InternLM-XComposer2-4KHD: A Pioneering Large Vision-Language Model Handling Resolutions from 336 Pixels to 4K HD

Xiaoyi Dong[*1,2], **Pan Zhang**[*1], **Yuhang Zang**[*1], **Yuhang Cao**[1,2], **Bin Wang**[1], **Linke Ouyang**[1],
Songyang Zhang[1], Haodong Duan[1], Wenwei Zhang[1], Yining Li[1], Hang Yan[1], Yang Gao[1], Zhe Chen[1]
Xinyue Zhang[1], Wei Li[1], Jingwen Li[1], Wenhai Wang[1,2], Kai Chen[1], Conghui He[3], Xingcheng Zhang[3],
Jifeng Dai[4,1], Yu Qiao[1], Dahua Lin[1,2,5], Jiaqi Wang[1,✉]
[1]Shanghai Artificial Intelligence Laboratory, [2]The Chinese University of Hong Kong,
[3]SenseTime Group, [4]Tsinghua University, [5] CPII under InnoHK
internlm@pjlab.org.cn

Abstract

The Large Vision-Language Model (LVLM) field has seen significant advancements, yet its progression has been hindered by challenges in comprehending fine-grained visual content due to limited resolution. Recent efforts have aimed to enhance the high-resolution understanding capabilities of LVLMs, yet they remain capped at approximately 1500 $\times$ 1500 pixels and constrained to a relatively narrow resolution range. This paper represents InternLM-XComposer2-4KHD, a groundbreaking exploration into elevating LVLM resolution capabilities up to 4K HD (3840 $\times$ 1600) and beyond. Concurrently, considering the ultra-high resolution may not be necessary in all scenarios, it supports a wide range of diverse resolutions from 336 pixels to 4K standard, significantly broadening its scope of applicability. Specifically, this research advances the patch division paradigm by introducing a novel extension: dynamic resolution with automatic patch configuration. It maintains the training image aspect ratios while automatically varying patch counts and configuring layouts based on a pre-trained Vision Transformer (ViT) (336 $\times$ 336), leading to dynamic training resolution from 336 pixels to 4K standard. Our research demonstrates that scaling training resolution up to 4K HD leads to consistent performance enhancements without hitting the ceiling of potential improvements. InternLM-XComposer2-4KHD shows superb capability that matches or even surpasses GPT-4V and Gemini Pro in 10 of the 16 benchmarks. The InternLM-XComposer2-4KHD model series with 7B parameters are publicly available at `https://github.com/InternLM/InternLM-XComposer`.

1 Introduction

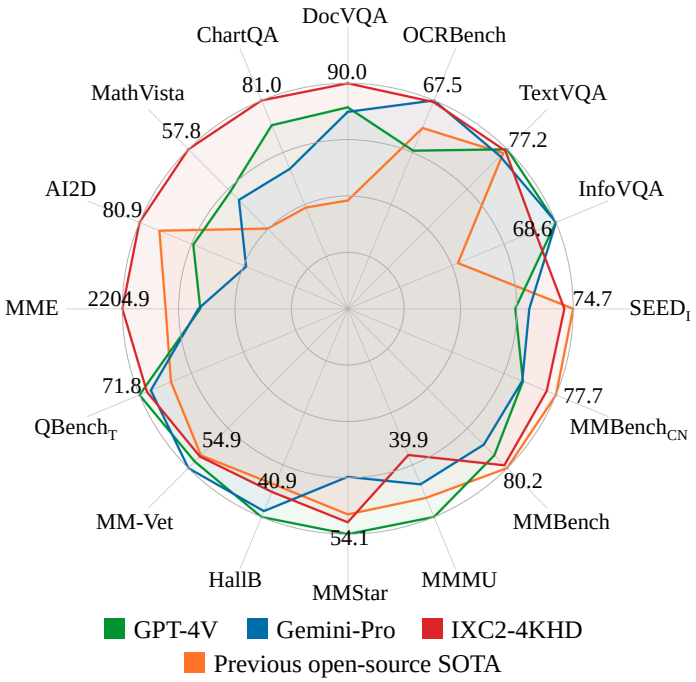
In recent years, the progress in Large Language Models (LLMs) (73; 92; 93; 39; 91; 10; 78; 29; 21) has provoked the development of Large Vision-Language Models (LVLMs). These models have demonstrated proficiency in tasks such as image captioning (17; 14) and visual-question-answering (VQA) (57; 31; 33; 107). Nevertheless, due to their limited resolution, they struggle with processing images containing fine details, such as charts (68), tables (87), documents (70), and infographics (69). This limitation constrains their practical applicability in real-world scenarios.

Recent advancements have aimed at enhancing the resolution of Large Vision-Language Models (LVLMs). Some approaches (66; 36; 97; 48) involve adapting high-resolution vision encoders

* indicates equal contribution.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

 https://doi.org/10.52202/079017-1348



Benchmark	Min	Median	Max
InfoVQA	355^2	1444^2	8541^2
DocVQA	246^2	1948^2	8243^2
OCRBench	20^2	408^2	5899^2
HallBench	147^2	724^2	4898^2
Qbench$_T$	312^2	873^2	4750^2
MathVista	58^2	444^2	4275^2
TextVQA	512^2	883^2	3990^2
MMMU	46^2	414^2	3166^2
MMStar	86^2	418^2	2598^2
AI2D	152^2	480^2	1500^2
ChartQA	245^2	668^2	1199^2
SEED$_I$	362^2	432^2	512^2
MMVet	181^2	422^2	512^2
MME	114^2	420^2	512^2
MMBench$_{CN}$	68^2	417^2	512^2
MMBench	68^2	417^2	512^2

Figure 1: **(a) Overview of InternLM-XComposer2-4KHD (IXC-4KHD) performance on benchmarks with different resolutions.** Our model based on InternLM2-7B (91) *matches or even surpasses GPT-4V (74) and Gemini Pro (90) in 10 of the 16 benchmarks*. **(b) Image resolution statistic of 16 benchmarks.** We report the minimum (Min), median, and maximum (Max) image area (resolution). Both the inter-/intra-benchmark resolution diversity are large, and we sort them by the maximum resolution.

directly. However, the Vision Transformer (ViT) architecture falls short when dealing with images of varying resolutions and aspect ratios, thereby restricting its ability to handle diverse inputs effectively. Alternatively, some methods (50; 59; 37; 51; 99; 55; 46) maintain the vision encoder's resolution, segmenting high-resolution images into multiple low-resolution patches. Yet, these methods are constrained by an inadequate resolution, typically around 1500 $\times$ 1500, which does not satisfy the demands of daily content, *e.g.*, website screenshots (85), document pages (70), and blueprints (69). Furthermore, they are confined to either a few predefined high-resolution settings (36; 97; 48; 50; 51; 55; 46; 66; 59) or a limited range of resolutions (101; 37; 99), thereby restricting their utility across a variety of applications.

