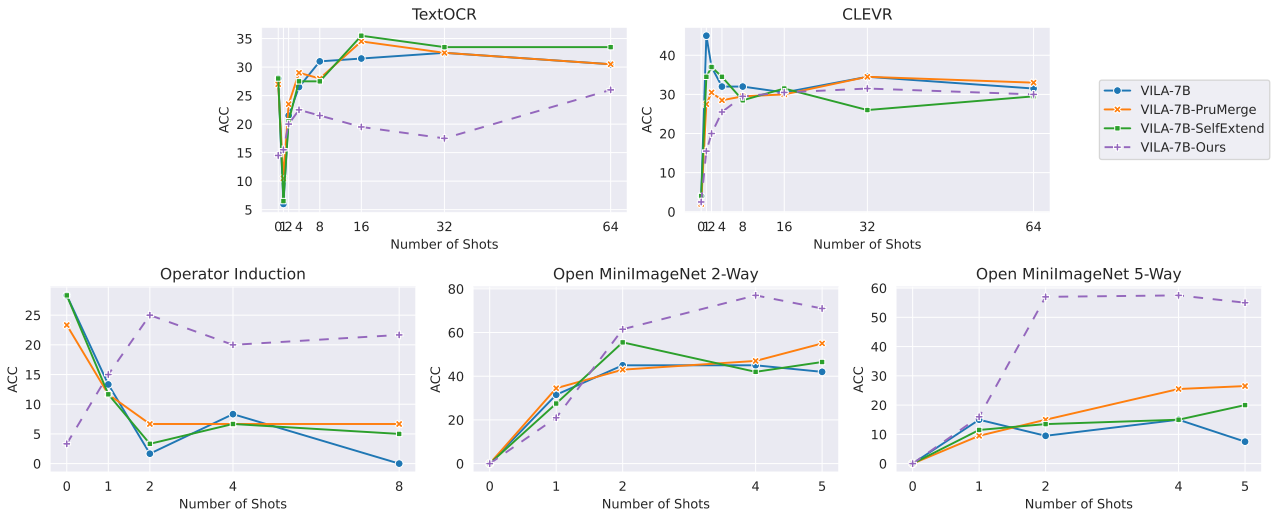

Figure 1. Performance of VILA-7B on VL-ICL tasks. The original model saturates quickly *w.r.t.* shots. SelfExtend or PruMerge cannot effectively improve tasks that require complex reasoning while our method can substantially improve these tasks.

3.2. Video Understanding

Understanding videos with VLLMs involves sampling multiple frames from the video and treating each frame as an image input (Maaz et al., 2023; Lin et al., 2023; Zhang et al., 2023a; Lin et al., 2024). A larger number of frames offers more comprehensive information, necessitating VLLMs to handle more frames for better video analysis.

Setup We use VILA-7B (Lin et al., 2024) and Video-LLaVA (Lin et al., 2023) and experiment on two popular video question-answering datasets MSVD (Chen & Dolan, 2011) and TGIF-QA (Jang et al., 2017). We follow Lin et al. (2023) to use GPT-3.5 to evaluate predictions. We use 8 as the default number of frames as in Lin et al. (2023) and further extend it to 16 using SelfExtend (Jin et al., 2024a) or PruMerge (Shang et al., 2024) to fit in the context window.

Results As shown in Table 2, we have two key observations: (1) Increasing the number of frames using both SelfExtend and PruMerge improves performance, highlighting the necessity of accommodating more frames with a longer context window. (2) PruMerge (efficient token pruning and merging) generally performs worse than SelfExtend (position encoding extrapolation) but better than the original model. This indicates that while visual token reduction with PruMerge results in some information loss, it still provides benefits from incorporating more frames. Note that 8 frames already exceed the context length of VILA and thus all methods give substantial improvement.

4. A Simple Baseline for Long-Context Vision-Language Tasks

Motivated by previous experiments, we present a simple yet effective *divide-and-conquer* strategy for long-context

Table 2. Video-QA results on MSVD and TGIF-QA datasets.

