NaVLM-PVC: Progressive Visual Compression for Efficient Native-Resolution Encoding in MLLMs

Shichu Sun[1,4*], Yichen Zhang[2*], Haolin Song[3,5*], Zonghao Guo[2†], Chi Chen[2], Yidan Zhang[3,6], Yuan Yao[2], Zhiyuan Liu[2], and Maosong Sun[2†]

[1] School of Advanced Interdisciplinary Sciences, University of Chinese Academy of Sciences, China
[2] Tsinghua University, China
[3] University of Chinese Academy of Sciences, China
[4] Institute of Software, Chinese Academy of Sciences, China
[5] Institute of Automation, Chinese Academy of Sciences, China
[6] Aerospace Information Research Institute, Chinese Academy of Sciences, China

Abstract. Visual encoding followed by token condensing has become the standard architectural paradigm in multi-modal large language models (MLLMs). Many recent MLLMs increasingly favor global native-resolution visual encoding over slice-based methods. To investigate this trend, we systematically compare their behavior on vision-language understanding and attention patterns, revealing that global encoding enhances overall capability but at the expense of greater computational overhead. To address this issue, we present NaVLM-PVC, an MLLM centered upon our proposed Progressive Visual Compression (PVC) method, which can be seamlessly integrated into standard Vision Transformer (ViT) to enable efficient native-resolution encoding. The PVC approach consists of two key modules: (i) refined patch embedding, which supports flexible patch-size scaling for fine-grained visual modeling, (ii) windowed token compression, hierarchically deployed across ViT layers to progressively aggregate local token representations. Jointly modulated by these two modules, a widely pretrained ViT can be reconfigured into an efficient architecture while largely preserving generality. Evaluated across extensive benchmarks, the transformed ViT, termed ViT-PVC, demonstrates competitive performance with MoonViT while reducing TTFT (time-to-first-token) by 2.4×, when developed within an identical MLLM architecture. Building upon ViT-PVC, NaVLM-PVC also achieves competitive performance to Qwen2-VL, while further reducing TTFT by 1.9×. We will release all code and checkpoints to support future research on efficient MLLMs.

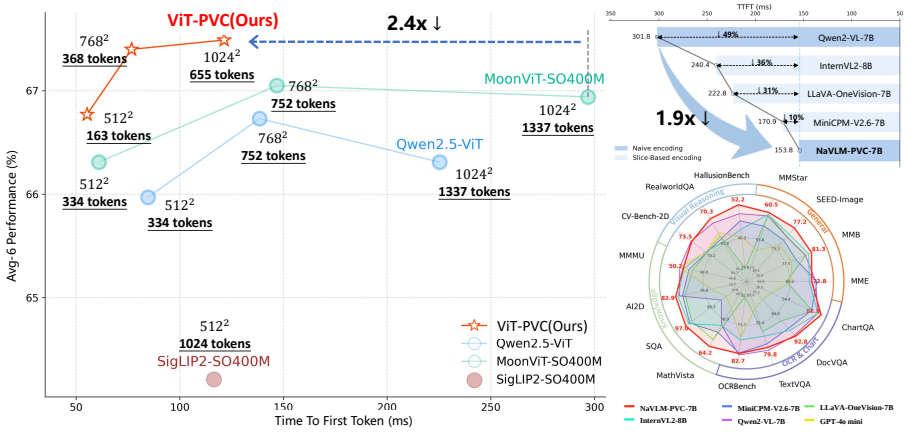
Keywords: Multimodal Large Language Model

[*] Equal contribution [†] Corresponding author

This work is supported by the AI9Stars community

1 Introduction

Recent advances in Multimodal Large Language Models (MLLMs) [23, 50, 59, 60] have significantly expanded the capabilities of vision-language understanding across diverse scenario, including optical character recognition [19, 42], remote sensing [22, 58, 66], and mobile agents [48, 61]. To efficiently support such broad applicability, visual encoding followed by token condensing has become the general vision embedding paradigm in MLLMs, enabling more standardized and scalable multi-modal training.



(a) Latency-performance trade-off of ViT-PVC

(b) Trade-off of NaVLM-PVC

Fig. 1: ViT-PVC and NaVLM-PVC exhibit a superior trade-off between efficiency and performance. (a) Within the LLaVA training paradigm, ViT-PVC achieves higher average performance across 6 benchmarks, such as MMBench and AI2D, compared to state-of-the-art vision encoders, while maintaining substantially greater computational efficiency (*e.g.*, achieving a 2.4× reduction in latency relative to MoonViT). (b) NaVLM-PVC attains performance comparable to advanced MLLMs (*e.g.*, Qwen2-VL) across 15 diverse benchmarks, while delivering 1.9× efficiency gains.

This encode-then-compress framework, however, entails considerable computational overhead, as the vision encoder must process an excessive number of tokens before any reduction, which is exacerbated when image resolution increases [12, 67, 68]. Early MLLMs alleviated this problem through slice-based encoding [17, 25, 28, 71], which encodes smaller image crops independently to reduce computation. However, this line of methods inevitably leads to a fragmented semantic context and limited global awareness. Previous studies [20, 66] manually design cross-slice interaction modules to supplement global information, yet these adapters are generally lack large-scale pretraining and so far have found limited adoption in industrial MLLMs.

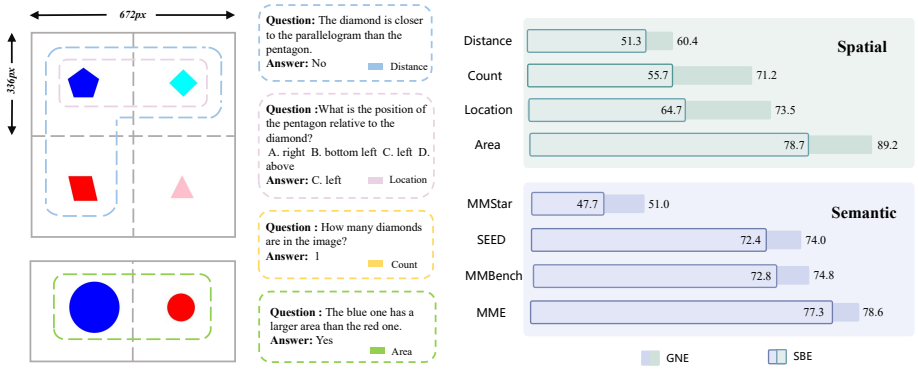
By contrast, there is a growing trend toward adopting global native-resolution encoding in recent state-of-the-art MLLMs [18, 37, 39, 59], which processes the entire image in a single forward pass. Although intuitively well-motivated, the principles behind this approach remain insufficiently explored.

In this work, we conduct controlled pilot experiments to investigate the mechanisms of global native-resolution versus slice-based encoding in terms of spatial and semantic understanding. Moreover, we further analyze the internal patterns of attention activation. Our experiments indicate that global native-resolution encoding yields stronger cross-modal understanding, yet its associated computational overhead remains substantial.