In this work, we introduce InternLM-XComposer2-4KHD, a pioneering model that for the first time expands the resolution capabilities of Large Vision-Language Models (LVLMs) to 4K HD and even higher, thereby setting a new standard in high-resolution vision-language understanding. Designed to handle a broad range of resolutions, InternLM-XComposer2-4KHD supports images with any aspect ratio from 336 pixels up to 4K HD, facilitating its deployment in real-world contexts.

InternLM-XComposer2-4KHD follows patch division (50; 46) paradigm and enhances it by incorporating an innovative extension: dynamic resolution with automatic patch configuration. To be specific, scaling the resolution of Large Vision-Language Models (LVLMs) to 4K HD and even higher standard is far beyond merely increasing the number of patches. It involves a nuanced approach to overcoming specific challenges: (1) **Dynamic Resolution and Automatic Patch Configuration**: Addressing the scarcity of high-resolution training data, our framework introduces a strategy that dynamically adjusts resolution alongside an automatic layout configuration. During training, it maintains the original aspect ratios of images while adaptively altering patch (336 $\times$ 336) layouts and counts. This results in a training resolution that exceeds the original image resolutions, reaching up to 4KHD, addressing the shortfall of high-resolution data. (2) **Handling Variability in Patch Layouts**: Despite the apparent simplicity of dynamic resolution training, the variability in patch layouts can heavily confuse LVLMs. To mitigate this, we introduce a newline token after each row of patch tokens to clearly delineate patch layouts, reducing training ambiguity and significantly boosting performance. (3) **Inference Beyond 4K Resolution:** Our observations reveal that, even when trained on images up to 4K resolution, the model can achieve additional performance improvements during inference by processing images at higher resolutions.

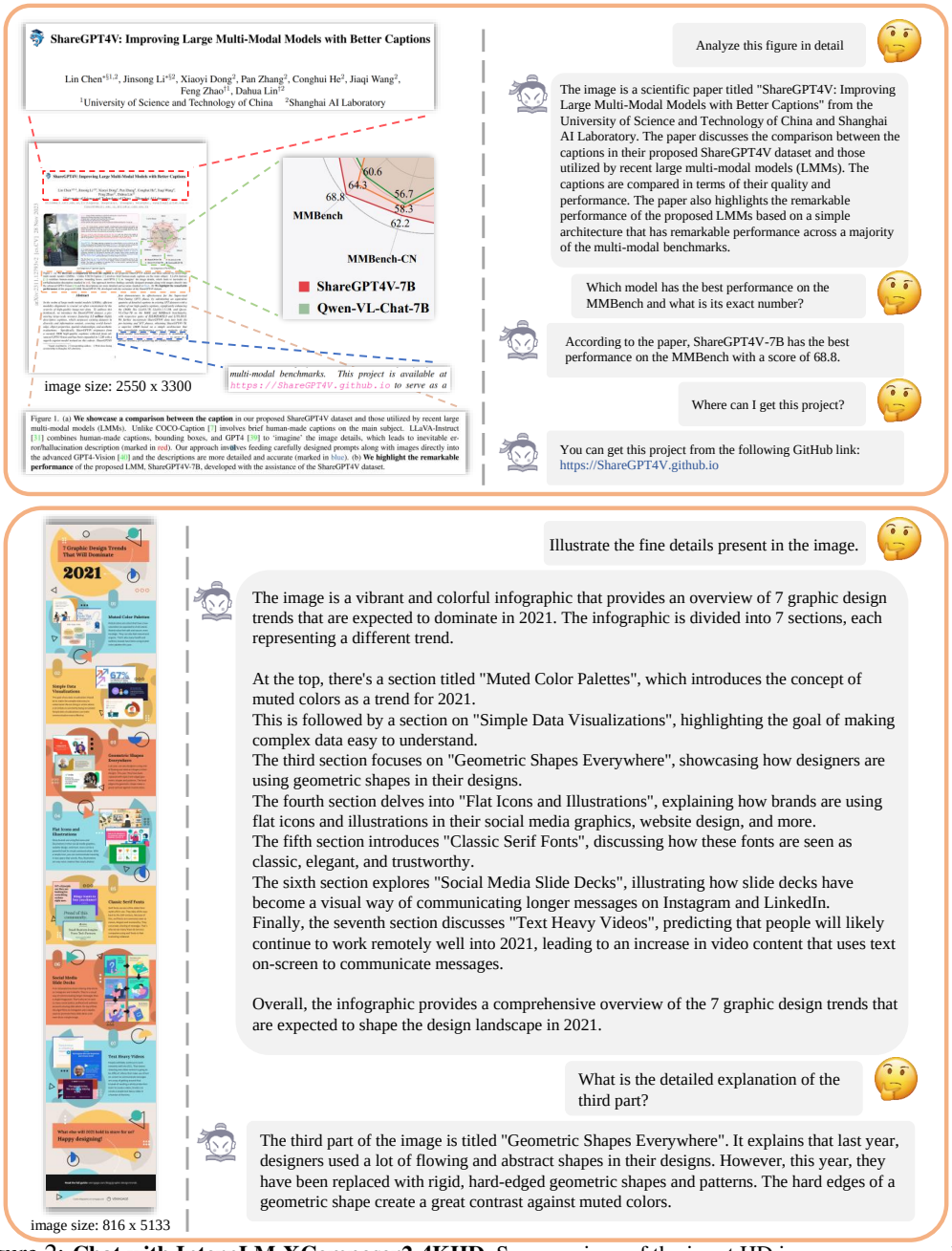


Figure 2: **Chat with InternLM-XComposer2-4KHD**. Some regions of the input HD images are zoomed in for better visualization. For more results please refer to the Supplementary materials.