	Method	# Tokens / Frame	# Frames	MSVD		TGIF-QA	
				Accuracy	Score	Accuracy	Score
Video-LLaVA	Original	256	8	69.02	3.93	68.75	3.82
	+PruMerge	100	16	69.95	3.94	68.91	3.84
	+SelfExtend	256	16	**70.65**	3.96	69.01	3.87
	+Ours	256	16	70.61	**3.97**	**69.82**	**3.91**
VILA-7B	Original	576	8	57.15	2.26	21.32	2.51
	+PruMerge	100	16	75.16	3.99	48.73	3.11
	+SelfExtend	576	16	75.25	4.00	51.46	3.19
	+Ours	576	16	**75.63**	**4.01**	**55.47**	**3.26**

vision-language tasks. The idea is straightforward: we first ask the VLLM to caption the input images separately, and then aggregate the captions as the input for the downstream tasks. We formally describe the procedures in Algorithm 1. This strategy divides the long input sequence into multiple single-image, short-context tasks, which naturally mitigates the long-context problem. Additionally, it allows VLLMs trained on single image-text pairs (e.g., LLaVA) to handle tasks involving multiple images.

Algorithm 1 Proposed Method

Input: Set of input images $\{I_1, I_2, \ldots, I_n\}$, task query P, VLLM M
Output: Generated answer Y
Initialize an empty list of captions C
for each image I_i in $\{I_1, I_2, \ldots, I_n\}$ **do**
    Generate caption C_i for image I_i using model M
    Append caption C_i to list C
end for
Concatenate all captions in list C with task prompt P to form input query Q_{CP}
Generate answer Y from input Q_{CP} using model M

For multimodal in-context learning, our method is particularly useful for tasks requiring more reasoning and induction, such as Operator Induction and Open MiniImageNet. For instance, we improved the performance of 5-way Open MiniImageNet from below 30% to over 50%. This is because these tasks demand stronger interaction between different modalities, such as associating a visual concept with a fixed random word or deducing the operator from an image of digits. Our method successfully bypasses this issue by translating images to captions and thus all inputs are of text modality. However, our method does not handle well tasks that require more fine-grained perception such as CLEVR and TextOCR since captioning only captures coarse-grained information. For video tasks, our method generally outperforms SelfExtend and PruMerge, highlighting the effectiveness of the proposed strategy.

5. Related Works

Context Extension for LLMs LLMs are typically pre-trained with a fixed context length (e.g., 4096 for LLaMA2 (Touvron et al., 2023)). To extend this context length, various approaches have been proposed, which can be broadly categorized into fine-tuning and training-free methods. Fine-tuning approaches often focus on efficient attention mechanisms (Liu et al., 2023c; Munkhdalai et al., 2024; Chen et al., 2024b) or new architectures (Peng et al., 2023a; Gu & Dao, 2023) to mitigate the quadratic complexity of standard attention. Training-free methods include techniques such as position encoding interpolation or extrapolation (Jin et al., 2024a; Chen et al., 2023; Peng et al., 2023b), context chunking or pruning strategies (Xiao et al., 2024; Dai et al., 2024; An et al., 2024), etc.

Long-Context Evaluations Various ways have been proposed to evaluate the long-context ability of LLMs: (1) Perplexity-based evaluation: This involves assessing the perplexity of long-sequence language modeling performance on datasets such as the book corpus dataset (Rae et al., 2020), and lower perplexity indicates more stable performance on long context tasks, (2) Needle-in-a-haystack: A common strategy to test in-context retrieval ability, this method involves placing a random fact or statement (the "needle") in the middle of a long context window (the "haystack") and asking the model to retrieve this statement, (3) Downstream tasks: Researchers have developed various benchmarks (Bai et al., 2023b; Li et al., 2023; An et al., 2023) to evaluate long-context downstream tasks, such as summarization and multi-document QA. Additionally, other types of long-context evaluations, such as many-shot in-context learning, are being explored (Agarwal et al., 2024).

Long-Context VLLMs Unlike LLMs, long-context VLLMs are relatively under-explored. Among closed-source models, GPT-4V (OpenAI, 2023) and Gemini 1.5 (Reid et al., 2024) have extended context lengths to 100k and even 10 million tokens. For open-source models, Liu et al. (2024) have curated a large dataset of videos and books, using RingAttention (Liu et al., 2023c) to train on long sequences, progressively increasing the context size to 1 million tokens. To the best of our knowledge, there is no specific strategy designed for VLLM context extension.

6. Conclusion

In this work, we present an initial empirical analysis of long-context VLLMs. Our findings indicate that VLLMs do not necessarily retain the long-context abilities of their base LLMs after fine-tuning for multimodal alignment, and that context extension or efficient token pruning methods do not effectively address complex multimodal long-context inputs. To address this, we propose a simple strategy that mitigates the multimodal long-context problem by individually understanding each image and then aggregating them in text form. While we do not claim this to be the optimal solution, we offer it as a baseline and encourage the development of more effective approaches in the community.

Acknowledgement

Yongshuo Zong is supported by the United Kingdom Research and Innovation (grant EP/S02431X/1), UKRI Centre for Doctoral Training in Biomedical AI at the University of Edinburgh, School of Informatics. For the purpose of open access, the author has applied a creative commons attribution (CC BY) licence to any author accepted manuscript version arising.

References

Abdin, M., Jacobs, S. A., Awan, A. A., Aneja, J., Awadallah, A., Awadalla, H., Bach, N., Bahree, A., Bakhtiari, A., Behl, H., et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Agarwal, R., Singh, A., Zhang, L. M., Bohnet, B., Chan, S., Anand, A., Abbas, Z., Nova, A., Co-Reyes, J. D., Chu, E., et al. Many-shot in-context learning. *arXiv preprint arXiv:2404.11018*, 2024.

An, C., Gong, S., Zhong, M., Li, M., Zhang, J., Kong, L., and Qiu, X. L-eval: Instituting standardized evaluation for long context language models. *arXiv preprint arXiv:2307.11088*, 2023.

An, C., Huang, F., Zhang, J., Gong, S., Qiu, X., Zhou, C., and Kong, L. Training-free long-context scaling of large language models. *arXiv preprint arXiv:2402.17463*, 2024.

Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., and Zhou, J. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023a.

Bai, Y., Lv, X., Zhang, J., Lyu, H., Tang, J., Huang, Z., Du, Z., Liu, X., Zeng, A., Hou, L., Dong, Y., Tang, J., and Li, J. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*, 2023b.

Chen, D. and Dolan, W. B. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, 2011.

Chen, S., Wong, S., Chen, L., and Tian, Y. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*, 2023.

Chen, Y., Qian, S., Tang, H., Lai, X., Liu, Z., Han, S., and Jia, J. Longlora: Efficient fine-tuning of long-context large language models. *ICLR*, 2024a.

Chen, Y., Qian, S., Tang, H., Lai, X., Liu, Z., Han, S., and Jia, J. Longlora: Efficient fine-tuning of long-context large language models. *ICLR*, 2024b.

Dai, J., Huang, Z., Jiang, H., Chen, C., Cai, D., Bi, W., and Shi, S. Sequence can secretly tell you what to discard. *arXiv preprint arXiv:2404.15949*, 2024.

Ding, Y., Zhang, L. L., Zhang, C., Xu, Y., Shang, N., Xu, J., Yang, F., and Yang, M. Longrope: Extending llm context window beyond 2 million tokens. *arXiv preprint arXiv:2402.13753*, 2024.

Dong, X., Zhang, P., Zang, Y., Cao, Y., Wang, B., Ouyang, L., Wei, X., Zhang, S., Duan, H., Cao, M., et al. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*, 2024a.

Dong, X., Zhang, P., Zang, Y., Cao, Y., Wang, B., Ouyang, L., Zhang, S., Duan, H., Zhang, W., Li, Y., et al. Internlm-xcomposer2-4khd: A pioneering large vision-language model handling resolutions from 336 pixels to 4k hd. *arXiv preprint arXiv:2404.06512*, 2024b.

Dua, D., Wang, Y., Dasigi, P., Stanovsky, G., Singh, S., and Gardner, M. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *NAACL*, 2019.

Gu, A. and Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. *ICLR*, 2021.

Jang, Y., Song, Y., Yu, Y., Kim, Y., and Kim, G. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *CVPR*, 2017.

Jiang, D., He, X., Zeng, H., Wei, C., Ku, M., Liu, Q., and Chen, W. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*, 2024.

Jin, H., Han, X., Yang, J., Jiang, Z., Liu, Z., Chang, C.-Y., Chen, H., and Hu, X. Llm maybe longlm: Self-extend llm context window without tuning. *arXiv preprint arXiv:2401.01325*, 2024a.

Jin, P., Takanobu, R., Zhang, C., Cao, X., and Yuan, L. Chat-univi: Unified visual representation empowers large language models with image and video understanding. *CVPR*, 2024b.

Li, J., Wang, M., Zheng, Z., and Zhang, M. Loogle: Can long-context language models understand long contexts? *arXiv preprint arXiv:2311.04939*, 2023.

Lin, B., Zhu, B., Ye, Y., Ning, M., Jin, P., and Yuan, L. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.

Lin, J., Yin, H., Ping, W., Lu, Y., Molchanov, P., Tao, A., Mao, H., Kautz, J., Shoeybi, M., and Han, S. Vila: On pre-training for visual language models. *CVPR*, 2024.

Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023a.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *NeurIPS*, 2023b.

Liu, H., Zaharia, M., and Abbeel, P. Ring attention with blockwise transformers for near-infinite context. *arXiv preprint arXiv:2310.01889*, 2023c.

Liu, H., Yan, W., Zaharia, M., and Abbeel, P. World model on million-length video and language with ringattention. *arXiv preprint arXiv:2402.08268*, 2024.

Maaz, M., Rasheed, H., Khan, S., and Khan, F. S. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023.

McKinzie, B., Gan, Z., Fauconnier, J.-P., Dodge, S., Zhang, B., Dufter, P., Shah, D., Du, X., Peng, F., Weers, F., et al. Mm1: Methods, analysis & insights from multimodal llm pre-training. *arXiv preprint arXiv:2403.09611*, 2024.

Munkhdalai, T., Faruqui, M., and Gopal, S. Leave no context behind: Efficient infinite context transformers with infini-attention. *arXiv preprint arXiv:2404.07143*, 2024.

OpenAI, R. Gpt-4 technical report. *arXiv*, pp. 2303–08774, 2023.

Peng, B., Alcaide, E., Anthony, Q., Albalak, A., Arcadinho, S., Cao, H., Cheng, X., Chung, M., Grella, M., GV, K. K., et al. Rwkv: Reinventing rnns for the transformer era. *arXiv preprint arXiv:2305.13048*, 2023a.

Peng, B., Quesnelle, J., Fan, H., and Shippole, E. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*, 2023b.

Rae, J. W., Potapenko, A., Jayakumar, S. M., and Lillicrap, T. P. Compressive transformers for long-range sequence modelling. *ICLR*, 2020.

Reid, M., Savinov, N., Teplyashin, D., Lepikhin, D., Lillicrap, T., Alayrac, J.-b., Soricut, R., Lazaridou, A., Firat, O., Schrittwieser, J., et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Shang, Y., Cai, M., Xu, B., Lee, Y. J., and Yan, Y. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. *arXiv preprint arXiv:2403.15388*, 2024.

Sun, Q., Cui, Y., Zhang, X., Zhang, F., Yu, Q., Luo, Z., Wang, Y., Rao, Y., Liu, J., Huang, T., et al. Generative multimodal models are in-context learners. *arXiv preprint arXiv:2312.13286*, 2023.

Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Xiao, G., Tian, Y., Chen, B., Han, S., and Lewis, M. Efficient streaming language models with attention sinks. *ICLR*, 2024.

Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., and Chen, E. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023.

Zhang, H., Li, X., and Bing, L. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023a.

Zhang, P., Wang, X. D. B., Cao, Y., Xu, C., Ouyang, L., Zhao, Z., Ding, S., Zhang, S., Duan, H., Yan, H., et al. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*, 2023b.

Zong, Y., Mac Aodha, O., and Hospedales, T. Self-supervised multimodal learning: A survey. *arXiv preprint arXiv:2304.01008*, 2023.

Zong, Y., Bohdal, O., and Hospedales, T. VL-ICL Bench: The devil in the details of benchmarking multimodal in-context learning. *arXiv preprint arXiv:2403.13164*, 2024.

A. Appendix

We present the breakdown results of LongBench (Bai et al., 2023b) in Table 3, and raw results of Figure 1 in Table 4 to 8.

Table 3. Breakdown of LongBench Performance.

Models	Single-Doc QA		Multi-Doc QA		Summarization		Few-shot Learning			Synthetic		Code		Avg.
	Qspr.	MulFi	HQA	WMQA	GRpt	MulN	TREC	TriQA	SMSM	PsgC	PsgR	Lcc	Repo	
Llama2-Chat-7B	22.26	37.34	27.83	31.56	26.90	26.38	64.50	83.34	41.42	2.84	6.50	58.54	52.15	37.86
VILA-7B	25.87	38.21	39.51	36.26	27.97	16.09	56.00	84.45	41.95	6.33	8.67	64.18	48.88	37.08
Vicuna-7B	23.30	38.27	21.36	17.40	27.65	26.75	66.00	82.53	40.66	2.00	5.00	61.53	47.79	34.95
LLaVA-v1.5-7B	16.47	41.51	23.93	19.58	33.08	25.55	55.67	85.07	39.20	5.33	9.00	52.73	42.98	34.32
InternLM2-Chat-1.8B	35.77	48.94	46.04	33.37	26.97	21.90	63.67	82.06	36.16	6.67	15.33	56.29	50.90	39.07
InternLM-XComposer2-1.8B	27.83	43.07	46.83	38.36	28.11	23.15	60.00	84.88	34.32	8.48	18.33	55.97	51.98	38.25
QwenLM-7B	23.41	30.07	27.10	16.79	27.76	21.03	62.67	86.01	41.54	6.82	11.86	71.03	55.41	36.62
Qwen-VL-7B	18.06	24.50	13.92	12.20	30.74	23.86	61.67	64.26	39.72	4.33	11.67	39.62	31.46	29.63
Phi-3B	41.49	52.17	50.93	40.61	32.65	23.56	57.33	87.35	36.42	8.00	81.67	62.93	56.29	46.39
Phi-3B-Vision	43.39	49.69	46.25	34.06	31.55	24.60	52.33	85.21	38.71	5.67	53.33	47.96	38.56	37.86

Table 4. Detailed results on VL-ICL TextOCR subset.

Model	0-Shot	1-Shot	2-Shot	4-Shot	8-Shot	16-Shot	32-Shot	64-Shot
VILA-2.7B	0.50	8.00	12.00	17.50	18.00	19.00	16.00	17.50
VILA-2.7B-PruMerge	3.00	8.00	9.00	12.00	12.50	11.50	13.00	14.50
VILA-2.7B-SelfExtend	0.50	8.00	11.00	13.50	17.00	22.00	23.50	23.50
VILA-2.7B-SelfExtend-PruMerge	0.50	5.50	11.00	14.00	18.50	22.00	23.50	22.00
VILA-2.7B-ours	2.50	10.00	15.50	21.50	21.00	16.00	17.50	16.00
VILA-7B	28.00	6.00	21.50	26.50	31.00	31.50	32.50	30.50
VILA-7B-PruMerge	27.00	10.50	23.50	29.00	28.00	34.50	32.50	30.50
VILA-7B-SelfExtend	28.00	6.50	20.50	25.50	27.50	35.50	33.50	33.50
VILA-7B-SelfExtend-PruMerge	27.00	5.00	22.50	29.50	32.50	32.50	35.50	33.50
VILA-7B-Ours	14.50	15.50	20.00	22.50	21.50	19.50	17.50	26.00

Table 5. Detailed results on VL-ICL CLEVR subset.

Model	0-Shot	1-Shot	2-Shot	4-Shot	8-Shot	16-Shot	32-Shot	64-Shot
VILA-2.7B	0.00	29.50	30.00	29.50	29.50	28.50	27.00	30.00
VILA-2.7B-PruMerge	0.00	24.00	27.00	32.00	29.50	38.00	29.00	32.00
VILA-2.7B-SelfExtend	0.00	30.00	32.50	31.00	28.50	32.00	32.50	30.00
VILA-7B	4.00	45.00	37.00	32.00	32.00	30.50	34.50	31.50
VILA-7B-PruMerge	2.00	27.50	30.50	28.50	29.50	30.00	34.50	33.00
VILA-7B-SelfExtend	4.00	34.50	37.00	34.50	28.50	31.50	26.00	29.50
VILA-7B-Ours	2.50	15.50	20.00	25.50	29.50	30.50	31.50	30.00

Table 6. Detailed results on VL-ICL Operator Induction subset.

Model	0-Shot	1-Shot	2-Shot	4-Shot	8-Shot
VILA-2.7B	16.67	13.33	16.67	11.67	11.67
VILA-2.7B-PruMerge	6.67	15.00	13.33	10.00	20.00
VILA-2.7B-SelfExtend	16.67	13.33	18.33	11.67	13.33
VILA-2.7B-Ours	10.00	16.67	13.33	15.00	17.00
VILA-7B	28.33	13.33	1.67	8.33	0.00
VILA-7B-PruMerge	23.33	11.67	6.67	6.67	6.67
VILA-7B-SelfExtend	28.33	11.67	3.33	6.67	5.00
VILA-7B-Ours	3.33	15.00	25.00	20.00	21.67

Table 7. Detailed results on VL-ICL 2-way Open MiniImageNet subset.

Model	0-Shot	1-Shot	2-Shot	4-Shot	5-Shot
VILA-2.7B	0.00	6.00	49.00	49.00	43.00
VILA-2.7B-PruMerge	0.00	5.00	38.00	45.00	47.00
VILA-2.7B-SelfExtend	0.00	4.50	29.50	41.50	41.50
VILA-2.7B-Ours	0.00	3.50	34.00	39.50	48.50
VILA-7B	0.00	31.50	45.50	45.00	42.00
VILA-7B-PruMerge	0.00	34.50	43.00	47.00	55.00
VILA-7B-SelfExtend	0.00	27.50	55.50	42.00	46.50
VILA-7B-Ours	0.00	21.00	61.50	77.00	71.00

Table 8. Detailed results on VL-ICL 5-way Open MiniImageNet subset.

Model	0-Shot	1-Shot	2-Shot	4-Shot	5-Shot
VILA-2.7B	0.00	5.00	14.50	14.00	12.50
VILA-2.7B-PruMerge	0.00	5.00	24.50	22.50	17.50
VILA-2.7B-SelfExtend	0.00	3.00	21.50	18.00	18.50
VILA-2.7B-Ours	0.00	19.50	21.00	22.00	22.00
VILA-7B	0.00	15.00	9.50	15.00	7.50
VILA-7B-PruMerge	0.00	9.50	15.00	25.50	26.50
VILA-7B-SelfExtend	0.00	11.50	13.50	15.00	20.00
VILA-7B-Ours	0.00	16.00	57.00	57.50	55.00