To address this issue, we introduce NaVLM-PVC, an MLLM built upon a novel Progressive Visual Compression (PVC) framework, which can be seamlessly integrated into standard ViT to enable efficient native-resolution encoding. The PVC framework consists of two key components. **(i) Refined patch embedding (RPE).** As the tokenizer of images, the ViT patch-embedding layer fundamentally determines the granularity of visual tokenization. By scaling the patch size to finer levels via an equivalent weight-transformation scheme [4], this module supports more detailed visual modeling while preserving compatibility with pretrained ViTs. **(ii) Windowed token compression (WTC).** We insert lightweight, learnable local compressors within the encoder to progressively merge tokens inside local windows (*e.g.*, 2×2) across ViT layers, reducing sequence length during encoding. Initialized with average pooling and trained to produce content-adaptive weights, the module preserves local semantics while lowering ViT computation and LLM prefill cost. Jointly modulated by these two modules, a widely pretrained ViT can be reconfigured into a more efficient architecture, *e.g.*, via adjusting patch size, compressor count, and insertion indices.

Under the PVC framework, a widely pretrained ViT could be reconfigured into ViT-PVC. In the LLaVA setting [17, 25] with a Qwen2-7B [51] as the LLM, ViT-PVC achieves competitive performance across 6 benchmarks under varying input resolutions, while reducing time-to-first-token (TTFT) by up to 2.4× relative to MoonViT [50], as shown in Fig. 1(a). (Please refer to Appendix A.1 for details.) Building on ViT-PVC, NaVLM-PVC attains comparable performance on 15 vision–language benchmarks compared to Qwen2-VL, yet delivers 1.9× lower TTFT latency, shown in Fig. 1(b).

Our contributions are threefold: (1) We conduct controlled probes of global native-resolution versus slice-based encoding, evidencing superior cross-modal understanding for the former and clarifying its underlying mechanisms. (2) We introduce a PVC framework, which jointly modulates refined patch embedding and windowed token compression to reconfigure widely pretrained ViTs into efficient native-resolution ones. (3) Evaluations on extensive vision–language benchmarks demonstrates the effectiveness and efficiency of ViT-PVC and NaVLM-PVC.



(a) Example from the ShapeGrid bench

(b) Performance comparison of two methods

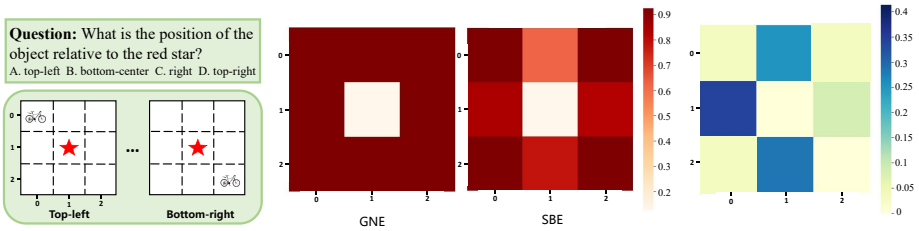
Fig. 2: ShapeGrid and model performance. (a) Examples from ShapeGrid bench, with each subset matched with color boxes. (b) Performance comparison between global native-resolution encoding (GNE) and slice-based encoding (SBE) across different general benchmarks and ShapeGrid subsets.

2 Pilot Experiment

First, we start with a series of controlled pilot experiments that systematically contrast global native-resolution visual encoding (GNE) and slice-based encoding (SBE) with respect to their capacities in semantic understanding and spatial perception. To guarantee experimental fairness, we adopt an identical model architecture, utilize an equivalent training corpus, and perform evaluation on a consistent suite of general-purpose benchmarks. Moreover, to decouple spatial perception from semantic understanding, we introduce a synthetic probe benchmark for spatial analysis.

2.1 Experimental Protocol

Probe benchmarking. For evaluating general semantic understanding, we adopt a set of widely used multi-modal benchmarks, including MMBench, *etc.*, shown in Fig. 2(b). These benchmarks comprehensively cover diverse capabilities such as object attribution recognition, position relation recognition, optical character recognition (OCR), and visual reasoning. To assess spatial perception, it is important to note that the slicing operation intrinsic to slice-based visual encoding tends to fragment objects in natural images, which in turn introduces semantic discontinuities and makes it difficult to disentangle spatial perception from semantic understanding. To solve this problem, we construct a controlled synthetic dataset, ShapeGrid. The ShapeGrid bench is generated from a template pool of parameterized geometric shapes, incorporating controlled variations in color, scale, and position. Final images are composed using predefined layouts aligned with slice-based encoding strategies. The dataset supports four spatial perception tasks as shown in Fig. 2(a). See details in Appendix A.2.2.



(a) Example from the Sudoku subset (b) Heatmap of accuracy (c) Attention bias of GNE minus SBE

Fig. 3: Illustration and analysis on ShapeGrid-Sudoku subset. (a) Example from the Sudoku subset. (b) Accuracy heatmap of models with global native-resolution encoding (GNE) vs. slice-based encoding (SBE) on the Sudoku subset. (c) Attention score bias map showing the difference in attention activation between GNE and SBE.

Model configuration and training setup. We adopt a standard MLLM architecture with SigLIP2-SO400M [54] as the vision encoder, a pixel-unshuffle [11] projector, and Qwen2-7B [51] as the LLM. Training follows a two-stage paradigm [71], first optimizing the vision encoder and projector while freezing the LLM, then fine-tuning all parameters. Both GNE and SBE methods are configured with a maximum resolution of 1008×1008 pixels. Details are provided in Appendix A.2.1.

2.2 Experimental Analysis

Overall performance. In Fig. 2(b), GNE method achieves significantly higher performance than SBE method, with clear improvements on both the general semantic benchmarks and the ShapeGrid. On average, it improves general semantic understanding by 2.1% (69.6% vs. 67.5%) and spatial perception by 11.0% (73.6% vs. 62.6%), indicating that preserving holistic visual context consistently benefits both semantic understanding and spatial perception.

Behavior on spatial directional perception. To better evaluate spatial positional understanding, we extend the original location subset into a Sudoku-style variant, where directional localization is explicitly isolated for clear evaluation. As for data construction, each image follows a 3×3 layout with a fixed central anchor and surrounding objects sampled from the template pool and additional real-world categories, which contains 8,000 image-query pairs focused on relative direction prediction, as shown in Fig. 3(a). Further details are provided in the Appendix A.2.2.

As shown in Fig. 3(b), the GNE method maintains consistently high accuracy on the Sudoku subset across all evaluated directions, demonstrating a well-balanced and robust capacity for spatial reasoning. In contrast, SBE method suffers systematic degradation along the vertical and horizontal axes, forming a clear cross-shaped bias. These results reveal an inherent systematic flaw in SBE approach. The trend that GNE alleviates such flaws more effectively can also

be observed when comparing Qwen2.5-VL (GNE) and MiniCPM-o 2.6 (SBE), which is shown in Appendix A.2.3.

Attention pattern analysis. To further investigate this phenomenon, we analyze the attention activation patterns of GNE and SBE methods. Specifically, we compute the average attention from answer text tokens to image tokens within the ground-truth cell and visualize the differences in Fig. 3(c). The results reveal a clear anisotropy in SBE: attention to the top, bottom, left, and right positions is markedly weaker, whereas the four corners exhibit similar activation levels. This suggests that image partitioning in SBE disrupts spatial uniformity, which in turn introduces a systematic directional bias. In contrast, GNE maintains a more evenly distributed and coherent attention pattern, thereby preserving holistic spatial relationships and enabling more accurate localization.