Furthermore, scaling the training resolution up to 4K standard results in a consistent improvement in performance, highlighting the potential for training even beyond 4K resolution. This underscores the capacity for further enhancing model capabilities and suggests a promising trajectory for advancing the frontiers of high-resolution image processing within the domain of large vision-language models.

We evaluate our InternLM-XComposer2-4KHD on 16 diverse benchmarks spanning various domains, including 5 challenging OCR datasets (InfographicVQA(69), DocVQA(70), OCRBench(58), TextVQA(87), and ChartQA(68)). Compared to previous open-source LVLM models and closed-source APIs, our approach achieves SOTA results in 6 of 16 benchmarks, demonstrating competitive performance despite only 7B parameters. As shown in Figure 1, InternLM-XComposer2-4KHD even surpasses the performance of GPT4V (74) and Gemini Pro (90) across ten benchmarks. Notably, our method exhibits excellent performance on 5 challenging OCR datasets, over existing open-source LVLMs by a substantial margin.

https://doi.org/10.52202/079017-1348

2 Related Works

Large Vision-Language Models (LVLMs). Large Language Models (LLMs) (9; 76; 73; 23; 41; 92; 93; 39; 91; 108; 6; 78; 10) have gained significant attention due to their impressive performance in various language-related tasks such as text generation and question answering. Following this enthusiasm, recent Large Vision-Language Models (LVLMs) have emerged(74; 19; 16; 18; 28; 32; 113; 25; 110; 7; 47; 77; 102; 4), combining LLMs with vision encoders (79; 109; 89) to leverage the complementary strengths of language and vision modalities. By fusing textual and visual representations, LVLMs can ground language in visual contexts, enabling a more comprehensive understanding and generation of multimodal content (14; 20; 51; 5; 95; 27; 11; 60).

LVLMs for High-Resolution Understanding. Large Vision-Language Models (LVLMs) often employ CLIP-ViT as the visual encoder for vision-dependent tasks. However, the visual encoder's reliance on low resolutions, such as 224×224 or 336×336 pixels, limits its effectiveness for high-resolution tasks like OCR and document/chart perception. To enhance high-resolution understanding, recent works have primarily employed the following strategies: (1) High-resolution (HR) visual encoders or dual encoders catering to high-resolution (HR) and low-resolution (LR) inputs (66; 97; 36; 48). For instance, Vary (97) introduces a new image encoder supporting HR inputs, which are then concatenated with LR embeddings from the original CLIP visual encoder. Similarly, CogAgent (36) and Mini-Gemini (48) also separate HR and LR images using distinct vision encoders, subsequently merging their features using a cross-attention module. In contrast, our approach offers a more simplified solution and shows advantages for varying resolutions and aspect ratio inputs. (2) Cropped image patches (50; 59; 99; 101; 37; 51; 46). For example, Monkey (50) employs sliding windows to segment images into patches, subsequently processing them with LoRA fine-tuning. TextMonkey (59) further proposes shifted window attention and token resampler to consider the connections among different patches. Fuyu (7) eliminates the need for the image encoder by directly processing a raw image patch sequence. These approaches are confined to either a few predefined high-resolution settings (36; 97; 48; 50; 51; 55; 46; 66; 59) or a limited range of resolutions (37; 99). Conversely, our method devises a dynamic resolution and automatic path configuration strategy to support the scaling from 336 pixels to 4K resolution, and the maximum resolution is larger than previous approaches (*e.g.*, 1.5k for Monkey (50) and 1.2k for UReader (101)). For the first time, our approach discussed the challenges and solutions for handling variability in image feature patch layouts, ensuring effective training with dynamic high resolutions.

LVLMs for Document Understanding. Document understanding involves analyzing and comprehending various digital documents, such as figures, tables, and academic papers. Many document understanding tasks require models to handle high-resolution inputs, complex layouts, various aspect ratios, and diverse document formats. To enhance the capabilities of LVLMs for document understanding, several works have collected and constructed high-quality document instruction tuning data, including LLaVAR (112), mPLUG-DocOwl (100) and TGDoc (96). DocPediaDocPedia (30) processes document inputs in the frequency domain. Some previous works have improved document understanding ability by designing special modules for high-resolution inputs, such as HR and LR encoders (36; 97) or cropped image patches (101; 59; 99). Our InternLM-XComposer2-4KHD first scales to 4K resolution inputs and demonstrates strong document understanding ability on OCR-related benchmarks. Also, our approach also achieves comparable results to state-of-the-art open-sourced LVLMs on other general LVLM benchmarks like perception and reasoning (61; 57; 33; 15).

3 Method

3.1 Model Architecture.

The model architecture of InternLM-XComposer2-4KHD mainly follows the design of InternLM-XComposer2(27) (XComposer2 / IXC2 in the following for simplicity), including a light-weight Vision Encoder OpenAI ViT-Large/14, Large Language Model InternLM2-7B, and Partial LoRA for efficient alignment.

3.2 High-Resolution Input.

Dynamic Patch Configuration. Utilizing a static input image size for processing high-resolution images, particularly those with varying aspect ratios, is neither efficient nor effective. To overcome

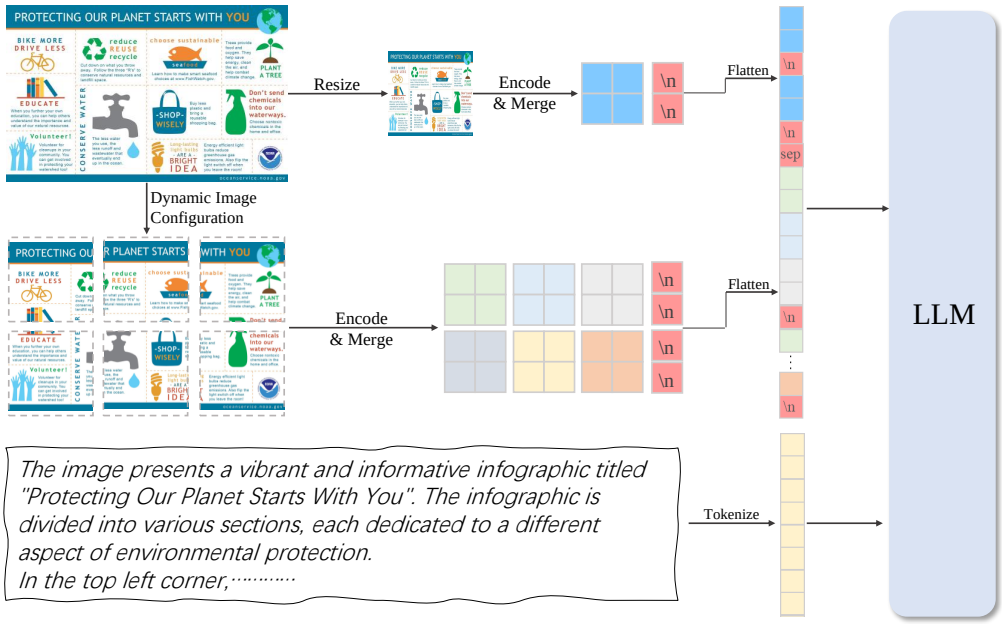


Figure 3: The framework of InternLM-XComposer2-4KHD. Our model processes the high-resolution image with a Dynamic Image Partition strategy and concatenates the image tokens with text tokens as LLM input.

this limitation, we introduce a dynamic patch configuration approach via image partitioning, as shown in Figure 3. Our method strategically segments the image into smaller patches, while maintaining the integrity of the original image's aspect ratio.

Given a maximum patch number $\mathcal{H}$, the image x with size $[h, w]$ is resized and padded to the new image $\hat{x}$ with size $[p_h \times 336, p_w \times 336]$. This process is subject to the following constraints:

$$p_w \times p_h \leq \mathcal{H};\ p_h = \lceil p_w \times h/w \rceil \tag{1}$$

here p_w and p_h represent the number of patches in each row and column, respectively. We then split the $\hat{x}$ into $p_h \times p_w$ non-overlapped patches. Each patch is a small image with 336×336 size and we treat these patches as individual inputs for the ViT.

In the following, we use 'HD-$\mathcal{H}$' to represent our high-resolution setting with the constraint of $\mathcal{H}$ patches. For example, the 'HD-9' allows up to 9 patches, including a range of resolutions such as 1008×1008, 672×1344, 336×3024, *etc.*

Global-Local Format. For each input image, we present it to the model with two views. The first is the global view, where the image is resized to a fixed size (in our case, 336 × 336). This provides a macro understanding of the image. Empirically, we have found this to be crucial for the LVLM to correctly understand the image. The second view is the local view. We divide the image into patches using the previously mentioned Dynamic Image Partition strategy and extract features from each patch. Following feature extraction, the patches are reassembled into a large feature map. The feature map is then flattened to the final local features after a straightforward token merging process.

Patch Layout Indicator. Given that an image will have a dynamic patch layout in our method, the number of tokens for each row can vary across different images. This variation will confuse the LVLM, making it difficult to determine which tokens belong to the same row of the image and which ones belong to the next row. This confusion hinders the LVLM's ability to understand the 2D structure of the image, which is crucial for comprehending structural image content such as documents, charts, and tables. To address this issue, we introduce a learnable newline ('\n') token at the end of each row of the image features before the flattening. Finally, we concatenate the global and local views, inserting a special separate ('sep') token between them to distinguish the two views.

3.3 Pre-Training

During the pre-training phase, the LLM is frozen while both the vision encoder and Partial LoRA are fine-tuned to align the visual tokens with the LLM following XComposer2(27). The pre-training data

https://doi.org/10.52202/079017-1348

Table 1: **Pre-Training Datasets**. The data are collected from diverse sources for the three objectives.

Task	Dataset
General Semantic Alignment	ShareGPT4V-PT (14), COCO (17), Nocaps (1), TextCaps (86), SBU (75), LAION400M (80), CC 3M (83)
World Knowledge Alignment	Concept Data (110)
Vision Capability Enhancement	WanJuan (35), Flicker(103), MMC-Inst(54), RCTW-17(84), CTW(106), LSVT(88), ReCTs(111), ArT(22)

Table 2: **Supervised Fine-Tuning Datasets**. We collect data from diverse sources to empower the model with different capabilities. The image resolution is also different for different tasks.

Task	Resolution Setting	Dataset
Caption	HD-25	ShareGPT4V (14), COCO (17),Nocaps (1)
General QA	HD-25	VQAv2 (3), GQA (38), OK-VQA (67), VD (26), RD(13), VSR(53),
Science QA	HD-25	AI2D (42), SQA (63), TQA(43), IconQA(65)
Chart QA	HD-25	DVQA (40), ChartQA, ChartQA-AUG (68)
Math QA	HD-25	MathQA (104), Geometry3K(62), TabMWP(64), CLEVR-MATH(52)/Super(49)
World Knowledge QA	HD-25	A-OKVQA (81),KVQA (82), ViQuAE(44)
OCR QA	HD-25	TextVQA(87), OCR-VQA(72), ST-VQA(8)
HD-OCR QA	HD-55	InfoVQA(69), DocVQA(70)
Conversation	–	LLaVA-150k (56), LVIS-Instruct4V (94), ShareGPT-en&zh (21), InternLM-Chat(91)

mainly follow the design in XComposer2 which is curated with **three objectives** in mind: 1) general semantic alignment, 2) world knowledge alignment, 3) vision capability enhancement. In this paper, we focus on high-resolution and structural image understanding. So we collected OCR and chart data from diverse sources to enhance this specific capability, as shown in Table.1.

In practice, we employ the OpenAI CLIP ViT-L-14-336 as the vision encoder. We keep the ViT resolution as 336×336 while adopting Dynamic Patch Configuration to handle higher-resolution images. For pretraining, we use the 'HD-25' configuration, which resizes the input image to a random larger resolution to generate more patches, with the constraint that the total number of patches does not exceed 25. For each image or patch tokens, the image token number is decreased to $1/4$ with a simple **merge operation**. We concatenate the nearby 4 tokens into a new token through the channel dimension, then align it with the LLM by an MLP. The 'sep' and '\n' tokens are randomly initialized. For the Partial LoRA, we set a rank of 256 for all the linear layers in the LLM decoder block. Our training process involves a batch size of 4096 and spans across 2 epochs. The learning rate linearly increases to 2×10^{-4} within the first 1% of the training steps. Following this, it decreases to 0 according to a cosine decay strategy. To preserve the pre-existing knowledge of the vision encoder, we apply a layer-wise learning rate (LLDR) decay strategy, and the decay factor is set to 0.90.

3.4 4KHD Supervised Fine-tuning (SFT)

After pre-training, we empower the model to understand high-resolution images and solve diverse challenges. Different from conventional perception tasks (*e.g.*, VQAv2, GQA) which typically answer questions based on the noticeable object in the image. OCR-related tasks depend on a detailed understanding of text within a high-resolution image. For instance, in InfoVQA, the length of the longer side of 50% of the images exceeds 2000 pixels. Low-resolution inputs can distort the dense text information, causing the model to fail in its understanding. However, we have observed a resolution saturation phenomena with perception tasks, where a higher resolution makes minor gains.

To address this, we introduce a mixed-resolution strategy for more efficient training. For tasks requiring high resolution, we employ the 'HD-55' setting during training. This allows for the input of 4K (3840×1600) images without necessitating additional image compression. These tasks are referred to as the HD-OCR QA tasks in Table 2. For the other tasks, we apply the 'HD-25' resolution setting for them. As in pre-training, we adopt the dynamic-resolution strategy during SFT, images are resized to fall within a range between their original size and the size specified by the 'HD' setting. This dynamic approach enhances the robustness of the LVLM against differences in input resolution, thereby enabling the LVLM to utilize a larger resolution during inference. For instance, we have observed that using the 'HD30' setting yields better results on most OCR-related tasks when the LVLM is trained under the 'HD25' setting.

In practice, we jointly train all the components with a batch size of 2048 over 3500 steps. Data from multiple sources are sampled in a weighted manner, with the weights based on the number of data from each sourced (See Appendix B.1 for more details). As the 'HD-55' setting has double image tokens than the 'HD-25', we adjust the data loader to enable different batch sizes for them and adjust their weight accordingly. The maximum learning rate is set to 5×10^{-5}, and each component has its

Table 3: **Comparison with closed-source APIs and previous open-source SOTAs.** Our InternLM-XComposer2-4KHD gets SOTA results in 6 of the 16 benchmarks with only 8B parameters, showing competitive results with current closed-source APIs. The best results are **bold** and the second-best results are underlined.

Method	Doc VQA	Chart QA	Info VQA	Text VQA	OCR Bench	MM Star	Math Vista	AI2D	MMMU	MME	MMB EN	MMB CN	SEED Image	QBench Test	MM-Vet	Hall Bench
Open-Source	(37)	(37)	(37)	(36)	(36)	(55)	(55)	(55)	(20)	(2)	(55)	(55)	(55)	(110)	(95)	(50)
Previous SOTA	8B	8B	8B	18B	18B	35B	35B	35B	40B	34B	35B	35B	35B	8B	17B	10B
	82.2	70.2	44.5	76.1	59.0	52.1	39.0	78.9	51.6	2050.2	**81.1**	79.0	**75.7**	64.4	54.5	39.3
Closed-source API																
GPT-4V	88.4	78.5	75.1	**78.0**	51.6	**57.1**	47.8	75.5	**56.8**	1,926.5	77.0	74.4	69.1	**74.1**	56.8	**46.5**
Gemini-Pro	88.1	74.1	**75.2**	74.6	**68.0**	42.6	45.8	70.2	47.9	1,933.3	73.6	74.3	70.7	70.6	**59.2**	45.2
IXC2-VL (27)	57.7	72.6	34.4	70.1	53.2	55.4	57.6	**81.2**	41.4	**2,220.4**	80.7	**79.4**	74.9	72.5	46.7	41.0
IXC2-4KHD	**90.0**	**81.0**	68.6	77.2	67.5	54.1	**57.8**	80.9	39.7	2,204.9	80.2	77.7	74.7	71.8	54.9	40.9

Table 4: **Comparison with open-source SOTA methods.** IXC2-4KHD outperforms competitors in most benchmarks. The best results are **bold** and the second-best results are underlined.

Method	LLM	MMStar	MathVista	AI2D	MME^P	MME^C	MMB	MMB^{CN}	$SEED^I$	$QBench^T$	MM-Vet
Qwen-VL-Chat	Qwen-7B	37.5	33.8	63.0	1,487.5	360.7	60.6	56.7	58.2	61.7	47.3
ShareGPT4V	Vicuna-7B	33.0	25.8	58.0	1,567.4	376.4	68.8	62.2	69.7	-	37.6
Monkey	Qwen-7B	38.3	34.8	62.5	1,522.4	401.4	72.4	67.5	68.9	-	33.0
CogVLM-17B	Vicuna-7B	36.5	34.7	63.3	-	-	65.8	55.9	68.8	-	54.5
LLaVA-XTuner	InernLM2-20B	-	24.6	65.4	-	-	75.1	73.7	70.2	-	37.2
LLaVA-1.5	Vicuna-13B	32.8	26.1	61.1	1,531.3	295.4	67.7	63.6	68.2	61.4	35.4
LLaVA-Next	Vicuna-13B	38.3	32.4	72.2	1,445.0	296.0	70.0	68.5	71.4	-	44.9
InternLM-XC (27)	InernLM-7B	-	29.5	56.9	1,528.4	391.1	74.4	72.4	66.1	64.4	35.2
IXC2-VL	InernLM2-7B	**55.4**	57.6	**81.2**	**1,712.0**	530.7	**80.7**	**79.4**	**74.9**	**72.5**	46.7
IXC2-4KHD	InernLM2-7B	54.1	**57.8**	80.9	1,655.9	**548.9**	80.2	77.7	74.7	71.8	**54.9**

own unique learning strategy. For the vision encoder, we set the LLDR to 0.9, which aligns with the pretraining strategy. For the LLM, we employ a fixed learning rate scale factor of 0.2. This slows down the update of the LLM, achieving a balance between preserving its original capabilities and aligning it with vision knowledge. It takes almost 40 hours with 256 A100 GPUs.

4 Experiments

In this section, we validate the benchmark performance of our InternLM-XComposer2-4KHD (IXC2-4KHD in the following for simplicity) after supervised fine-tuning.

4.1 LVLM Benchmark results.

In Table 3 and Table 4, we compare our IXC2-4KHD on a list of benchmarks with both SOTA open-source LVLMs and closed-source APIs. Here we report results in DocVQA(70), ChartQA(68), InfographicVQA(69), TextVQA(87), OCRBench(58), MMStar(15), MathVista(61), MMMU(107), AI2D(42), MME (31), MMBench (MMB) (57), MMBench-Chinese (MMB^{CN}) (57), SEED-Bench Image Part ($SEED^I$)(45), QBench-Testset ($QBench^T$)(98), MM-Vet (105), HallusionBench (HallB)(34). The evaluation is mainly conducted on the OpenCompass VLMEvalKit(24) for the unified reproduction of the results.

Comparison with Closed-Source APIs. As demonstrated in Table 3, IXC2-4KHD exhibits competitive performance across a variety of benchmarks, rivaling that of Closed-Source APIs. Owing to its high-resolution input, IXC2-4KHD achieves a score of 90.0% on DocVQA and 81.0% on ChartQA, thereby surpassing GPT-4V and Gemini-Pro with a non-trivial margin. In the challenging InfographicVQA task, our model is the first open-source model that is close to the performance of Closed-Source APIs, exceeding the performance of previous open-source models by nearly 20%. In addition to OCR-related tasks, IXC2-4KHD is a general-purpose Large Vision-Language Modal that excels in semantic-level tasks, demonstrating competitive results.

Comparison with Open-Source Models. We also conduct a comprehensive comparison with open-source LVLMs under a similar model scale. As shown in Table 4, our model significantly outperforms existing open-source models, achieving competitive results across all benchmarks.

High-resolution Understanding Evaluation. Then we compare IXC2-4KHD with models that are specifically designed for high-resolution understanding tasks. We report the results of 5 high-resolution benchmarks in Table 5, as a general LVLM, IXC2-4KHD shows superb performance on

https://doi.org/10.52202/079017-1348

Table 5: **High-resolution Evaluation.** IntenrLM-XComposer2-4KHD has the largest input resolution and outperforms open-source LVLMs which are specifically tuned for document understanding.

Model	Model Size	Max Resolution	DocVQA^{Test}	ChartQA^{Test}	InfoVQA^{Test}	TextVQA^{Val}	OCRBench
TextMonkey(59)	9B	896x896	73.0	66.9	28.6	65.6	55.8
LLaVA-UHD (99)	13B	1008x672	—	—	—	67.7	—
CogAgent (36)	17B	1024x1024	81.6	68.4	44.5	76.1	59.0
UReader (101)	7B	1120x896	65.4	59.3	42.2	57.6	—
DocOwl 1.5 (37)	8B	1344x1344	82.2	70.2	50.7	68.6	—
IXC2-4KHD	8B	3840x1600	**90.0** (+7.8)	**81.0** (+10.8)	**68.6** (+17.9)	**77.2** (+1.2)	**67.5** (+8.5)

Table 6: **Influence of Inference Resolution.** The model achieves better performance on text-related tasks when the inference resolution is higher than its training resolution.

Train	Eval	Doc	Info	Text	Chart	MMB	MME	SEED*
HD9	HD9	79.4	50.5	73.8	78.2	79.5	2,201	76.6
	HD16	83.0	58.6	74.3	75.8	79.3	2,198	76.7
HD16	HD16	84.9	60.8	75.7	80.1	80.2	2,129	75.7
	HD25	85.9	62.1	75.8	79.1	80.1	2,100	75.4
HD25	HD25	87.0	63.6	76.0	80.3	78.5	2,209	74.9
	HD30	87.4	64.6	76.2	79.4	78.9	2,173	74.3

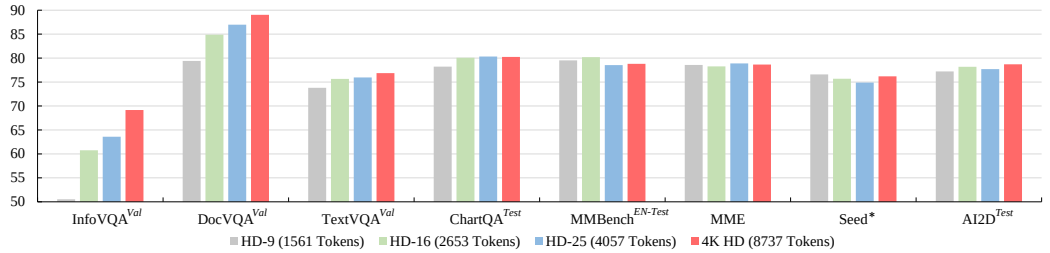


Figure 4: **Influence of Training Resolution.** High-resolution training is critical for HD-OCR tasks, while its gain on other tasks is minor.

these tasks and outperforms competitors with a large margin. For example, IXC2-4KHD gets 68.6% on InfographicVQA, surpassing recent DocOwl 1.5 with +17.9%. For the OCRBench, IXC2-4KHD gets 67.5%, outperforms CogAgent with +8.5%.

4.2 Dive into Resolution

High-Resolution Training is Critical for HD-OCR tasks. We study four resolution settings: HD-9 (1561 image tokens at most, we simply the statement in the following), HD-16 (2653 tokens), HD-25 (4057 tokens), and 4KHD (8737 tokens). Here we report the validation set of InfoVQA, DocVQA, and TextVQA, test set of ChartQA and AI2D, MMBench EN-Test, and a 2k subset of SEEDBench (we denote it as SEED*). In the following, we report results on the above benchmarks by default.

As illustrated in Fig.4, we note a significant improvement in the HD-OCR tasks as the resolution increases. For instance, the model achieves only a 50.5% score on the InfographicVQA with the HD-9 setting. However, when we switch to the HD-16 setting, we observe a performance gain of +10.2%. The performance continues to improve as the resolution increases, with saturation not observed even for the 4KHD setting. Due to computational constraints, we defer the exploration of the upper bound of improvement to future work. In terms of other OCR-related tasks, the performance gain attributable to increased resolution is relatively minor. For the perception-related benchmarks, performance is saturated on the resolution that only has negligible difference between the four settings.

Higher Inference Resolution Leads to better results on Text-related Tasks. An intriguing observation from our experiments is that our model, when inferring with a higher resolution, tends to yield improved results on text-related tasks. We present the results of HD-9, HD-16, and HD-25 in Table 6. For instance, IXC2-HD9 achieves a 50.5% score on InfographicVQA. When we infer with HD16, we see a performance gain of +8.1%, without additional training. Similar improvements are also observed with IXC2-HD16 and IXC2-HD25. We posit that the dynamic image token length used in training enhances the robustness of the LVLM, leading to better results when the text in the image is more 'clear' in the higher-resolution input. Conversely, the results on ChartQA consistently degrade under this setting. This could be due to the model becoming confused about the chart structure when

Table 7: **(a) Influence of Indicator '\n' in the Image Features.** '\n' helps LVLM understand structural images when the input resolution is dynamic and large. **(b) Ablation on Token Merging Operation.** Both the simple concatenation operation and the C-Abstractor works well.

Model	'\n'	Doc	Info	Text	Chart	MMB	MME	SEED*
HD9	×	79.5	50.3	74.0	78.2	79.1	2206	75.9
HD9	✓	79.4	50.5	73.8	78.2	79.5	2201	**76.6**
4KHD	×	88.1	67.4	75.9	80.4	79.9	**2232**	76.4
4KHD	✓	**89.0**	**69.3**	**77.2**	**81.0**	**80.2**	2205	76.2

Strategy	Doc	Info	Text	Chart	MMB	MME	SEED*
Re-Sampler	86.2	67.1	75.3	78.8	79.6	2124	74.2
C-Abstractor	88.6	69.5	77.1	80.6	80.4	2236	76.7
Concat	89.0	69.3	77.2	81.0	80.2	2205	76.2

Table 8: **Influence of Global-View in the Input.** Global-view is critical for most benchmarks.

Model	Doc	Info	Text	Chart	MMB	MME	SEED*
HD9	79.4	50.5	73.8	78.2	79.5	2201	76.6
+ w/o global-view	78.1	47.9	71.2	77.9	75.1	2019	76.2

Table 9: **Strategy-Level comparison between LLaVA-Next and our IXC2-4KHD.** Our strategy reaches better performance under a similar image token number constrain.

Model Nmae	HD Strategy	Image Tokens	Max Resolution	DocVQA	TextVQA	ChartQA	MMBench
LLaVA-Next	LLaVA-Next	2880	672x672	78.2	52.0	69.5	72.1
IXC-LLaVA	LLaVA-Next	2880	672x672	78.9	71.5	74.1	77.6
IXC-4KHD	HD9	1440	1008x1008	79.4	73.8	78.2	79.5
IXC-4KHD	HD16	2448	1344x1344	84.9	75.7	80.1	80.2

the resolution is increased. Additionally, similar to the observation from Figure 4, the impact of resolution on perception-related benchmarks appears to be quite minor.

4.3 High-Resolution Strategy Ablation

The Role of Global-View. We first examine the impact of the global view in our Global-Local Format. As indicated in Table 8(a), we find that the global view is essential for the LVLM to accurately comprehend the input image. When it is removed, the model performs worse across all benchmarks. For instance, the model experiences a -4.4% drop in performance on the MMBench EN-Test without the global view. We contend that the global view offers a general macro understanding of the image, which the model struggled to derive from the large number of tokens in the local view.

The Role of the Newline Token. We incorporate a special newline token at the end of each row of the image features before the flattening operation. This token serves as an indicator of the image's 2D structure. We examine its impact on both the HD-9 and 4KHD strategies in Table 7(a). When a fixed high-resolution strategy HD-9 is employed, we observe that the benefit derived from the newline token is minor. This could be attributed to the LVLM's ability to handle limited differences in image ratios after training. However, when we implement a more challenging 4KHD (HD-25 + HD-55) strategy, which exhibits significant diversity in both image ratio and token number, the LVLM demonstrates a notable decline in performance on OCR-related tasks without the newline indicator. This finding supports our hypothesis that the LVLM struggles to comprehend the shape of the image when the image tokens are directly flattened into a 1D sequence. The newline token can assist the model in better understanding the structure of the image.

Influence of Token Merging Operation. In practice, we employ a simple merging operation that concatenates four adjacent tokens along the channel dimension. We have found this approach to be effective in reducing the number of image tokens efficiently. Here we study the influence of different token-merging operations under the 4KHD setting. In Table 7(b), we study two additional strategies: Re-Sampler(5) and C-Abstractor(12), with their default setting and the same compressing rate 0.25, *i.e.*, reducing an image with 576 tokens to 144 tokens. Results show that both concatenation and C-Abstractor work well and get similar results on most benchmarks, this observation is also consistent with the study in MM-1(71) that the influence of the connector is minor. However, the Re-Sampler performs worse than the other methods with a noticeable margin. We argue this is caused by the learnable queries used for gathering information requiring a great number of data for training, our pre-training data is somewhat lightweight for it to converge fully.

https://doi.org/10.52202/079017-1348

Table 10: **Inference Efficieny Analysis.** The image token number mainly inference the prefix speed, and their difference in the decoding part is neglectable.

HD	image tokens	prefix encoding time	per-token decoding speed	time to generation 2048 new tokens
HD9	1440	0.2845	0.0982	201.4
HD16	2448	0.3966	0.0983	201.7
HD25	3744	0.5513	0.0981	201.5

Strategy-Level Comparison. For a fair comparison, we trained a new model with the identical architecture, training strategy, and dataset as our original model, but with one key modification: we adopted the high-resolution strategy from LLaVA-Next. We name it as IXC-LLaVA and compare it with 1) LLaVA-Next 0530 official results and 2) IXC2-4KHD under HD-9/16 setting. The results in Table.9 demonstrate that IXC-LLaVA achieves promising performance across all six benchmarks, leveraging the benefits of additional training data and advanced IXC architecture design. However, it is outperformed by IXC2-4KHD HD-9, which utilizes fewer image tokens yet yields better results. This fair comparison underscores the efficiency and effectiveness of our proposed high-resolution strategy, highlighting its advantages over the LLaVA-Next approach.

Inference Efficiency Analysis. Our model processes high-resolution images with numerous image tokens, and here we study its inference efficiency in real-world usages. The model inference process consists of two stages: encoding the prefix (model input) and autoregressively decoding new tokens (model output). Correspondingly, the inference efficiency considers two parts: time to encode the prefix and speed to decode each token. Here we report the prefix encoding time and per-token decoding speed under different HD settings. We test the speed with a Nvidia-A100 80G. With the results in Table 10, we have three observations: 1) Prefix encoding time increases linearly with the number of prefix tokens. 2) Decoding speed remains relatively constant, regardless of prefix length, thanks to optimizations on transformers from research communities and companies, including kv-cache and flash-attention. 3) When generating 2048 tokens, total inference time usage is nearly identical across HD9 to HD55, as encoding time is much smaller. Based on the above analysis, we believe the inference efficiency of our model is acceptable. Besides, we believe some targeted designs can further improve efficiency, while our paper focuses on enabling LVLM to understand high-resolution images with a general and effective solution, and we would leave the efficiency exploration in future work.

5 Conclusion

In this paper, we propose the InternLM-Xcomposer2-4KHD that exceeds the performance of previous open-source models on OCR-related tasks and also achieves competitive results on general-purpose LVLM benchmarks. Thanks to our dynamic resolution and automatic patch configuration, our model supports a maximum training resolution of up to 4K HD. We also integrate a global view patch to support the macro understanding and a learnable newline token to handle the various input image resolutions. Our model's performance continues to improve as the training resolution increases for HD-OCR tasks. Notably, we do not observe any performance saturation even for the 4KHD setting, and we have not explored the upper bound due to the computational burden increasing with higher-resolution inputs. In future work, we plan to explore efficient solutions for accurate LVLM training and inference, enabling our model to handle even higher resolutions while maintaining computational efficiency.

6 Acknowledgment

This project is funded in part by Shanghai Artificial lntelligence Laboratory, the National Key R&D Program of China (2022ZD0160201), the Centre for Perceptual and Interactive Intelligence (CPII) Ltd under the Innovation and Technology Commission (ITC)'s InnoHK. Dahua Lin is a PI of CPII under the InnoHK.

References

[1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957, 2019.

[2] 01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. Yi: Open foundation models by 01.ai, 2024.

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *International Conference on Computer Vision (ICCV)*, 2015.

[4] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv.org*, 2023.

[5] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv.org*, 2023.

[6] Baichuan. Baichuan 2: Open large-scale language models. *arXiv.org*, 2023.

[7] Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. Introducing our multimodal models, 2023.

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

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:1877–1901, 2020.

[10] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*, 2024.

[11] Yuhang Cao, Pan Zhang, Xiaoyi Dong, Dahua Lin, and Jiaqi Wang. DualFocus: Integrating macro and micro perspectives in multi-modal large language models. *arXiv preprint arXiv:2402.14767*, 2024.

[12] Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. Honeybee: Locality-enhanced projector for multimodal llm. *arXiv preprint arXiv:2312.06742*, 2023.

[13] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv.org*, 2023.

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

[15] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024.

[16] Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, Siamak Shakeri, Mostafa Dehghani, Daniel Salz, Mario Lucic, Michael Tschannen, Arsha Nagrani, Hexiang Hu, Mandar Joshi, Bo Pang, Ceslee Montgomery, Paulina Pietrzyk, Marvin Ritter, AJ Piergiovanni, Matthias Minderer, Filip Pavetic, Austin Waters, Gang Li, Ibrahim Alabdulmohsin, Lucas Beyer, Julien Amelot, Kenton Lee, Andreas Peter Steiner, Yang Li, Daniel Keysers, Anurag Arnab, Yuanzhong Xu, Keran Rong, Alexander Kolesnikov, Mojtaba Seyedhosseini, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. Pali-x: On scaling up a multilingual vision and language model, 2023.

[17] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server, 2015.

[18] Xi Chen, Xiao Wang, Lucas Beyer, Alexander Kolesnikov, Jialin Wu, Paul Voigtlaender, Basil Mustafa, Sebastian Goodman, Ibrahim Alabdulmohsin, Piotr Padlewski, Daniel Salz, Xi Xiong, Daniel Vlasic,

 https://doi.org/10.52202/079017-1348

Filip Pavetic, Keran Rong, Tianli Yu, Daniel Keysers, Xiaohua Zhai, and Radu Soricut. Pali-3 vision language models: Smaller, faster, stronger, 2023.

[19] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. Pali: A jointly-scaled multilingual language-image model, 2023.

[20] 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.

[21] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.

[22] Chee Kheng Chng, Yuliang Liu, Yipeng Sun, Chun Chet Ng, Canjie Luo, Zihan Ni, ChuanMing Fang, Shuaitao Zhang, Junyu Han, Errui Ding, et al. Icdar2019 robust reading challenge on arbitrary-shaped text-rrc-art. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1571–1576. IEEE, 2019.

[23] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv.org*, 2022.

[24] OpenCompass Contributors. Opencompass: A universal evaluation platform for foundation models. `https://github.com/open-compass/opencompass`, 2023.

[25] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

[26] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[27] 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.

[28] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model. In *arXiv preprint arXiv:2303.03378*, 2023.

[29] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, 2022.

[30] Hao Feng, Qi Liu, Hao Liu, Wengang Zhou, Houqiang Li, and Can Huang. DocPedia: Unleashing the power of large multimodal model in the frequency domain for versatile document understanding. *arXiv preprint arXiv:2311.11810*, 2023.

[31] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

[32] Chaoyou Fu, Renrui Zhang, Zihan Wang, Yubo Huang, Zhengye Zhang, Longtian Qiu, Gaoxiang Ye, Yunhang Shen, Mengdan Zhang, Peixian Chen, Sirui Zhao, Shaohui Lin, Deqiang Jiang, Di Yin, Peng Gao, Ke Li, Hongsheng Li, and Xing Sun. A challenger to gpt-4v? early explorations of gemini in visual expertise. *arXiv preprint arXiv:2312.12436*, 2023.

[33] Yuying Ge, Yixiao Ge, Ziyun Zeng, Xintao Wang, and Ying Shan. Planting a seed of vision in large language model. *arXiv preprint arXiv:2307.08041*, 2023.

[34] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination & visual illusion in large vision-language models, 2023.

[35] Conghui He, Zhenjiang Jin, Chaoxi Xu, Jiantao Qiu, Bin Wang, Wei Li, Hang Yan, Jiaqi Wang, and Da Lin. Wanjuan: A comprehensive multimodal dataset for advancing english and chinese large models. *ArXiv*, abs/2308.10755, 2023.

[36] Wenyi Hong, Weihan 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. *arXiv preprint arXiv:2312.08914*, 2023.

[37] 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.
[38] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
[39] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.
[40] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656, 2018.
[41] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.
[42] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer, 2016.
[43] Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, pages 4999–5007, 2017.
[44] Paul Lerner, Olivier Ferret, Camille Guinaudeau, Hervé Le Borgne, