

V2PE: Improving Multimodal Long-Context Capability of Vision-Language Models with Variable Visual Position Encoding

Junqi Ge^{1,4*}, Ziyi Chen^{1,4*}, Jintao Lin^{3,4*}, Jinguo Zhu^{4*},
 Xihui Liu³, Jifeng Dai^{1,4}, Xizhou Zhu^{1,2,4†}

¹Tsinghua University ²SenseTime Research ³University of Hong Kong
⁴Shanghai AI Laboratory

Abstract

Vision-Language Models (VLMs) have shown promising capabilities in handling various multimodal tasks, yet they struggle in long-context scenarios, particularly tasks involving videos, high-resolution images, or lengthy image-text documents. In our work, we first conduct an empirical analysis of VLMs’ long-context capabilities using our augmented long-context multimodal datasets. Our findings reveal that directly applying the positional encoding mechanism used for textual tokens to visual tokens is suboptimal, and VLM performance degrades sharply when the position encoding exceeds the model’s context window. To address this, we propose Variable Visual Position Encoding (V2PE), a novel positional encoding approach that employs variable and smaller increments for visual tokens, enabling more efficient management of long multimodal sequences. Our experiments demonstrate the effectiveness of V2PE in enhancing VLMs’ ability to effectively understand and reason over long multimodal contexts. We further integrate V2PE with our augmented long-context multimodal datasets to fine-tune the open-source VLMs. The fine-tuned model achieves strong performance on both standard and long-context multimodal tasks. Notably, when the sequence length of the training dataset is increased to 256K tokens, the model is capable of processing multimodal sequences up to 1M tokens, highlighting its potential for real-world long-context applications. We shall release the code, model weights, and datasets to facilitate further research.

1. Introduction

With the rapid advancement of Large Language Models (LLMs) [14, 29, 83, 99, 108], Vision-Language Models (VLMs) have made substantial strides [19, 82, 103, 115], excelling at tasks like visual captioning [17], visual question answering [78], and complex visual reasoning [136]. Despite this progress, existing research [112, 119, 132, 144]

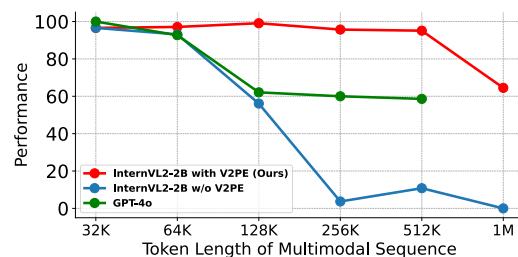


Figure 1. Performance on the image retrieval task using a token length of up to 1M on MM-NIAH [119] across different VLMs. The GPT-4o [81] version is 2024-08-06, while InternVL2-2B [20] models are both fine-tuned from the official release, using our augmented data, with token lengths reaching up to 256K per sample.

reveals that VLMs struggle to generalize effectively when confronted with multimodal long-sequence inputs (e.g., long videos [121, 126], high-resolution images [45, 122], and lengthy image-text documents [3, 105, 151]). This limitation emerges even in relatively straightforward, such as object counting and passkey duplication, significantly restricting VLMs’ potential applications and impeding the enhancement of user experience [22, 114, 119, 120].

Recent efforts have attempted to extend VLMs’ capabilities to process multiple images or handle long multimodal sequences. However, these approaches either permit only a small number of images (typically fewer than five) [52, 65, 149] or primarily target video data (e.g., LongVA [144], LongVILA [132], LongLLAVA [120], VideoXL [93]). These studies are only limited to specific application scenarios, highlighting the challenges VLMs face in handling complex and long-sequence multimodal data, which makes a key research question particularly urgent: *Why do VLMs perform poorly in long-context scenarios, and how can we unlock their capacity for comprehensive multimodal understanding and reasoning over long sequences?*

To investigate this, we first construct a large-scale pool of long-context multimodal datasets to evaluate and analyze VLM capabilities systematically. By extending the se-

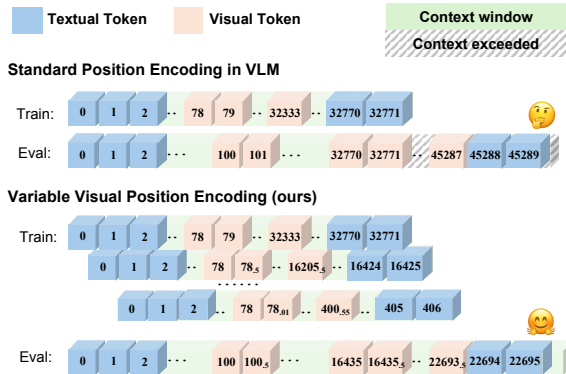


Figure 2. Illustration from our proposed Variable Visual Position Encoding (V2PE). Unlike the standard position encoding used in most VLMs, which shares the same stepwise positional increment for both visual and textual tokens, our proposed Variable Visual Positional Encoding (V2PE) uses smaller and variable positional increments specifically for visual tokens compared to textual tokens. This flexible design enables VLMs to handle long-context multimodal sequences within a limited context window.

quence length of existing instruction-tuning datasets (e.g., DocVQA [78], ChartQA [77], SQA [72]) to 32K-64K tokens, we adapt these datasets specifically to train and validate VLMs for enhanced long-context capabilities. We also extend long-context benchmarks from 64K to 1M tokens (e.g., by employing the official MM-NIAH [119] curation code) to validate VLM performance with longer contexts, providing a deeper understanding of the limitations.

Our empirical analysis shows that the design of positional encodings of visual tokens plays an essential role in long-context scenarios, which is often overlooked in previous studies. Specifically, (1) directly applying the LLM positional encoding mechanism to visual tokens is suboptimal, and (2) the performance of VLMs degrades significantly when positional encodings for visual tokens exceed the trained context window, and previous position encoding extension methods applicable to LLMs provide only marginal improvements. These findings highlight the need for specialized positional encoding methods to manage visual tokens in long-context multimodal scenarios effectively.

To address these challenges, we propose Variable Visual Position Encoding (V2PE), a novel approach for handling visual token positions in VLMs. Considering the continuity nature of pixel space, adjacent visual tokens exhibit greater similarity than adjacent text tokens. Thus, V2PE uses smaller positional increments for visual tokens than text tokens (see Fig. 2). Furthermore, V2PE employs variable positional increments for visual tokens during training, enabling the model to learn and adapt to position encoding in various scenarios. This variable adjustment allows the model to effectively handle different numbers and complexities of image inputs during inference, thereby enhancing its stability and adaptability in long-context processing.

In experiments, we apply V2PE to enhance the long-context capability of open-source VLMs [19, 20], and fine-tune them using our extended multimodal datasets. The resulting models not only maintain strong performance on standard short-context multimodal benchmarks, but also excel in tasks requiring long-context handling, which outperform traditional token compression and other position encoding extension methods. In particular, after further fine-tuning on multimodal sequences up to 256K tokens, our model achieves promising performance in multimodal retrieval tasks involving sequences as long as 1M tokens, as shown in Fig. 1. The main contributions of this paper are as follows:

- We construct mixed datasets to improve VLMs’ long-context capability by augmenting existing multimodal instruction tuning datasets and conduct a thorough investigation into why current VLMs struggle with long-context multimodal inputs, revealing that directly applying LLM positional encoding to visual tokens is ineffective.
- We propose Variable Visual Position Encoding (V2PE), a novel positional encoding strategy that employs variable and smaller increments for visual tokens, significantly enhancing VLMs’ ability to understand and reason over extended multimodal contexts.
- We apply our V2PE method and extend training data on the open-source VLMs. The fine-tuned VLMs perform exceptionally well on both general multimodal benchmarks and long-context multimodal tasks, with the capacity to handle sequences of up to 1M tokens, significantly longer than training sequence.

2. Related Works

Vision-Language Models (VLMs). With the development of Large Language Models (LLMs) [6, 10, 14, 36, 83, 99, 101, 104, 107, 108, 123, 124, 134, 148], the integration of vision and language modalities is significantly catalyzed. This progress gives rise to Vision-Language Models (VLM), which can perceive visual contents, conduct visual reasoning, and engage in multi-modal dialogue with humans. Both proprietary commercial VLMs [5, 51, 79, 82, 95, 102, 103, 129] and open-source VLMs [46, 73, 75, 106, 113] have witnessed this significant evolution. For commercial entities, GPT-4V [82] incorporates visual inputs to extend GPT-4 [83] with capabilities of handling multi-modal content, while Google’s Gemini series [102, 103] are able to process 1 million multi-modal tokens with significant performance. For open-source initiatives, BLIP series [27, 58, 59], LLaVA series [63–65], Qwen-VL series [7, 115], MiniCPM-V series [135], InternVL series [19, 20, 39], and others [9, 25, 26, 34, 35, 48, 61, 71, 86, 117, 140, 142] have also impacted the AGI landscape in the research community by linking large language models [6, 14, 107] and large vision models [4, 19, 30, 89] in processing both vi-

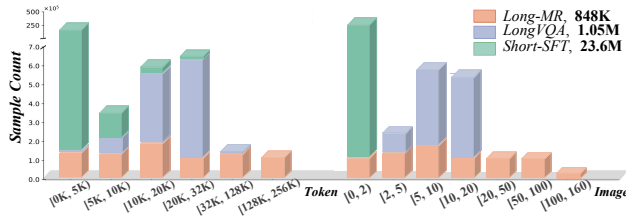


Figure 3. Statistics of our mixed training dataset (Long-MR, Long-VQA, and short SFT from InternVL2). The left and right figures illustrate the distributions of token counts and image counts per training sample, respectively.

sual and textual modality. Specifically, there are also some works advance VLMs’ power by improving the scale and quality of data [15, 42, 68], training on high-resolution images [45, 122] or optimizing vision foundation models [19, 139].

Long Context Modeling. Due to practical demands, a large body of research [13, 21, 47, 67, 69, 143, 147] has been developed to extend the context length of LLMs. One mainstream direction is to increase the length of LLM’s context window by extrapolation of position encoding [11, 16, 32, 85, 87, 96, 97, 146, 150], which allows the LLM to process unseen positions during inference. Another line of research focuses on alleviating the complexity of self-attention by using sparse attention [8, 18, 23, 31, 88, 138], linear-complexity state-space modules [40, 41], or approximate attention computation [24, 55, 91, 116] as the alternatives for dense global attention. The context can also be extended by utilizing external memory to retrieve relevant tokens from the compression of past inputs, such as chunked input [80, 118, 131, 141] or cached KV activations [2, 49, 109, 127, 128]. Other works [43, 127, 130] also adopt sliding windows to achieve an infinite context by only maintaining the activations for the very first and the latest tokens. Recent efforts [52, 120, 132, 144] aim to improve VLMs’ ability to process multiple images and long multimodal sequences. However, these studies are often limited to specific applications, revealing challenges faced by VLMs in processing complex and long-context data.

Position Encoding in Transformer. To address the lack of sequential information inherent in self-attention mechanisms, Transformers [110] utilize position encoding as a fundamental component to provide position information. The common position encoding design can be broadly categorized into absolute position encoding (APE) and relative position encoding (RPE). The absolute position encoding [12, 53, 56, 110, 111, 145] is simple and intuitive, as it embeds each absolute position position into a position vector which is then added into the representation of input sequences, but this strategy cannot be applied effectively for long sequences. To improve the modeling of long-term dependencies, relative positional encod-

ing [28, 74, 87, 90, 92, 98, 100, 133] focuses on utilizing distance between tokens as the position information. Some research works [44, 87, 100] aim to utilize relative position for length extrapolation to train Transformers on short sequences and inference on longer contexts. Another research direction is Rotary Position Encoding (RoPE), which embeds position information by rotating the key-query products according to their relative distances. Methods like Position Interpolation [16], NTK-Aware scaling [11], and LongRoPE [32] make progress on top of RoPE [98] to improve its poor extrapolation capability for longer sequences. For VLMs, most studies [7, 20, 132] adopt the same position embedding methods used in LLMs, and there are also some works [38, 76, 125] to explore different diagrams of the positional encoding schemes.

3. Method

3.1. Augmented Long-context Multimodal Datasets

We introduce two augmented long-context multimodal datasets: *Long Visual Question Answering* and *Long multimodal Retrieval*. These datasets aim to enhance VLMs’ long-context training and establish a systematic evaluation framework, thereby addressing the challenges associated with long-context understanding that extend beyond the scope of existing training data.

Long Visual Question Answering (Long-VQA). The Long-VQA dataset aims to evaluate the capabilities of VLMs in understanding and reasoning over long multimodal sequences within general visual question-answering tasks. We extended 17 widely adopted datasets (e.g., DocVQA [78], GQA [50], SQA [72]), expanding their context from short sequences to those containing up to 64K tokens. The tasks involve answering questions that require commonsense reasoning, factual knowledge, and interpretation of visual information from charts, documents, and real-world texts.

To extend existing datasets, we interleaved images from multiple samples into a single input, increasing sequence length and complexity to simulate real-world scenarios involving extraneous information. To reduce ambiguity, we refined questions to be more specific, incorporating instructions like “Based on page n , answer the following question”, which helps models focus on relevant content.

Long-VQA contains 533K samples: 392K for training (up to 32K tokens) and 141K for validation (up to 64K tokens) to evaluate the generalization to longer contexts.

Long Multimodal Retrieval (Long-MR). Inspired by MM-NIAH (Multimodal Needle-in-a-Haystack) [119], we developed Long-MR by inserting a target image or textual segment into sequences of interleaved images and texts. Long-MR evaluates VLMs’ ability to retrieve specific targets from ultra-long multimodal sequences, requiring mod-

els to locate the inserted “needle” and answer associated questions. We generated two subsets of Long-MR: Long-MR-32K (488K samples, sequences up to 32K tokens) and Long-MR-256K (50K samples, sequences up to 256K tokens), following the data construction process of MM-NIAH.

To assess the limitation of VLMs’ long-context capabilities, we further extended MM-NIAH evaluation benchmark following the official code[84], generating testing samples with sequence lengths ranging from 64K to 1M tokens, named after MM-NIAH_{1M} benchmark. This pushes the testing capacity beyond the original MM-NIAH, which is limited to sequences up to 64K tokens.

We combined the training splits of Long-VQA and Long-MR with short-context instruction-tuning datasets, as utilized in InternVL2, to create a mixed training set for our experiments. Fig. 3 illustrates the distribution of token sequence lengths and the number of images in the mixed training dataset, highlighting our focus on long-context samples.

More details about the construction process, examples, and statistics of the datasets are provided in Appendix Sec. C.

3.2. Variable Visual Position Encoding

Position Encoding in Vision-Language Models. Position encoding is essential in transformer architectures, enabling models to capture sequential relationships by providing tokens with positional information. It usually involves two sequential steps: *Position Index Derivation* f_{pos} , which assigns a positional index p_i to each token x_i , and *Position Embedding Computation* g_{emb} , which transforms these indices into position embeddings that influence the attention mechanism.

Formally, the multimodal input in VLMs can be represented as a sequence of N interleaved textual and visual tokens:

$$\mathbf{X} = [x_0, x_1, \dots, x_{N-1}], \quad (1)$$

where each token x_i is either a textual token x_i^{txt} or a visual token x_i^{vis} .

The Position Index Derivation function f_{pos} is recursively defined to capture the sequential nature of token positions:

$$p_i = \begin{cases} 0, & \text{if } i = 0, \\ f_{\text{pos}}(p_{i-1}, x_i), & \text{for } i = 1, 2, \dots, N - 1. \end{cases} \quad (2)$$

In existing LLMs and VLMs, the position index increments uniformly by 1 for each token, regardless of its modality:

$$p_i = p_{i-1} + 1, \quad \text{for } i = 1, 2, \dots, N - 1. \quad (3)$$

The Position Embedding Computation g_{emb} then transforms these position indices into embeddings. VLMs typically adopt the same position embedding methods used in

Large Language Models (LLMs), such as Relative Position Encoding [92] or Rotary Position Embedding (RoPE) [98]. These embeddings are incorporated into the token representations at each layer to provide positional context during attention computations. For instance, in RoPE encoding, the token representation \mathbf{h}_i integrates positional information as:

$$\mathbf{h}'_i = \mathbf{h}_i \otimes g_{\text{emb}}(p_i), \quad (4)$$

where \otimes represents element-wise multiplication, and \mathbf{h}'_i is subsequently used as query or key embeddings in the Transformer attention mechanism.

Variable Position Index Derivation. The uniform increment of position indices in current VLMs does not account for the differences in information complexity and redundancy between textual and visual tokens. Visual tokens often exhibit higher redundancy and greater similarity with adjacent tokens, suggesting they may require smaller positional increments than textual tokens. Moreover, the large number of visual tokens can cause position indices to exceed the model’s pre-trained context window, leading to degraded performance.

To address these issues, we propose a modality-specific recursive function for position index derivation, assigning position indices differently for textual and visual tokens:

$$p_i = p_{i-1} + \begin{cases} 1, & \text{if } x_i \text{ is a textual token,} \\ \delta, & \text{if } x_i \text{ is a visual token,} \end{cases} \quad (5)$$

where δ is a smaller increment ($\delta < 1$) that reduces the rate at which position indices increase for visual tokens. The standard increment of 1 is retained for textual tokens to maintain their positional distinctions.

During training in our experiments, δ is randomly selected for each image from a set of fractional values:

$$\delta \in \Delta = \left\{ 1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}, \frac{1}{64}, \frac{1}{128}, \frac{1}{256} \right\}. \quad (6)$$

Note that, δ remains constant within a single image to preserve the relative positional relationships among its visual tokens. For inputs containing multiple images, δ can be independently chosen for each image.

During inference, δ can be flexibly selected based on the input sequence length, allowing us to balance task performance and ensure that position indices remain within the model’s valid context range. Specifically, for long input sequences, a smaller δ can be employed to control the increase in position indices, preventing from exceeding the trained positional embedding range, as illustrated in Fig. 2.

Discussion and Comparison with Previous Methods. Our Variable Visual Position Encoding (V2PE) offers several advantages over existing long-context methods:

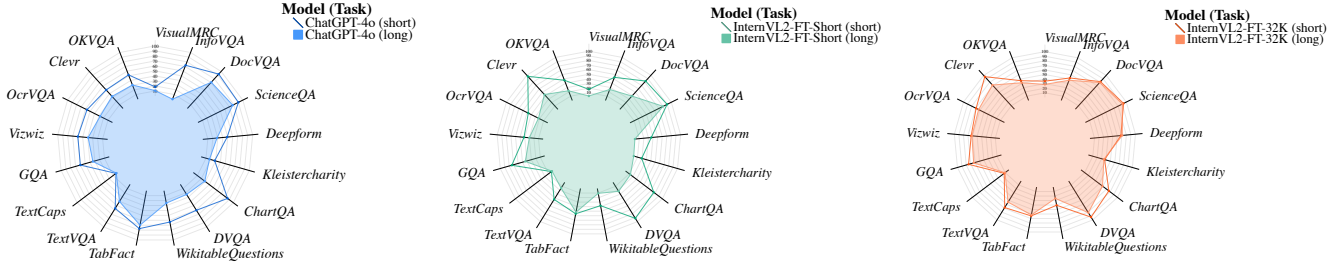


Figure 4. Performance on Long-VQA (long) with sequence lengths up to 32K tokens and its corresponding standard VQA (short) benchmarks. Compared with GPT-4o and InternVL2-FT-Short, the InternVL2-FT-32K, enhanced with our proposed Long-VQA dataset, effectively narrows the performance gap between short and long-context tasks.

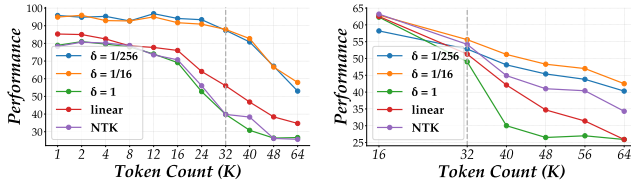


Figure 5. Performance on image retrieval task in MM-NIAH (left) and QA task in Long-VQA (right) with different positional increments. Additionally, we include the results of InternVL2-FT-32K when utilizing linear interpolation (linear) [16] and NTK-Aware Scaled RoPE (NTK) [11] position encoding extension methods.

1) Unlike approaches that reduce the number of visual tokens through attention pooling or feature pooling, which potentially result in information loss, V2PE retains all visual tokens within the VLM, preserving the richness and granularity of visual content.

2) Position encoding extension methods (*e.g.*, Positional Interpolation [16] and Extrapolation [96]) adjust position embeddings during inference to accommodate longer sequences. However, extrapolation extends position indices beyond the model’s trained range, while interpolation methods (linear [16] / NTK-aware [11]) face performance degradation with unseen position increment, potentially requiring additional fine-tuning [85]. In contrast, V2PE allows the VLM to adapt position indices with arbitrary intervals while evaluating zero-shot in longer contexts by dynamically selecting δ during training. This strategy avoids unexpected position embeddings, ensures consistent model performance across varying input lengths and avoids additional fine-tuning in longer contexts.

4. Experiment

4.1. Analysis of VLMs’ Long-Context Capabilities

We first analyze the long-context capabilities of existing VLM by using our constructed long-context multimodal datasets (see Sec. 3.1). Note that this sub-section does not use V2PE.

Experimental Setup. We fine-tune InternVL2-2B on two different training datasets and obtain four finetuned models: 1) *InternVL2-FT-Short*: This model is fine-tuned on

Table 1. Performance on Long-VQA task with sequence lengths up to 64K tokens at various positional increments δ .

Model	δ	16K	32K	40K	48K	56K	64K	Avg
InternVL2-2B	—	52.8	27.3	23.0	23.9	22.6	21.0	28.4
InternVL2-FT-32K	—	63.0	56.1	38.3	35.5	34.6	30.6	43.0
InternVL2-V2PE-32K	1/256	58.2	52.8	48.1	45.4	43.8	40.3	48.1
	1/128	60.0	53.7	49.5	46.5	45.3	41.3	49.4
	1/64	61.0	54.5	50.1	47.4	45.9	42.2	50.2
	1/32	61.9	55.4	51.4	47.5	46.4	42.3	50.8
	1/16	62.9	55.6	51.2	48.3	47.0	42.5	51.3
	1/8	63.2	55.8	52.0	47.6	45.8	41.9	51.1
	1/4	63.4	55.7	51.3	47.0	45.1	39.7	50.4
	1/2	63.3	54.9	46.8	36.8	33.3	29.2	44.1
1/1	62.3	49.0	30.0	26.5	27.0	25.9	36.8	
InternVL2-FT-32K (Linear [16])	—	62.5	51.3	42.1	34.7	31.4	25.9	40.9
InternVL2-FT-32K (NTK [11])	—	63.2	54.2	44.9	41.0	40.4	34.3	45.3

the original instruction-tuning dataset (short-context) used by InternVL2-2B. It serves as a baseline to evaluate performance without exposure to long-context data. 2) *InternVL2-FT-32K*: This model is fine-tuned on a mixed training dataset introduced in Sec. 3.1, incorporating our augmented long-context training data to enhance its long-context capacity. 3) *InternVL2-32K-s1/16*: To explore the effect of positional encoding increments on visual tokens, we reduce the positional increment to 1/16 for visual tokens, holding it fixed across training and inference, while maintaining training on the same mixed dataset as InternVL2-FT-32K. 4) *InternVL2-32K-s1/256*: Similarly, we further reduce the positional increments to 1/256 and keep it fixed to examine its impact. Notably, the positional increments for this model also remain fixed during training, without using V2PE.

Evaluation Benchmarks. All models are evaluated on the Long-VQA validation split and the MM-NIAH benchmark. Additionally, performance on the standard multimodal VQA benchmarks is assessed if needed.

Effectiveness of Augmented Long-Context Multimodal Data. We compare the performance of InternVL2-FT-Short and InternVL2-FT-32K on both the Long-VQA validation set and standard short-context benchmarks, as shown in Fig. 4. The results reveal that InternVL2-FT-Short, which is trained exclusively on short-context data, performs competitively on standard benchmarks but suffers significant degradation on long-context tasks. Even advanced models, such as GPT-4o, exhibit notable declines when the input to-

Table 2. Performance on standard VQA benchmarks with different positional increment δ . The highest score is marked in **bold**.

Model	δ	ChartQA	DocVQA	A12D	InfoVQA	SQA	POPE	MMMU _{val}	MMBench _{EN}	SEED _I	Avg
InternVL2-2B	—	71.7	86.9	74.1	58.9	94.1	85.2	36.3	73.4	70.9	72.4
InternVL2-FT-Short	—	76.0	85.7	74.0	57.1	94.0	88.9	34.8	73.0	66.6	72.2
InternVL2-FT-32K	—	75.6	84.8	73.6	56.5	94.7	88.7	35.3	73.8	65.1	72.0
InternVL2-V2PE-32K	1/256	76.2	81.6	72.3	51.7	94.3	88.0	35.7	73.2	71.0	71.6
	1/128	76.5	82.3	72.6	52.9	94.2	88.1	34.7	73.5	70.9	71.7
	1/64	76.4	83.0	72.6	53.6	94.4	88.0	35.4	73.4	70.8	72.0
	1/32	76.9	83.6	72.6	54.3	94.3	88.1	35.7	73.3	71.0	72.2
	1/16	77.2	84.3	73.0	55.2	94.5	88.1	35.7	73.3	71.0	72.5
	1/8	77.0	84.6	73.1	55.4	94.6	88.1	36.0	72.9	71.0	72.5
	1/4	77.3	84.7	73.3	56.5	94.6	88.2	36.0	73.6	71.0	72.8
	1/2	76.8	84.9	73.5	56.0	94.9	88.7	35.8	73.8	71.2	72.9
	1/1	76.6	84.9	73.7	55.2	94.8	88.4	36.1	73.7	71.1	72.7

Table 3. Performance on the image retrieval task in MM-NIAH with token lengths up to 1M at various position increments δ . Test samples exceeding 64K tokens are from our extended MM-NIAH_{1M}. The designation “NA” indicates that the GPT-4 results are not applicable due to the extremely long context length.

Model	δ	1K	16K	32K	64K	128K	256K	512K	1M
InternVL2-2B	—	23.9	33.7	28.7	26.0	17.3	21.8	5.3	0.0
GPT-4o	—	100.0	90.4	100.0	92.7	62.1	60.0	58.6	NA
InternVL2-V2PE-32K	1/256	100.0	84.7	79.3	81.5	61.2	56.9	36.7	6.0
	1/64	96.0	83.9	78.1	79.5	43.6	39.3	23.4	3.4
	1/16	96.0	82.4	73.1	65.4	57.6	41.3	19.8	0.0
	1/4	96.0	84.2	76.2	57.6	48.9	21.5	2.8	0.0
	1/1	83.0	61.2	42.0	27.5	25.5	1.5	0.0	0.0
InternVL2-V2PE-256K	1/256	96.0	97.1	96.6	97.1	99.1	95.7	95.1	64.5
	1/64	96.0	94.7	96.6	97.1	97.1	94.1	94.1	62.8
	1/16	96.0	94.7	93.1	94.2	100	96.2	92.9	45.1
	1/4	96.0	94.4	100.0	97.1	75.5	75.2	38.3	2.5
	1/1	91.0	92.8	96.6	93.1	56.1	3.7	10.8	0.0
InternVL2-FT-32K with linear interpolation	—	85.3	76.0	56.0	34.7	26.4	21.9	16.3	0.5
InternVL2-FT-32K with NTK interpolation	—	78.3	70.7	39.8	25.7	12.3	16.8	14.8	0.0

ken count is increased, underscoring a common limitation in handling extended sequences. In contrast, InternVL2-FT-32K, trained on the mixed dataset including long-context data, effectively narrows the performance gap between short and long-context tasks. This suggests it is difficult for short-context VLMs to inherently generalize to long-context multimodal tasks without further finetuning. Targeted exposure to long-context data is crucial for achieving robust performance across varying input lengths.

Impact of Visual Position Encoding. To assess the impact of position encoding increments for visual tokens, we evaluate the models InternVL2-FT-32K, InternVL2-32K-s1/16, and InternVL2-32K-s1/256 on both the Long-VQA validation set and MM-NIAH benchmark.

Experimental results, presented in Fig. 5, indicate that InternVL2-FT-32K continues to experience performance degradation as input token sequences grow up to 64K tokens, despite being trained on sequences up to 32K tokens. When the input sequence length exceeds the maximum token count during training, performance declines sharply. Even applying positional encoding extension techniques developed for large language models (LLMs) give limited benefits. However, models with reduced positional increments for visual tokens show significant improvements. Both InternVL2-32K-s1/16 and InternVL2-

32K-s1/256 maintain stable performance. We hypothesize that reducing the position increments by factors of 16 and 256 effectively prevents the position indices of visual tokens from exceeding the model’s trained context window, thereby mitigating performance decline.

4.2. Effectiveness of Our Proposed V2PE

Experimental Setup. We fine-tune the InternVL2-2B model using our proposed V2PE method in a two-stage training procedure. 1) In the first stage, we fine-tune the released InternVL2-2B model on our mixed training dataset, incorporating V2PE. During training, as described in Eq. 6, the positional increment δ is randomly selected, enabling the model to adapt to varying positional increments for visual tokens flexibly. The model obtained after this stage is denoted as *InternVL2-V2PE-32K*. 2) In the second stage, we further fine-tune the model on the Long-MR-256K dataset, with sequence lengths up to 256K tokens. To preserve the model’s general performance in shorter context, we retain 50% of the data from the first stage. To optimize memory usage, we apply the Ring Attention [66], enabling model parallelism across multiple GPUs. The model obtained after the second stage is denoted as *InternVL2-V2PE-256K*.

Evaluation Benchmarks. We evaluate our trained models on various long-context and standard benchmarks, in-

Table 4. Comparison with existing MLLMs on general MLLM benchmarks. “#Param” denotes the number of parameters. The designation “—” indicates that the corresponding score is not released. Numbers in gray are results evaluated by ourselves. δ is fixed 1/2 during testing.

Model	#Param	ChartQA	DocVQA	AI2D	InfoVQA	SQA	POPE	MMMU _{val}	MMBench _{EN}	SEED _I	Avg
InternVL2-2B [20]	2.0B	71.7	86.9	74.1	58.9	94.1	85.2	36.3	73.4	70.9	72.4
InternVL2-2B-FT32K	2.0B	75.6	84.9	73.7	56.5	94.7	88.5	35.3	73.4	70.9	71.1
DeepSeek-VL-1.3B [71]	2.0B	47.4	36.5	51.5	20.6	68.4	85.9	33.8	66.4	66.0	52.9
Qwen2-VL-2B [115]	2.0B	73.5	90.1	74.7	65.5	77.7	88.2	41.1	74.9	72.9	73.1
Aquila-VL-2B [42]	2.2B	32.0	85.0	75.1	58.3	95.1	83.1	46.9	79.0	73.9	69.8
MiniCPM-V-2 [135]	2.8B	55.6	71.9	62.9	36.0	80.7	86.3	38.2	64.1	67.1	62.5
Vintern-3B-beta [33]	3.7B	68.3	78.1	69.1	52.2	75.0	87.4	46.7	70.6	70.0	68.6
Llama 3.2 11B [108]	11B	83.4	88.4	91.1	—	—	—	50.7	68.0	—	—
Qwen2-VL-72B [115]	73B	88.3	96.5	88.1	84.5	91.2	87.2	64.5	86.9	77.9	85.0
GPT-4o [81]	—	85.7	92.8	84.7	—	90.1	97.2	69.1	82.1	76.7	—
InternVL2-V2PE-32K ($\delta = 1/2$)	2.0B	76.8	84.9	73.5	56.0	94.9	88.7	35.8	73.8	71.2	72.9

Table 5. Comparison with existing MLLMs on long context MLLM benchmarks. “#Param” denotes the number of parameters. The designation “—” indicates that the corresponding score is not released. Numbers in gray are results evaluated by ourselves. δ is fixed 1/256 during testing.

Model	#Param	MM-NIAH			Milebench				VideoMME
		Image	Text	Avg	T	S	NI	Avg	
InternVL2-2B [20]	2.0B	23.0	18.9	21.0	58.2	54.5	37.0	49.9	48.2
InternVL2-2B-FT32K	2.0B	73.0	81.1	77.05	63.6	56.0	96.2	71.9	47.8
Phi-3-Vision [1]	2.7B	24.9	26.3	25.6	46.9	50.0	73.8	56.9	40.7
LongLLaVA [120]	9B	28.8	57.1	43.0	47.3	46.8	50.0	48.0	43.7
LongLLaVA [120]	13B	27.7	55.7	41.7	52.7	52.1	50.0	51.6	51.6
VILA1.5 [62]	13B	5.4	11.0	8.2	21.4	48.5	7.7	25.9	47.4
LongVILA [132]	7B	33.7	34.1	33.9	65.3	55.1	91.9	70.8	60.1
Gemini-1.5 [103]	—	28.5	82.1	55.2	50.2	58.3	97.9	68.8	69.6
GPT-4o [81]	—	52.3	75.0	63.7	56.2	63.5	99.8	73.2	64.7
Claude3-Opus [5]	—	—	—	—	37.4	48.1	85.3	56.9	59.7
InternVL2-V2PE-32K ($\delta = 1/256$)	2.0B	78.1	85.5	81.8	65.3	55.1	96.7	72.4	52.3

cluding the Long-VQA validation split, MM-NIAH [119] (including our augmented MM-NIAH_{IM}), and widely used VQA benchmarks such as ChartQA [77], DocVQA [78], AI2D [54], InfoQA [54], SQA [72], POPE [60], MMMU [137], MMBench_{EN} [70], and SEED_{Image} [57]. To systematically assess VLMs’ performance on long-context tasks, we also compare our model with other state-of-the-art models on two additional long-context benchmarks: MileBench [94] for multi-image and video understanding, and Video-MME [37] for video analysis tasks.

Effectiveness of V2PE. We compare the performance of InternVL2-V2PE-32K and InternVL2-V2PE-256K on the Long-VQA validation split, MM-NIAH (including MM-NIAH_{IM}), and other standard multimodal benchmarks. The results are shown in Tab. 1, Tab. 2, and Tab. 3.

We demonstrate how varying positional increments impact the models’ performance on these benchmarks. Specifically, MM-NIAH_{IM} is evaluated on the image-retrieval task where both the “needle” and “question” are images, posing a stringent challenge for multimodal models to handle visual tokens in long-context sequences.

On standard short-context benchmarks (see Fig. 2), V2PE does not offer obvious advantages. We observe a performance drop on specific tasks, such as DocVQA and InfoQA, when the positional increment is reduced. In Appendix Sec. A, we investigate this issue by applying V2PE in the pre-training stage of VLMs. The results indicate that the marginal benefits of V2PE on short-context benchmarks

stem from the fact that InternVL2 was pre-trained and fine-tuned (SFT) using the default position encoding, making it inherently adapted to $\delta = 1$.

However, on long-context benchmarks like Long-VQA and MM-NIAH (see Fig. 1 and Fig. 3), V2PE demonstrates a clear advantage. For instance, on Long-VQA, the best performance is typically achieved when the positional increments are reduced to around 1/8 or 1/16. Moreover, as the token sequence length increases, smaller positional increments tend to be optimal.

In the MM-NIAH benchmark, involving longer token sequences, V2PE demonstrates optimal performance with positional increments of 1/256. Notably, InternVL2-V2PE-32K markedly enhances performance on sequences of 256K tokens, boosting scores from 1.5 to 56.9. This enhancement is also evident in InternVL2-V2PE-256K, which records the highest scores across various token sequence lengths when using a positional increment of 1/256. For sequences of 512K tokens, the score surges from 10.8 to 95.1. Additionally, with the same positional increment of 1/256, the InternVL2-V2PE-256K model attains a significant score of 64.5 on sequences reaching 1M tokens.

From our previous experimental results, we observe a qualitative relationship among the choice of δ , the context length, and model performance: with longer context, selecting a smaller δ improves model’s performance. Therefore, in later experiments, we set δ to 1/2 for evaluating general and 1/256 for long context benchmarks respectively, based

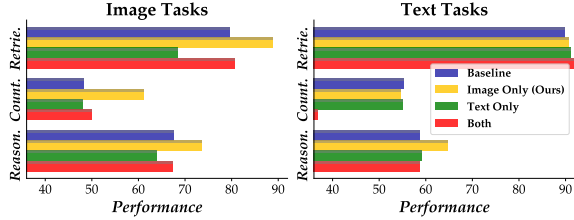


Figure 6. Performance on MM-NIAH benchmarks with different position encoding strategies across six task types: “image retrieval”, “text retrieval”, “image counting”, “text counting”, “image reasoning” and “text reasoning”. “Image only”, “Text only”, “Both” and “Baseline” represent applying V2PE to visual tokens, textual tokens, both simultaneously, and neither, respectively.

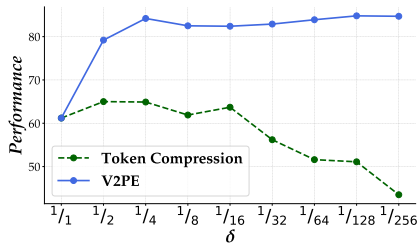


Figure 7. Performance on image retrieval task in MM-NIAH benchmark using V2PE and token compression strategies. Token compression reduces the number of visual tokens by a ratio δ , whereas V2PE employs a smaller positional increment δ while preserving all visual tokens.

on the context length. Furthermore, we provide theoretical analysis on optimal δ selection in the Appendix Sec. B

4.3. Comparison with Other VLMs.

We evaluate the performance of InternVL2-V2PE-32K against state-of-the-art vision-language models (VLMs) on both standard and long-context multimodal benchmarks, with results presented in Tab. 4 and Tab. 5, respectively. Despite built on a relatively small 2B-parameter model and incorporating a substantial amount of long-context training data, InternVL2-V2PE-32K retains highly competitive results on standard short-context multimodal benchmarks.

On long-context multimodal benchmarks, the model performs exceptionally well, demonstrating that V2PE can significantly enhance long-context abilities even built on a relative small model. These results validate the potential of V2PE for improving the performance of VLMs on a wide range of long-context multimodal tasks.

4.4. Ablation Study

Can V2PE be applied to textual tokens? We evaluate the impact of V2PE on visual tokens, textual tokens, both concurrently, and neither, with results on six MM-NIAH tasks presented in Fig. 6. Applying V2PE solely to textual tokens yields marginal improvements in language understanding but degrades performance on vision-centric tasks. When applied to both modalities, model performance exhibits instability. Notably, restricting V2PE to visual tokens consistently enhances performance across all tasks.

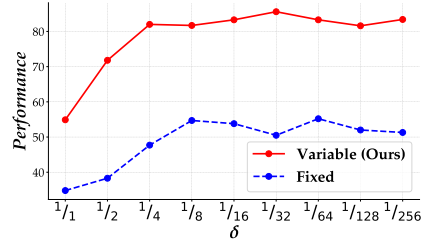


Figure 8. Performance on image retrieval task in MM-NIAH benchmark with variable versus fixed position increments for visual tokens in the VLMs. “Variable” denotes training with variable increments, allowing evaluation with arbitrary fixed δ , while “Fixed” means training and evaluating with the same fixed δ .

Is V2PE equivalent to visual token compression? We also compare V2PE with the token compression strategy, where visual tokens undergo pooling operations to reduce token number, allowing more images to be processed ideally. We test various compression ratios and evaluate the model’s performance on the MM-NIAH image retrieval task. The results, shown in Fig. 7, reveal that while token compression results in a rapid performance drop when the number of visual tokens is reduced 16 times or more, our method demonstrates stable performance across all settings.

Can positional increment be fixed during training? We train a set of InternVL2-FT-32K models with both V2PE method and fixed positional increments under the same settings. Their performance in the image retrieval task of the MM-NIAH benchmark is shown in Fig. 8. The results indicate that fixing the positional increment does not lead to optimal task performance. In contrast, the variable adjustment of positional increments provides consistent improvements.

Is V2PE superior to the position encoding extension methods? To compare V2PE with existing position encoding extension methods, we evaluated both InternVL2-V2PE-32K and InternVL2-FT-32K models using two widely adopted position extension techniques: linear interpolation and NTK-Aware Scaled RoPE. The results, presented in Fig. 5, Tab. 1, and Tab. 3, show that V2PE not only offers greater stability but also achieves superior task performance, especially in long-context scenarios.

5. Conclusion

We investigate the long-context capabilities of the existing VLMs, using our augmented long-context datasets, and find that positional encoding for visual tokens is critical for the long-context capabilities of VLMs. Based on this observation, we introduce V2PE, a novel position encoding strategy that applies smaller, variable positional increments to visual tokens, enabling more efficient handling of long-context multimodal sequences. By leveraging V2PE and our augmented long-context datasets, we fine-tune the open-source VLMs successfully, which shows significant improvements on both general and long-context multimodal benchmarks.

References

- [1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. 7
- [2] Muhammad Adnan, Akhil Arunkumar, Gaurav Jain, Prashant Nair, Ilya Soloveychik, and Purushotham Kamath. Keyformer: Kv cache reduction through key tokens selection for efficient generative inference. *Proceedings of Machine Learning and Systems*, 6:114–127, 2024. 3
- [3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736, 2022. 1
- [4] Dosovitskiy Alexey. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2
- [5] Anthropic. The claude 3 model family: Opus, sonnet, haiku. <https://www.anthropic.com>, 2024. 2, 7
- [6] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 2
- [7] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 2, 3
- [8] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020. 3
- [9] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024. 2
- [10] Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiusi Du, Zhe Fu, et al. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024. 2
- [11] @bloc97. Ntk-aware scaled rope, 2023. 3, 5
- [12] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 3
- [13] Aydar Bulatov, Yuri Kuratov, Yermek Kapushev, and Mikhail S Burtsev. Scaling transformer to 1m tokens and beyond with rmt. *arXiv preprint arXiv:2304.11062*, 2023. 3
- [14] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*, 2024. 1, 2
- [15] Lin Chen, Jinsong 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. 3
- [16] Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*, 2023. 3, 5
- [17] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 1
- [18] Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. Longlora: Efficient fine-tuning of long-context large language models. *arXiv preprint arXiv:2309.12307*, 2023. 3
- [19] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023. 1, 2, 3
- [20] 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. 1, 2, 3, 7
- [21] Alexis Chevalier, Alexander Wettig, Anirudh Ajith, and Danqi Chen. Adapting language models to compress contexts. *arXiv preprint arXiv:2305.14788*, 2023. 3
- [22] Hao-Tien Lewis Chiang, Zhuo Xu, Zipeng Fu, Mithun George Jacob, Tingnan Zhang, Tsang-Wei Edward Lee, Wenhao Yu, Connor Schenck, David Rendleman, Dhruv Shah, et al. Mobility v1a: Multimodal instruction navigation with long-context vlms and topological graphs. *arXiv preprint arXiv:2407.07775*, 2024. 1
- [23] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019. 3
- [24] Krzysztof Choromanski, Valerii Likhoshesterov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020. 3
- [25] Xiangxiang Chu, Limeng Qiao, Xinyang Lin, Shuang Xu, Yang Yang, Yiming Hu, Fei Wei, Xinyu Zhang, Bo Zhang, Xiaolin Wei, et al. Mobilevlm: A fast, reproducible and strong vision language assistant for mobile devices. *arXiv preprint arXiv:2312.16886*, 2023. 2
- [26] Xiangxiang Chu, Limeng Qiao, Xinyu Zhang, Shuang Xu, Fei Wei, Yang Yang, Xiaofei Sun, Yiming Hu, Xinyang Lin, Bo Zhang, et al. Mobilevlm v2: Faster and stronger baseline for vision language model. *arXiv preprint arXiv:2402.03766*, 2024. 2
- [27] Wenliang Dai, Junnan Li, D Li, AMH Tiong, J Zhao, W Wang, B Li, P Fung, and S Hoi. Instructblip: Towards

- general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2, 2023. 2
- [28] Zihang Dai. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019. 3
- [29] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shitong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanxia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yudian Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. 1
- [30] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *International Conference on Machine Learning*, pages 7480–7512. PMLR, 2023. 2
- [31] Jiayu Ding, Shuming Ma, Li Dong, Xingxing Zhang, Shao-han Huang, Wenhui Wang, Nanning Zheng, and Furu Wei. Longnet: Scaling transformers to 1,000,000,000 tokens. *arXiv preprint arXiv:2307.02486*, 2023. 3
- [32] Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. Longrope: Extending llm context window beyond 2 million tokens. *arXiv preprint arXiv:2402.13753*, 2024. 3
- [33] Khang T. Doan, Bao G. Huynh, Dung T. Hoang, Thuc D. Pham, Nhat H. Pham, Quan T. M. Nguyen, Bang Q. Vo, and Suong N. Hoang. Vintern-1b: An efficient multimodal large language model for vietnamese, 2024. 7
- [34] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*, 2024. 2
- [35] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Songyang Zhang, Haodong Duan, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Zhe Chen, Xinyue Zhang, Wei Li, Jingwen Li, Wenhui Wang, Kai Chen, Conghui He, Xingcheng Zhang, Jifeng Dai, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer2-4khd: A pioneering large vision-language model handling resolutions from 336 pixels to 4k hd. *arXiv preprint arXiv:2404.06512*, 2024. 2
- [36] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. *arXiv preprint arXiv:2103.10360*, 2021. 2
- [37] Chaoyou Fu, Yuhang Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024. 7
- [38] Mingze Gao, Jingyu Liu, Mingda Li, Jiangtao Xie, Qingbin Liu, Bo Zhao, Xi Chen, and Hui Xiong. Tc-llava: Rethinking the transfer from image to video understanding with temporal considerations. *arXiv preprint arXiv:2409.03206*, 2024. 3
- [39] Zhangwei Gao, Zhe Chen, Erfei Cui, Yiming Ren, Weiyun Wang, Jinguo Zhu, Hao Tian, Shenglong Ye, Junjun He, Xizhou Zhu, et al. Mini-internvl: A flexible-transfer pocket multimodal model with 5% parameters and 90% performance. *arXiv preprint arXiv:2410.16261*, 2024. 2
- [40] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 3
- [41] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021. 3

- [42] Shuhao Gu, Jialing Zhang, Siyuan Zhou, Kevin Yu, Zhaohu Xing, Liangdong Wang, Zhou Cao, Jintao Jia, Zhuoyi Zhang, Yixuan Wang, et al. Infinity-mm: Scaling multimodal performance with large-scale and high-quality instruction data. *arXiv preprint arXiv:2410.18558*, 2024. 3, 7
- [43] Chi Han, Qifan Wang, Wenhan Xiong, Yu Chen, Heng Ji, and Sinong Wang. Lm-infinite: Simple on-the-fly length generalization for large language models. *arXiv preprint arXiv:2308.16137*, 2023. 3
- [44] Adi Haviv, Ori Ram, Ofir Press, Peter Izsak, and Omer Levy. Transformer language models without positional encodings still learn positional information. *arXiv preprint arXiv:2203.16634*, 2022. 3
- [45] Wenyi Hong, Weihang Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, et al. Cogagent: A visual language model for gui agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14281–14290, 2024. 1, 3
- [46] Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, et al. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *arXiv preprint arXiv:2403.12895*, 2024. 2
- [47] Zhiyuan Hu, Yuliang Liu, Jinman Zhao, Suyuchen Wang, Yan Wang, Wei Shen, Qing Gu, Anh Tuan Luu, See-Kiong Ng, Zhiwei Jiang, et al. Longrecipe: Recipe for efficient long context generalization in large language models. *arXiv preprint arXiv:2409.00509*, 2024. 3
- [48] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36:72096–72109, 2023. 2
- [49] Xinting Huang and Nora Hollenstein. Long-range language modeling with selective cache. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4838–4858, 2023. 3
- [50] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 3
- [51] HyperGAI Research Team. Introducing hpt: A family of leading multimodal llms. <https://www.hypergai.com/blog/introducing-hpt-a-family-of-leading-multimodal-llms>, 2024. 2
- [52] Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhui Chen. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*, 2024. 1, 3
- [53] Guolin Ke, Di He, and Tie-Yan Liu. Rethinking positional encoding in language pre-training. *arXiv preprint arXiv:2006.15595*, 2020. 3
- [54] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European conference on computer vision*, pages 235–251. Springer, 2016. 7
- [55] Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*, 2020. 3
- [56] Shun Kiyono, Sosuke Kobayashi, Jun Suzuki, and Kentaro Inui. Shape: Shifted absolute position embedding for transformers. *arXiv preprint arXiv:2109.05644*, 2021. 3
- [57] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 7
- [58] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 2
- [59] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2
- [60] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 7
- [61] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*, 2024. 2
- [62] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26689–26699, 2024. 7
- [63] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 2
- [64] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.
- [65] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 1, 2
- [66] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with ringattention. *arXiv preprint arXiv:2402.08268*, 2024. 6
- [67] Jiaheng Liu, Zhiqi Bai, Yuanxing Zhang, Chenchen Zhang, Yu Zhang, Ge Zhang, Jiakai Wang, Haoran Que, Yukang Chen, Wenbo Su, et al. E²-llm: Efficient and extreme length extension of large language models. *arXiv preprint arXiv:2401.06951*, 2024. 3
- [68] Tengfei Liu, Yongli Hu, Mingjie Li, Junfei Yi, Xiaojun Chang, Junbin Gao, and Baocai Yin. Tackling real-world complexity: Hierarchical modeling and dynamic prompting for multimodal long document classification. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2025. 3

- [69] Xiaoran Liu, Ruixiao Li, Mianqiu Huang, Zhigeng Liu, Yuerong Song, Qipeng Guo, Siyang He, Qiqi Wang, Linlin Li, Qun Liu, et al. Thus spake long-context large language model. *arXiv preprint arXiv:2502.17129*, 2025. 3
- [70] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multimodal model an all-around player? In *European Conference on Computer Vision*, pages 216–233. Springer, 2025. 7
- [71] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024. 2, 7
- [72] 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. 2, 3, 7
- [73] Gen Luo, Yiyi Zhou, Yuxin Zhang, Xiawu Zheng, Xiaoshuai Sun, and Rongrong Ji. Feast your eyes: Mixture-of-resolution adaptation for multimodal large language models. *arXiv preprint arXiv:2403.03003*, 2024. 2
- [74] Ang Lv, Kaiyi Zhang, Shufang Xie, Quan Tu, Yuhan Chen, Ji-Rong Wen, and Rui Yan. Are we falling in a middle-intelligence trap? an analysis and mitigation of the reversal curse. *arXiv preprint arXiv:2311.07468*, 2023. 3
- [75] Tengchao Lv, Yupan Huang, Jingye Chen, Yuzhong Zhao, Yilin Jia, Lei Cui, Shuming Ma, Yaoyao Chang, Shaohan Huang, Wenhui Wang, et al. Kosmos-2.5: A multimodal literate model. *arXiv preprint arXiv:2309.11419*, 2023. 2
- [76] Fan Ma, Xiaojie Jin, Heng Wang, Yuchen Xian, Jiashi Feng, and Yi Yang. Vista-llama: Reliable video narrator via equal distance to visual tokens. *arXiv preprint arXiv:2312.08870*, 2023. 3
- [77] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022. 2, 7
- [78] 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. 1, 2, 3, 7
- [79] Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruvi Shah, Xianzhi Du, Futang Peng, Floris Weers, et al. Mm1: Methods, analysis & insights from multimodal llm pre-training. *arXiv preprint arXiv:2403.09611*, 2024. 2
- [80] Amirkeivan Mohtashami and Martin Jaggi. Landmark attention: Random-access infinite context length for transformers. *arXiv preprint arXiv:2305.16300*, 2023. 3
- [81] OpenAI. Hello gpt-4o, 2023. Accessed: 2023-11-14. 1, 7
- [82] OpenAI. Gpt-4v(ision) system card. https://cdn.openai.com/papers/GPTV_System_Card.pdf, 2023. 1, 2
- [83] R OpenAI. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2(5), 2023. 1, 2
- [84] OpenGVLab. Needle in a multimodal haystack, 2024. 4
- [85] Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*, 2023. 3, 5
- [86] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 2
- [87] Ofir Press, Noah A Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*, 2021. 3
- [88] Jiezhong Qiu, Hao Ma, Omer Levy, Scott Wen-tau Yih, Sinong Wang, and Jie Tang. Blockwise self-attention for long document understanding. *arXiv preprint arXiv:1911.02972*, 2019. 3
- [89] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2
- [90] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 3
- [91] Hongyu Ren, Hanjun Dai, Zihang Dai, Mengjiao Yang, Jure Leskovec, Dale Schuurmans, and Bo Dai. Combiner: Full attention transformer with sparse computation cost. *Advances in Neural Information Processing Systems*, 34: 22470–22482, 2021. 3
- [92] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018. 3, 4
- [93] Yan Shu, Zheng Liu, Peitian Zhang, Minghao Qin, Junjie Zhou, Zhengyang Liang, Tiejun Huang, and Bo Zhao. Video-xl: Extra-long vision language model for hour-scale video understanding, 2024. 1
- [94] Dingjie Song, Shunian Chen, Guiming Hardy Chen, Fei Yu, Xiang Wan, and Benyou Wang. Milebench: Benchmarking mllms in long context. *arXiv preprint arXiv:2404.18532*, 2024. 7
- [95] StepFun Research Team. Step-1v: A hundred billion parameter multimodal large model. <https://platform.stepfun.com>, 2024. 2
- [96] Jianlin Su. Rectified rotary position embeddings. <https://github.com/bojone/rerope>, 2023. 3, 5
- [97] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568: 127063, 2024. 3
- [98] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568: 127063, 2024. 3, 4

- [99] Tianxiang Sun, Xiaotian Zhang, Zhengfu He, Peng Li, Qinyuan Cheng, Hang Yan, Xiangyang Liu, Yunfan Shao, Qiong Tang, Xingjian Zhao, et al. Moss: Training conversational language models from synthetic data. *arXiv preprint arXiv:2307.15020*, 7:3, 2023. 1, 2
- [100] Yutao Sun, Li Dong, Barun Patra, Shuming Ma, Shao-han Huang, Alon Benhaim, Vishrav Chaudhary, Xia Song, and Furu Wei. A length-extrapolatable transformer. *arXiv preprint arXiv:2212.10554*, 2022. 3
- [101] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models*. <https://crfm.stanford.edu/2023/03/13/alpaca.html>, 3(6):7, 2023. 2
- [102] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 2
- [103] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 1, 2, 7
- [104] InternLM Team. Internlm: A multilingual language model with progressively enhanced capabilities, 2023. 2
- [105] Changyao Tian, Xizhou Zhu, Yuwen Xiong, Weiyun Wang, Zhe Chen, Wenhai Wang, Yuntao Chen, Lewei Lu, Tong Lu, Jie Zhou, et al. Mm-interleaved: Interleaved image-text generative modeling via multi-modal feature synchronizer. *arXiv preprint arXiv:2401.10208*, 2024. 1
- [106] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578, 2024. 2
- [107] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2
- [108] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shrutu Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 1, 2, 7
- [109] Szymon Tworowski, Konrad Staniszewski, Mikołaj Pacek, Yuhuai Wu, Henryk Michalewski, and Piotr Miłoś. Focused transformer: Contrastive training for context scaling. *Advances in Neural Information Processing Systems*, 36, 2024. 3
- [110] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 3
- [111] Benyou Wang, Lifeng Shang, Christina Lioma, Xin Jiang, Hao Yang, Qun Liu, and Jakob Grue Simonsen. On position embeddings in bert. In *International Conference on Learning Representations*, 2020. 3
- [112] Hengyi Wang, Haizhou Shi, Shiwei Tan, Weiyi Qin, Wenyuan Wang, Tunyu Zhang, Akshay Nambi, Tanuja Ganu, and Hao Wang. Multimodal needle in a haystack: Benchmarking long-context capability of multimodal large language models. *arXiv preprint arXiv:2406.11230*, 2024. 1
- [113] 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. 2
- [114] Mingjie Wang, Jun Zhou, Yong Dai, Eric Buys, and Minglun Gong. Enhancing zero-shot counting via language-guided exemplar learning. *arXiv preprint arXiv:2402.05394*, 2024. 1
- [115] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1, 2, 7
- [116] Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020. 3
- [117] Weiyun Wang, Min Shi, Qingyun Li, Wenhai Wang, Zhenhang Huang, Linjie Xing, Zhe Chen, Hao Li, Xizhou Zhu, Zhiguo Cao, et al. The all-seeing project: Towards panoptic visual recognition and understanding of the open world. *arXiv preprint arXiv:2308.01907*, 2023. 2
- [118] Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. Augmenting language models with long-term memory. *Advances in Neural Information Processing Systems*, 36, 2024. 3
- [119] Weiyun Wang, Shuibo Zhang, Yiming Ren, Yuchen Duan, Tiantong Li, Shuo Liu, Mengkang Hu, Zhe Chen, Kaipeng Zhang, Lewei Lu, et al. Needle in a multimodal haystack. *arXiv preprint arXiv:2406.07230*, 2024. 1, 2, 3, 7
- [120] Xidong Wang, Dingjie Song, Shunian Chen, Chen Zhang, and Benyou Wang. Longllava: Scaling multi-modal llms to 1000 images efficiently via hybrid architecture. *arXiv preprint arXiv:2409.02889*, 2024. 1, 3, 7
- [121] Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. *arXiv preprint arXiv:2403.10517*, 2024. 1
- [122] Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, Jinrong Yang, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. Vary: Scaling up the vision vocabulary for large vision-language model. In *European Conference on Computer Vision*, pages 408–424. Springer, 2025. 1, 3
- [123] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 2
- [124] Tianwen Wei, Liang Zhao, Lichang Zhang, Bo Zhu, Lijie Wang, Haihua Yang, Biye Li, Cheng Cheng, Weiwei Lü,

- Rui Hu, et al. Skywork: A more open bilingual foundation model. *arXiv preprint arXiv:2310.19341*, 2023. 2
- [125] Xilin Wei, Xiaoran Liu, Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Jian Tong, Haodong Duan, Qipeng Guo, Jiaqi Wang, et al. Videorope: What makes for good video rotary position embedding? *arXiv preprint arXiv:2502.05173*, 2025. 3
- [126] Yuetian Weng, Mingfei Han, Haoyu He, Xiaojun Chang, and Bohan Zhuang. Longvlm: Efficient long video understanding via large language models. In *European Conference on Computer Vision*, pages 453–470. Springer, 2025. 1
- [127] Jeffrey Willette, Heejun Lee, Youngwan Lee, Myeongjae Jeon, and Sung Ju Hwang. Training-free exponential extension of sliding window context with cascading kv cache. *arXiv preprint arXiv:2406.17808*, 2024. 3
- [128] Yuhuai Wu, Markus N Rabe, DeLesley Hutchins, and Christian Szegedy. Memorizing transformers. *arXiv preprint arXiv:2203.08913*, 2022. 3
- [129] X.ai. Grok-1.5 vision preview. <https://x.ai/blog/grok-1.5v>, 2024. 2
- [130] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023. 3
- [131] Peng Xu, Wei Ping, Xianchao Wu, Lawrence McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. Retrieval meets long context large language models. *arXiv preprint arXiv:2310.03025*, 2023. 3
- [132] Fuzhao Xue, Yukang Chen, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, et al. Longvila: Scaling long-context visual language models for long videos. *arXiv preprint arXiv:2408.10188*, 2024. 1, 3, 7
- [133] L Xue. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*, 2020. 3
- [134] Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*, 2023. 2
- [135] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 2, 7
- [136] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. CLEVRER: collision events for video representation and reasoning. In *ICLR*, 2020. 1
- [137] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. 7
- [138] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297, 2020. 3
- [139] 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
- [140] Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Wenwei Zhang, Hang Yan, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*, 2023. 2
- [141] Peitian Zhang, Shitao Xiao, Zheng Liu, Zhicheng Dou, and Jian-Yun Nie. Retrieve anything to augment large language models. *arXiv preprint arXiv:2310.07554*, 2023. 3
- [142] Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, Songyang Zhang, Wenwei Zhang, Yining Li, Yang Gao, Peng Sun, Xinyue Zhang, Wei Li, Jingwen Li, Wenhai Wang, Hang Yan, Conghui He, Xingcheng Zhang, Kai Chen, Jifeng Dai, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*, 2024. 2
- [143] Peitian Zhang, Zheng Liu, Shitao Xiao, Ninglu Shao, Qiwei Ye, and Zhicheng Dou. Long context compression with activation beacon. *arXiv preprint arXiv:2401.03462*, 2024. 3
- [144] 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. 1, 3
- [145] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. 3
- [146] Yikai Zhang, Junlong Li, and Pengfei Liu. Extending llms’ context window with 100 samples. *arXiv preprint arXiv:2401.07004*, 2024. 3
- [147] Liang Zhao, Tianwen Wei, Liang Zeng, Cheng Cheng, Liu Yang, Peng Cheng, Lijie Wang, Chenxia Li, Xuejie Wu, Bo Zhu, et al. Longskywork: A training recipe for efficiently extending context length in large language models. *arXiv preprint arXiv:2406.00605*, 2024. 3
- [148] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023. 2
- [149] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigt-4: Enhancing vision-language

understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1

- [150] Dawei Zhu, Nan Yang, Liang Wang, Yifan Song, Wenhao Wu, Furu Wei, and Sujian Li. Pose: Efficient context window extension of llms via positional skip-wise training. *arXiv preprint arXiv:2309.10400*, 2023. 3
- [151] Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *Advances in Neural Information Processing Systems*, 36, 2024. 1