2.3 Conclusions on Pilot Experiment

In summary, we demonstrate that global native-resolution encoding consistently outperforms slice-based methods across both general semantic benchmarks and spatial perception probes. The superiority is particularly pronounced in directional localization, where global encoding eliminates the cross-shaped bias inherent in slice-based methods and maintains balanced attention distributions. These findings indicate that global encoding not only enhances overall capability in vision-language understanding but also provides mechanistic insights into its advantage in spatial reasoning. Nevertheless, global encoding still incurs substantial computational cost, as discussed in previous work [27,36,68], underscoring the urgent need for a native and efficient visual encoding paradigm.

3 NaVLM-PVC

Building upon conclusions of pilot experiments, we propose NaVLM-PVC, an MLLM equipped with our Progressive Visual Compression (PVC) approach for efficient native-resolution encoding. NaVLM-PVC follows the standard MLLM architecture of a vision encoder, a projector, and a large language model (LLM), as illustrated in Fig. 4. PVC framework is applied within the vision encoder, where it integrates two modules: Refined patch embedding (RPE) and Windowed token compression (WTC). By progressively condensing tokens while preserving global contextual information, these components reconfigure a pretrained ViT into an efficient and generalizable model, referred to as ViT-PVC.

3.1 ViT-PVC

ViTs are built upon the transformer architecture, where self-attention serves as the core operation [56]. Although this design offers strong representational capacity, a key limitation lies in its quadratic complexity with respect to token length. Higher input resolutions substantially increase visual token counts, resulting in elevated memory demands and inference latency during both visual

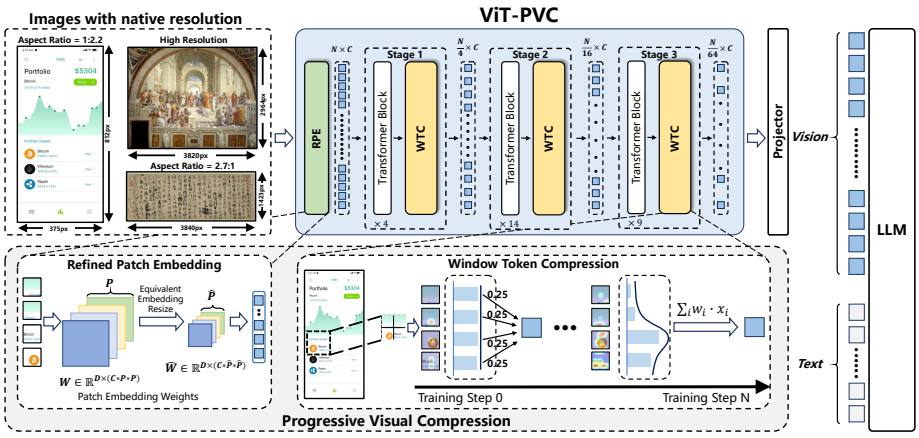


Fig. 4: Overview of the architecture of NaVLM-PVC. ViT-PVC first utilizes Refined Patch Embedding (RPE) to tokenize images with native-resolution into fine-grained tokens. Window Token Compression (WTC) modules are inserted at multiple stages to progressively reduce token length while learning local semantics. The final vision tokens are then projected into the LLM.

encoding and LLM pre-filling. A more fundamental alternative is to replace self-attention with linear variants [8] or RNN-style architectures [46, 65]. However, such designs typically require large-scale training from scratch and fail to inherit the well-established modeling capacity of pretrained transformer-based models. Therefore, our core idea is to keep the token length within a controllable range while enriching fine-grained visual features. To achieve this, we apply RPE to flexibly expand visual granularity, and then employ WTC to progressively condense tokens and reduce redundancy.

3.1.1 Refined patch embedding

Generally, the objective of this module is to reduce the patch size P employed in patch embedding (*e.g.*, from 14×14 to 10×10), which increases the modeling granularity of visual tokens, while preserving the representational equivalence of the original embeddings. Specifically, for an original patch embedding kernel weights $W \in \mathbb{R}^{D\times(C*P*P)}$, where C denotes the number of input channels (*e.g.*, 3 for RGB images), and D the embedding dimension, we define the transformed weights $\hat{W} \in \mathbb{R}^{D\times(C*\hat{P}*\hat{P})}$, where $\hat{P} < P$. Given a patch vector of the image $t \in \mathbb{R}^{1\times(C*P*P)}$ and its finer-level counterpart $\hat{t} \in \mathbb{R}^{1\times(C*\hat{P}*\hat{P})}$, we can derive a linear transformation matrix $B \in \mathbb{R}^{(C*P*P)\times(C*\hat{P}*\hat{P})}$ that maps the coarse patch to the finer one, *i.e.*, $\hat{t} = tB$. To ensure representational equivalence between the original patch embedding $tW^\top \in \mathbb{R}^{1\times D}$ and refined patch embedding $\hat{t}\hat{W}^\top \in \mathbb{R}^{1\times D}$, we require $tW^\top \approx \hat{t}\hat{W}^\top$. Since $\hat{t} = tB$, this condition can be rewritten as $B\hat{W}^\top \approx W^\top$. Following previous work [4], we utilize the least-squares solution to this relation and yields $\hat{W} = (B^\top)^+ W$, where $(B^\top)^+$

denotes the pseudo-inverse of $B^\top$. Through this weight transformation, we update a patch embedding weights W into new one $\hat{W}$, which could encode each finer-level patch vector into an image token.

So that, given an input image I of size $H_I \times W_I$, patch embedding with transformed weights $\hat{W} \in \mathbb{R}^{D\times(C*\hat{P}*\hat{P})}$ produces a token feature map of size $h_I \times w_I = (H_I/\hat{P}) \times (W_I/\hat{P})$, which can be flattened into N visual tokens $\{x_i = t_i W^\top\}_{i=1}^{N}$, where $N = h_I \times w_I$, $t_i \in \mathbb{R}^{1\times(C*\hat{P}*\hat{P})}$ is i-th image patch vector.

3.1.2 Windowed token compression

Following patch embedding, a large number of visual tokens are forwarded into transformer layers for further encoding. Especially, when the patch size decreases using RPE, the number of visual tokens increases quadratically, leading to heavy computational overhead. To mitigate the burden on the ViT backbone as well as the pre-filling stage of the LLM, we insert a set of lightweight compression layers within the ViT. These layers hierarchically reduce the token sequence length, thereby partitioning the backbone into J stages with progressive compression, as shown in Fig. 4.

Plain pooling strategy. In the j-th compression stage, we predefine a local window of size 2×2 and partition the entire feature map into non-overlapping windows. Within each window, an average pooling strategy is applied to aggregate the tokens as

$$x_{avg} = \frac{1}{2\times 2}\sum_{i=1}^{2\times 2} x_i, \tag{1}$$

so that after the j-th compression stage, the token feature map resolution is downsampled by $2\times$ and the number of visual tokens is condensed to $\{x_i\}_{i=1}^{N/4^j}$. Although this simple pooling operation assigns uniform weights to all tokens and ignores their semantic importance, our experiments show that it facilitates more stable and effective convergence. In contrast, parameterized methods such as pixel-unshuffle [60] intuitively provide enhanced local semantic modeling capacity but exhibit convergence challenges when inserted in early ViT layer (see Appendix A.5.1).

Enhanced pooling mechanism. To enhance the expressiveness of pooling operation while maintaining efficient convergence, we introduce a content-adaptive pooling mechanism. Specifically, it first performs feature average pooling to obtain an aggregated token $x_{avg} \in \mathbb{R}^{1\times D}$ as mentioned in Eq. (1). This token is then concatenated to each of the tokens $\{x_i\}_{i=1}^{4}$ within the local window to form extended representations $\hat{x}_i = [x_i; x_{avg}] \in \mathbb{R}^{1\times 2D}$, followed by passing through an MLP ($f_\theta : \mathbb{R}^{1\times 2D} \to \mathbb{R}^{1\times D}$) to compute channel-wise attention logits $a_i = f_\theta(\hat{x}_i)$. In the end, the Eq. (1) can be improved as a learnable weighted aggregation like

$$x_{avg} = \sum_{i=1}^{2\times 2} w_i x_i = \sum_{i=1}^{4} \frac{\exp(a_i)}{\sum_{i=1}^{4}\exp(a_i)} x_i. \tag{2}$$

During early training, attention weights w_i tend to be uniform by zero initializing the MLP f_θ, which effectively approximates the average pooling for stabilize optimization. As training progresses, f_θ gradually learns semantically meaningful aggregation weights, allowing the model to preserve key spatial and semantic structures while reducing the token count.

3.1.3 Joint modulation Building on the two proposed modules, the PVC framework reconfigures a pretrained ViT by jointly modulating key structural factors. The patch size P in refined patch embedding controls token granularity, while the number of J WTC and their positions j in ViT layers govern hierarchical token reduction. Following established design principles [30, 36], these factors systematically transform a native-resolution ViT into a more efficient architecture.

3.2 Vision Projector and LLM

MLP projector. Considering that ViT-PVC already integrates hierarchical token compression within the vision encoder, the role of the projector is simplified to the pure feature alignment. We concretely employ a simple MLP like [23, 25] to directly map the visual tokens into the language embedding space. Such a design avoids redundant manual engineering and offers a more standardized and scalable solution, facilitating large-scale MLLM training.

Large language model. Compared with slice-based methods such as [17, 60, 68], our design encodes images at native-resolution and thus obviates the need for inserting special tokens (*e.g.*, "," and "/n") to explicitly inform the LLM of slice layouts. Instead, we simply adopt the `<image>` and `</image>` placeholders to delimit the visual context, allowing all tokens to be seamlessly injected into the language model as contextual inputs.

4 Experiments

In this section, we introduce the training recipe, detail the experiment setting , analyze the model performance and further conduct the ablation study.

4.1 Training Recipe

Overall training strategy. We adopt a three-stage training paradigm to gradually enhance the capabilities of NaVLM-PVC, more details are shown in the Appendix A.4. Stage 1 focuses on vision-language alignment, where a plain ViT initialized from MoonViT-SO-400M is transformed into ViT-PVC and aligned with a Qwen2-7B LLM via an MLP projector. Stage 2 conducts joint multimodal pre-training, aligning vision and language features in a unified space. Stage 3 applies supervised fine-tuning with instruction and reasoning data [16] for coherent response generation. The overall training period is ∼300h with 32× 80G-A100 GPUs. Details are provided in Appendix A.3.

Table 1: Performance comparison of NaVLM-PVC and state-of-the-art MLLMs on general and knowledge benchmarks. The performance is reported from its technical paper or OpenCompass leaderboard. "*" indicates the performance we reproduce using VLMEval-Kit. "Ratio" denotes the compression ratio defined in Sec. 4.2.

Model	LLM	#Data	Ratio	General			Knowledge				
				MME	MMB	SEED-Image	MMStar	MMMU	AI2D	SQA	MathVista
Academic Open-source MLLMs											
FastVLM	Vicuna-7B	1.6M	**64**	-	-	-	-	37.3	-	74.1	-
VILA^2	Llama3-8B	57.5M	1	-	76.6	66.1	-	40.8	-	87.6	-
LLaVA-OneVision	Qwen2-7B	9.4M	4	1998.0	80.8	75.4	**61.7**	48.8	81.4	96.0	63.2
Cambrain-1	Llama-3-8B	9.5M	16	-	75.9	74.7	-	42.7	73.0	80.4	49.0
Valley2	Qwen2.5-7B	27.2M	4	-	80.7	-	61.0	**57.0**	82.5	-	**64.6**
Closed-source MLLMs											
GPT-4o-mini	-	-	-	2041.0*	76.8*	72.5*	50.0*	54.1*	73.7*	82.6*	52.5*
Gemini-1.5-Pro	-	-	-	-	73.9	-	59.1	60.6	79.1	-	58.3
Step-1.5V-mini	-	-	-	-	79.7	-	54.7	51.7	81.3	-	57.8
Commercial Open-source MLLMs											
InternVL2	InternLM2.5-7B	∼100M	4	2210.3	79.5	76.2*	60.7	54.1	**83.0**	84.6	58.3
POINTS	Qwen-2.5-7B	431.0M	4	2195.2	**83.2**	74.8	61.0	49.4	80.9	94.8	63.1
MiniCPM-V-2.6	Qwen2-7B	460.0M	16	**2348.4**	78.0	74.1	57.5	49.8	82.1	96.9*	60.6
Qwen2-VL	Qwen2-7B	∼700.0M	4	2326.8	80.7	75.3*	60.7	54.1	**83.0**	84.6*	58.2
DeepSeek-VL2	DeepSeekMoE-16B	∼203.0M	4	2123.0	79.9	76.8*	57.5	49.7	81.7	96.2*	61.9
NaVLM-PVC (ours)	Qwen2-7B	20.1M	**64**	2183.6	81.3	**77.2**	60.5	50.2	82.9	**97.0**	64.2

Table 2: Performance comparison of NaVLM-PVC and state-of-the-art MLLMs on OCR&Chart and visual reasoning benchmarks.

Model	LLM	#Data	Ratio	Visual Reasoning			OCR & Chart			
				HallusionBench	RealworldQA	CV-Bench-2D	OCRBench	TextVQA	DocVQA	ChartQA
Academic Open-source MLLMs										
FastVLM	Vicuna-7B	1.6M	**64**	-	-	-	-	67.4	62.8	-
VILA^2	Llama3-8B	57.5M	1	-	-	-	-	73.4	-	-
LLaVA-OneVision	Qwen2-7B	9.4M	4	31.6	66.3	69.7*	62.2	75.9*	87.5	80.0
Cambrain-1	Llama-3-8B	9.5M	16	-	64.2	-	62.4	71.7	77.8	73.3
Valley2	Qwen2.5-7B	27.2M	4	48.0	-	-	84.2	-	-	-
Closed-source MLLMs										
GPT-4o-mini	-	-	-	42.5*	67.2*	69.9*	78.5*	68.0*	78.2*	29.04*
Gemini-1.5-Pro	-	-	-	45.6	-	-	75.4	-	-	-
Step-1.5V-mini	-	-	-	46.7	-	-	77.3	-	-	-
Commercial Open-source MLLMs										
InternVL2	InternLM2.5-7B	∼100M	4	45.2	64.4	70.1*	79.4	77.4	91.6	83.3
POINTS	Qwen-2.5-7B	431.0M	4	46.0	67.3	-	72.0	-	-	-
MiniCPM-V-2.6	Llama-3-8B	460.0M	16	48.1	65.5*	69.7*	85.2	80.1	90.8	79.4*
Qwen2-VL	Qwen2-7B	∼700.0M	4	50.6	70.1	**76.0***	**86.6**	**84.3**	**94.5**	83.0
DeepSeek-VL2	DeepSeekMoE-16B	∼203.0M	4	43.8	65.4	-	83.2	83.4	92.3	**84.5**
NaVLM-PVC (ours)	Qwen2-7B	20.1M	**64**	**52.2**	**70.3**	73.5	82.7	79.8	92.8	82.8

Ablation training strategy. The MLLM uses SigLIP2-SO-400M as the vision encoder, an MLP projector, and Qwen2-7B as the LLM, with training and inference fixed at a resolution of 1024×1024. We follow the LLaVA pipeline, using LLaVA-Pretrain-558K for vision-language alignment and SFT-858K [71] for supervised fine-tuning. For the baseline, we insert a WTC layer using pixel-unshuffle in the last ViT layer. For the PVC incremental setting, an extra ViT pre-alignment stage is added before standard LLaVA training to mitigate the perturbation of ViT feature modeling introduced by PVC. The pre-alignment data is a subset of stage 1. More details are supplied in Appendix A.4.

4.2 Experiment Setting

Benchmarks. We evaluate our model on a comprehensive suite of multi-modal benchmarks, which are divided into four categories: (1) General benchmarks including MME [14], MMB [33], SEED [24] and MMStar [6]. (2) Knowledge benchmarks such as SQA [41], MMMU [69], AI2D [21] and MathVista [40]. (3) OCR&Chart benchmarks consist of OCRBench [35], TextVQA [47], DocVQA [44] and ChartQA [43]. (4) Visual reasoning benchmarks like HallusionBench [15], RealWorldQA [64] and CV-Bench [53].
Evaluation Metrics. (1) We present the amount of training image-text pairs (#Data) for fair comparison. For methods that reports the number of training tokens (#token), we roughly compute the number of image-text pairs by dividing #token by a standard maximum token limitation (*i.e.*, 4096) in a single batch. (2) We compute the compression ratio as the final LLM token count divided by the number of tokens after the ViT patch embedding, where a higher value indicates denser visual information and more efficient inference. (3) We report Time-to-First-Token (TTFT) under a fixed resolution of 1024×1024, averaged across 100 inferences, using FlashAttention2 [9] and bf16 data type on one single 80G-A100 GPU for the whole model. (4) In ablation study, we report the average accuracy of benchmarks within each sub-category.
Counterparts. The compared models are categorized as three manifold. (1) Commercial open-source MLLMs, trained large corpora (*e.g.* over 100M vision-language samples) by industry labs, which includes MiniCPM-V2.6 [68], Qwen2-VL [59], DeepSeek-VL2 [62], InternVL2 [7], POINTS [34], VILA^2 [13] and *etc.* (2) Academic-scale MLLMs like LLaVA-OneVision [23], Cambrain-1 [53], FastVLM [55], Valley2 [63]) and *etc.* (3) Closed-source MLLMs such as GPT-4o-mini [45], Gemini-1.5-Pro [49] and Step-1.5V-mini [52].

4.3 Model Performance

In Tab. 1, compared with commercial open-source MLLMs, our NaVLM-PVC achieves strong performance despite using one order of magnitude less training data than Qwen2-VL (20.1M vs. 700.0M). On general benchmarks, it demonstrates notable gains—such as improvements on MMB (81.3 vs. 80.7) and SEED-Image (77.2 vs 75.3), showing the effectiveness of our approach in general multimodal understanding. On knowledge benchmarks, especially MathVista, our model surpasses both InternVL2 and MiniCPM-V-2.6, highlighting the superior capability of our global encoding mechanism in capturing complex geometric and spatial relationships.

Visual reasoning and OCR&Chart tasks require more fine-grained visual representations. As illustrated in Tab. 2, although our NaVLM-PVC has the highest compression ratio (64), it significantly outperforms the slice-based encoding methods MiniCPM-V-2.6 (ratio = 16) and Intern-VL-2 (ratio = 4) on spatial reasoning datasets such as HallusionBench and CV-Bench. Moreover, on the OCR&Chart subset, our MLLM delivers comparable performance compared to Qwen2-VL and MiniCPM-V-2.6, demonstrating that our high-compression

Table 3: Evaluation on proposed modules including refined patch embedding (RPE) and windowed token compression (WTC). "# WTC" denotes the number of inserted WTC in ViT, "CA pooling" the brief of content-adaptive pooling described in Sec. 3.1. For all configurations, the WTC inserted at the 27th layer is implemented using pixel-unshuffle.

Method	Patch Size	# WTC	Layer Id	Tokens↓	TTFT (ms)↓	Avg↑	General↑	OCR&Chat↑	Visual Reasoning↑	Knowledge↑
Baseline	16	1	27	1024	233	62.1	69.8	61.8	**55.7**	59.7
+ WTC										
- Avg-pooling	16	2	4,18	256	82	59.2	68.9	53.4	53.2	59.6
- CA-pooling	16	2	4,18	256	83	60.7	**70.5**	55.6	55.2	**59.9**
- Pixel-unshuffle	16	2	4,18	256	83	57.5	64.9	53.5	52.9	57.5
+ RPE	8	3	4,18,27	256	160	**63.0**	70.3	**64.6**	55.6	59.8

global encoding not only preserves but also enhances fine-grained visual understanding required for text-rich and structured visual tasks.

4.4 Abalation Studies

Effect of Individual Components To better understand the contributions of the proposed components in ViT-PVC, we conduct a comprehensive ablation study. As shown in Tab. 3, introducing WTC using average pooling substantially improves inference efficiency (from 233 ms to 82 ms). However, the sharp token reduction (1024 to 256) weakens fine-grained representation, leading to severe OCR&Chart degradation. In contrast, our content-adaptive pooling alleviates this issue and improves performance on the general subset, highlighting the importance of dynamically modeling local semantics, while incurring nearly no extra cost compared to average pooling. Utilizing a parameterized module like pixel-unshuffle as WTC causes notable performance degradation due to convergence challenges. Combining RPE with 3 WTC layers yields a favorable trade-off, delivering 1.5× higher efficiency (160 ms vs. 233 ms) and boosting average accuracy (63.0 vs. 62.1) across all benchmarks.

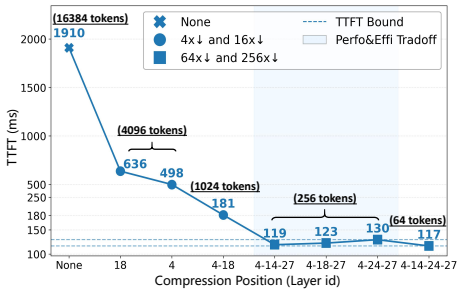


Fig. 5: Efficiency of varying compression position. Optimal trade-off are shown in the light blue region. 4× denotes a different compression ratio.

Table 4: Performance of baseline and ViT-PVC with the same Pre-alignment Stage

Encoding Strategy	Vision Encoder	PVC	Max Tokens↓	Avg↑	MME	POPE	HallusionBench	MMMU	AI2D	MMStar	TextVQA
GNE	MoonViT	✗	776	64.11	1522.97	85.38	39.94	46.44	78.04	49.93	72.93
GNE with PVC(Ours)	ViT-PVC	✓	380	**64.26**	1532.33	88.25	39.11	46.55	77.97	49.66	71.73

Joint modulation for superior trade-off. We evaluate the impact of different configurations of WTC on the TTFT latency of NaVLM-PVC under a patch size of $P = 10$ and an input resolution of 1024×1024. As shown in Fig. 5, adding 1 or 2 WTC layers significantly improves efficiency, with earlier inserted layers yielding greater latency reduction. Interestingly, the efficiency gain saturates after introducing 3 WTC layers, and further increasing their number or altering their placement offers negligible improvement.

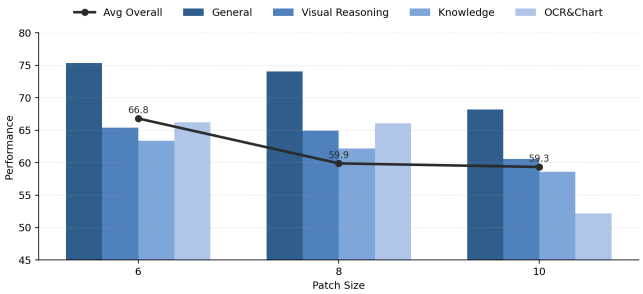
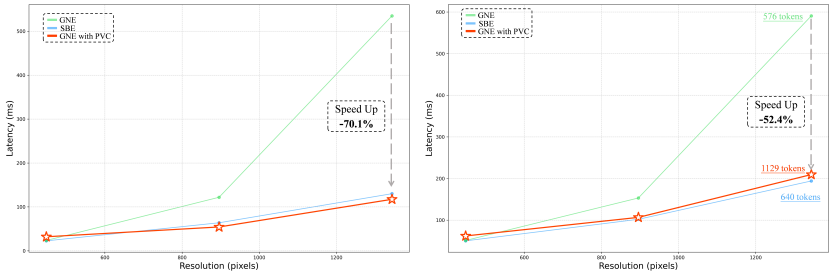


Fig. 6: Performance across different patch sizes. The average score decreases with larger patch size, with consistent drops in Visual Reasoning and OCR and Chart tasks.

Patch-Size Scaling Recent studies such as [4] [57] systematically investigated the effect of patch size in Vision Transformers across classification, detection, and segmentation tasks. Their results demonstrate a consistent trend: smaller patch sizes lead to stronger visual representation and improved downstream performance, provided the model is properly trained or adapted. We extend this investigation to multimodal large language model (MLLM) settings and observe a similar trend, as illustrated in Fig. 6. Based on empirical analysis and latency profiling, we adopt a patch size of 10 as a balanced configuration, which provides strong encoding capability while maintaining acceptable inference cost.

Pre-alignment stage ablations. To construct our ViT-PVC, we integrate the RPE and WTC modules into a well-pretrained vision encoder. This modification alters the original encoding flow and disrupts the pretrained feature distribution. To stabilize the representation space and preserve the benefits of pretraining, we introduce an additional pre-alignment stage using 4M samples. This stage effectively restores the embedding space and enables efficient optimization while retaining GNE's strong performance.

To further study the role of this pre-alignment process and ensure a fair comparison with existing approaches, we apply the same pre-alignment training to both MoonViT(GNE) and ViT-PVC(GNE with PVC), followed by identical pretraining (558K samples) and SFT (858K samples). As shown in Tab. 4, under the same data regime, ViT-PVC achieves comparable performance while using significantly fewer tokens.

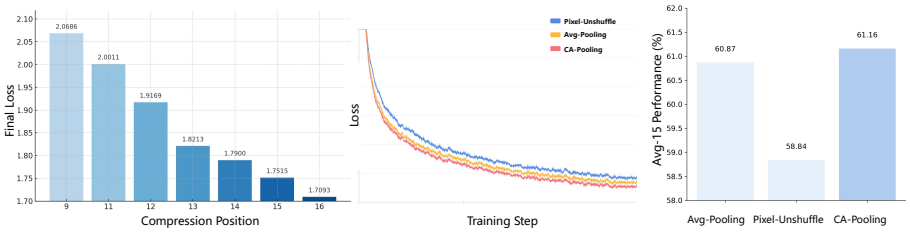


(a) Vision Encoder latency (b) MLLM latency

Fig. 7: (a) Evaluation of Vision Encoder latency of GNE(global native-resoluion encoding), SBE(slice-based encoding) and GNE with our proposed PVC method in different input resolution.(b) Evaluation of MLLM latency of GNE(global native-resoluion encoding), SBE(slice-based encoding) and GNE with our proposed PVC method in different input resolution.

Efficiency ablations of Encoding strategies. With the integration of our PVC framework, the proposed ViT-PVC(GNE with PVC) retains the advantages of GNE while progressively compressing visual tokens throughout the encoding stage. As illustrated in Fig. 7, this design enables faster inference than SBE methods, while the model maintains comparable performance to GNE models, as validated in Tab. 4. In essence, ViT-PVC combines the strengths of both GNE and SBE — achieving global contextual understanding with high efficiency.

Experiment on Different WTC As shown in Fig. 8(a), when integrating WTC with pixel-unshuffle into the ViT architecture, we observe that placing WTC at lower layers leads to greater disruption of training convergence. This demonstrates the point discussed in Appendix **??**, namely that inserting additional modules directly into the ViT disturbs the pre-trained feature modeling flow, thereby necessitating an extra pre-alignment stage to stabilize the network. Moreover, in Fig. 8(b), we find that avg-pooling, despite being parameter-free, achieves better convergence stability than pixel-unshuffle when inserted at the same positions (*i.e.*, 4-th and 18-th layers). Finally, our content-adaptive pooling further enhances convergence stability and efficiency, shown in Fig. 8(b), yielding faster convergence and achieving superior average performance across 15 benchmarks, as illustrated in Fig. 8(c).



(a) Final loss of different position (b) Loss curve of different WTC (c) Performance of difference WTC

Fig. 8: Evaluations on hyper-parameters of WTC type and layer position. All the ablation is conducted on SigLIP2 at 1024×1024 with patch size 16. (a) illustrates the final loss when using pixel-unshuffle in different layer position. (b) shows the training loss curve when using different WTC. (c) gives the performance comparison of different WTC.

5 Related works

Sliced-based Visual Encoding. Slice-based encoding improves efficiency by dividing high-resolution images into smaller slices, processing each independently with a ViT, and then concatenating their features [20, 29, 31, 70]. LLaVA-UHD [17] computes an optimal slicing strategy while preserving aspect ratios, and SPHINX [29] reduces sequence length by padding partial slices with "/n" tokens.

Native-Resolution Image Encoding. native-resolution encoding directly processes images at their original resolution without resizing [10], preserving holistic spatial information. Recent works such as Qwen2.5-VL [2] and MiMo-VL [48] improve this approach through window attention mechanism and position embedding like 2d-RoPE.

Visual Feature Compression. To handle the long token sequences output by ViTs, existing MLLMs commonly rely on projectors for feature compression and alignment. Common designs include MLPs [32], Q-Formers [26], and Re-samplers [1, 3], with recent variants exploring pooling and attention-based compression [5, 38].

6 Conclusion

We introduced NaVLM-PVC, an MLLM built upon Progressive Visual Compression (PVC) for efficient native-resolution encoding. By integrating refined patch embedding and windowed token compression, a standard ViT will be reconfigured into an efficient yet generalizable encoder. Extensive evaluations demonstrate competitive performance and significantly reduced inference latency compared with SoTA ViTs and MLLMs.

References

1. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023) 15
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025) 15
3. Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Han, S.C.T., Gong, Z., et al.: Jean-baptiste alayrac jeff donahue pauline luc antoine miech. arXiv preprint arXiv:2204.14198 (2022) 15
4. Beyer, L., Izmailov, P., Kolesnikov, A., Caron, M., Kornblith, S., Zhai, X., Minderer, M., Tschannen, M., Alabdulmohsin, I., Pavetic, F.: Flexivit: One model for all patch sizes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14496–14506 (2023) 3, 7, 13
5. Cha, J., Kang, W., Mun, J., Roh, B.: Honeybee: Locality-enhanced projector for multimodal llm. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13817–13827 (2024) 15
6. Chen, L., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., Duan, H., Wang, J., Qiao, Y., Lin, D., Zhao, F.: Are we on the right way for evaluating large vision-language models? (2024), https://arxiv.org/abs/2403.20330 11
7. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024) 11
8. Choromanski, K., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J., Mohiuddin, A., Kaiser, L., et al.: Rethinking attention with performers. arXiv preprint arXiv:2009.14794 (2020) 7
9. Dao, T.: Flashattention-2: Faster attention with better parallelism and work partitioning (2023), https://arxiv.org/abs/2307.08691 11
10. Dehghani, M., Mustafa, B., Djolonga, J., Heek, J., Minderer, M., Caron, M., Steiner, A., Puigcerver, J., Geirhos, R., Alabdulmohsin, I.M., et al.: Patch n'pack: Navit, a vision transformer for any aspect ratio and resolution. Advances in Neural Information Processing Systems **36**, 2252–2274 (2023) 15
11. Dong, X., Zhang, P., Zang, Y., Cao, Y., Wang, B., Ouyang, L., Zhang, S., Duan, H., Zhang, W., Li, Y., Yan, H., Gao, Y., Chen, Z., Zhang, X., Li, W., Li, J., Wang, W., Chen, K., He, C., Zhang, X., Dai, J., Qiao, Y., Lin, D., Wang, J.: Internlm-xcomposer2-4khd: A pioneering large vision-language model handling resolutions from 336 pixels to 4k hd (2024), https://arxiv.org/abs/2404.06512 5
12. Fan, Q., You, Q., Han, X., Liu, Y., Tao, Y., Huang, H., He, R., Yang, H.: Vitar: Vision transformer with any resolution. arXiv preprint arXiv:2403.18361 (2024) 2
13. Fang, Y., Zhu, L., Lu, Y., Wang, Y., Molchanov, P., Kautz, J., Cho, J.H., Pavone, M., Han, S., Yin, H.: Vila^2: Vila augmented vila. arXiv preprint arXiv:2407.17453 (2024) 11
14. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., Wu, Y., Ji, R.: Mme: A comprehensive evaluation benchmark for multimodal large language models (2024), https://arxiv.org/abs/2306.13394 11

15. Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., Manocha, D., Zhou, T.: Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models (2024), https://arxiv.org/abs/2310.14566 11
16. Guo, J., Zheng, T., Bai, Y., Li, B., Wang, Y., Zhu, K., Li, Y., Neubig, G., Chen, W., Yue, X.: Mammoth-vl: Eliciting multimodal reasoning with instruction tuning at scale (2025), https://arxiv.org/abs/2412.05237 9
17. Guo, Z., Xu, R., Yao, Y., Cui, J., Ni, Z., Ge, C., Chua, T.S., Liu, Z., Huang, G.: Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. In: European Conference on Computer Vision. pp. 390–406. Springer (2024) 2, 3, 9, 15
18. Hong, W., Yu, W., Gu, X., Wang, G., Gan, G., Tang, H., Cheng, J., Qi, J., Ji, J., Pan, L., et al.: Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. arXiv e-prints pp. arXiv–2507 (2025) 3
19. Hu, A., Xu, H., Ye, J., Yan, M., Zhang, L., Zhang, B., Li, C., Zhang, J., Jin, Q., Huang, F., et al.: mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. arXiv preprint arXiv:2403.12895 (2024) 2
20. Huang, R., Ding, X., Wang, C., Han, J., Liu, Y., Zhao, H., Xu, H., Hou, L., Zhang, W., Liang, X.: Hires-llava: Restoring fragmentation input in high-resolution large vision-language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29814–29824 (2025) 2, 15
21. Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., Farhadi, A.: A diagram is worth a dozen images (2016) 11
22. Kuckreja, K., Danish, M.S., Naseer, M., Das, A., Khan, S., Khan, F.S.: Geochat: Grounded large vision-language model for remote sensing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27831–27840 (2024) 2
23. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024) 2, 9, 11
24. Li, B., Ge, Y., Ge, Y., Wang, G., Wang, R., Zhang, R., Shan, Y.: Seed-bench: Benchmarking multimodal large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13299–13308 (2024) 11
25. Li, F., Zhang, R., Zhang, H., Zhang, Y., Li, B., Li, W., Ma, Z., Li, C.: Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. arXiv preprint arXiv:2407.07895 (2024) 2, 3, 9
26. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023) 15
27. Li, Y., Mao, H., Girshick, R., He, K.: Exploring plain vision transformer backbones for object detection (2022), https://arxiv.org/abs/2203.16527 6
28. Li, Z., Yang, B., Liu, Q., Ma, Z., Zhang, S., Yang, J., Sun, Y., Liu, Y., Bai, X.: Monkey: Image resolution and text label are important things for large multi-modal models. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26763–26773 (2024) 2
29. Liu, D., Zhang, R., Qiu, L., Huang, S., Lin, W., Zhao, S., Geng, S., Lin, Z., Jin, P., Zhang, K., et al.: Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. arXiv preprint arXiv:2402.05935 (2024) 15

30. Liu, F., Zhang, X., Peng, Z., Guo, Z., Wan, F., Ji, X., Ye, Q.: Integrally migrating pre-trained transformer encoder-decoders for visual object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6825–6834 (2023) 9
31. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26296–26306 (2024) 15
32. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023) 15
33. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? In: European conference on computer vision. pp. 216–233. Springer (2024) 11
34. Liu, Y., Tian, L., Zhou, X., Gao, X., Yu, K., Yu, Y., Zhou, J.: Points1.5: Building a vision-language model towards real world applications (2024), https://arxiv.org/abs/2412.08443 11
35. Liu, Y., Li, Z., Huang, M., Yang, B., Yu, W., Li, C., Yin, X.C., Liu, C.L., Jin, L., Bai, X.: Ocrbench: on the hidden mystery of ocr in large multimodal models. Science China Information Sciences **67**(12) (Dec 2024). https://doi.org/10.1007/s11432-024-4235-6, http://dx.doi.org/10.1007/s11432-024-4235-6 11
36. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows (2021), https://arxiv.org/abs/2103.14030 6, 9
37. Liu, Z., Dong, Y., Liu, Z., Hu, W., Lu, J., Rao, Y.: Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution. arXiv preprint arXiv:2409.12961 (2024) 3
38. Liu, Z., Dong, Y., Liu, Z., Hu, W., Lu, J., Rao, Y.: Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution (2025), https://arxiv.org/abs/2409.12961 15
39. Liu, Z., Dong, Y., Wang, J., Liu, Z., Hu, W., Lu, J., Rao, Y.: Ola: Pushing the frontiers of omni-modal language model with progressive modality alignment. arXiv e-prints pp. arXiv–2502 (2025) 3
40. Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.W., Galley, M., Gao, J.: Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts (2024), https://arxiv.org/abs/2310.02255 11
41. Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.W., Zhu, S.C., Tafjord, O., Clark, P., Kalyan, A.: Learn to explain: Multimodal reasoning via thought chains for science question answering. Advances in Neural Information Processing Systems **35**, 2507–2521 (2022) 11
42. Lv, T., Huang, Y., Chen, J., Zhao, Y., Jia, Y., Cui, L., Ma, S., Chang, Y., Huang, S., Wang, W., et al.: Kosmos-2.5: A multimodal literate model. arXiv preprint arXiv:2309.11419 (2023) 2
43. Masry, A., Long, D.X., Tan, J.Q., Joty, S., Hoque, E.: Chartqa: A benchmark for question answering about charts with visual and logical reasoning (2022), https://arxiv.org/abs/2203.10244 11
44. Mathew, M., Karatzas, D., Jawahar, C.V.: Docvqa: A dataset for vqa on document images (2021), https://arxiv.org/abs/2007.00398 11
45. OpenAI: Gpt-4o-mini: Advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence (Sep 2024), accessed: 2024-09-24 11

46. Peng, B., Alcaide, E., Anthony, Q., Albalak, A., Arcadinho, S., Biderman, S., Cao, H., Cheng, X., Chung, M., Grella, M., GV, K.K., He, X., Hou, H., Lin, J., Kazienko, P., Kocon, J., Kong, J., Koptyra, B., Lau, H., Mantri, K.S.I., Mom, F., Saito, A., Song, G., Tang, X., Wang, B., Wind, J.S., Wozniak, S., Zhang, R., Zhang, Z., Zhao, Q., Zhou, P., Zhou, Q., Zhu, J., Zhu, R.J.: Rwkv: Reinventing rnns for the transformer era (2023), https://arxiv.org/abs/2305.13048 7
47. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards vqa models that can read (2019), https://arxiv.org/abs/1904.08920 11
48. Team, C., Yue, Z., Lin, Z., Song, Y., Wang, W., Ren, S., Gu, S., Li, S., Li, P., Zhao, L., Li, L., Bao, K., Tian, H., Zhang, H., Wang, G., Zhu, D., Cici, He, C., Ye, B., Shen, B., Zhang, Z., Jiang, Z., Zheng, Z., Song, Z., Luo, Z., Yu, Y., Wang, Y., Tian, Y., Tu, Y., Yan, Y., Huang, Y., Wang, X., Xu, X., Song, X., Zhang, X., Yong, X., Zhang, X., Deng, X., Yang, W., Ma, W., Lv, W., Zhuang, W., Liu, W., Deng, S., Liu, S., Chen, S., Yu, S., Liu, S., Wang, S., Ma, R., Wang, Q., Wang, P., Chen, N., Zhu, M., Zhou, K., Zhou, K., Fang, K., Shi, J., Dong, J., Xiao, J., Xu, J., Liu, H., Xu, H., Qu, H., Zhao, H., Lv, H., Wang, G., Zhang, D., Zhang, D., Zhang, D., Ma, C., Liu, C., Cai, C., Xia, B.: Mimo-vl technical report (2025), https://arxiv.org/abs/2506.03569 2, 15
49. Team, G.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context (2024), https://arxiv.org/abs/2403.05530 11
50. Team, K., Du, A., Yin, B., Xing, B., Qu, B., Wang, B., Chen, C., Zhang, C., Du, C., Wei, C., et al.: Kimi-vl technical report. arXiv preprint arXiv:2504.07491 (2025) 2, 3
51. Team, Q.: Qwen2 technical report. arXiv preprint arXiv:2407.10671 (2024) 3, 5
52. Team, S.: Step-1.5v-mini. https://www.stepfun.com (2025), closed-source multimodal model by StepFun Team 11
53. Tong, P., Brown, E., Wu, P., Woo, S., IYER, A.J.V., Akula, S.C., Yang, S., Yang, J., Middepogu, M., Wang, Z., et al.: Cambrian-1: A fully open, vision-centric exploration of multimodal llms. Advances in Neural Information Processing Systems **37**, 87310–87356 (2024) 11
54. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025) 5
55. Vasu, P.K.A., Faghri, F., Li, C.L., Koc, C., True, N., Antony, A., Santhanam, G., Gabriel, J., Grasch, P., Tuzel, O., et al.: Fastvlm: Efficient vision encoding for vision language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 19769–19780 (2025) 11
56. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017) 6
57. Wang, F., Yu, Y., Wei, G., Shao, W., Zhou, Y., Yuille, A., Xie, C.: Scaling laws in patchification: An image is worth 50,176 tokens and more (2025), https://arxiv.org/abs/2502.03738 13
58. Wang, F., Chen, M., Li, Y., Wang, D., Wang, H., Guo, Z., Wang, Z., Shan, B., Lan, L., Wang, Y., et al.: Geollava-8k: Scaling remote-sensing multimodal large language models to 8k resolution. arXiv preprint arXiv:2505.21375 (2025) 2
59. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024) 2, 3, 11

60. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025) 2, 8, 9
61. Wu, Q., Xu, W., Liu, W., Tan, T., Liu, J., Li, A., Luan, J., Wang, B., Shang, S.: Mobilevlm: A vision-language model for better intra-and inter-ui understanding. arXiv preprint arXiv:2409.14818 (2024) 2
62. Wu, Z., Chen, X., Pan, Z., Liu, X., Liu, W., Dai, D., Gao, H., Ma, Y., Wu, C., Wang, B., et al.: Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. arXiv preprint arXiv:2412.10302 (2024) 11
63. Wu, Z., Chen, Z., Luo, R., Zhang, C., Gao, Y., He, Z., Wang, X., Lin, H., Qiu, M.: Valley2: Exploring multimodal models with scalable vision-language design (2025), https://arxiv.org/abs/2501.05901 11
64. XAI: RealWorldQA (Sep 2024), https://huggingface.co/datasets/xai-org/RealworldQA, accessed: 2024-09-24 11
65. Yang, S., Kautz, J., Hatamizadeh, A.: Gated delta networks: Improving mamba2 with delta rule. arXiv preprint arXiv:2412.06464 (2024) 7
66. Yao, K., Xu, N., Yang, R., Xu, Y., Gao, Z., Kitrungrotsakul, T., Ren, Y., Zhang, P., Wang, J., Wei, N., et al.: Falcon: A remote sensing vision-language foundation model. arXiv preprint arXiv:2503.11070 (2025) 2
67. Yao, T., Li, Y., Pan, Y., Mei, T.: Hiri-vit: Scaling vision transformer with high resolution inputs. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(9), 6431–6442 (2024) 2
68. Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., Zhu, H., Cai, T., Li, H., Zhao, W., He, Z., et al.: Minicpm-v: A gpt-4v level mllm on your phone. arXiv preprint arXiv:2408.01800 (2024) 2, 6, 9, 11
69. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., Wei, C., Yu, B., Yuan, R., Sun, R., Yin, M., Zheng, B., Yang, Z., Liu, Y., Huang, W., Sun, H., Su, Y., Chen, W.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi (2024), https://arxiv.org/abs/2311.16502 11
70. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11975–11986 (2023) 15
71. Zhang, Y., Liu, Y., Guo, Z., Zhang, Y., Yang, X., Zhang, X., Chen, C., Song, J., Zheng, B., Yao, Y., et al.: Llava-uhd v2: an mllm integrating high-resolution semantic pyramid via hierarchical window transformer. arXiv preprint arXiv:2412.13871 (2024) 2, 5, 10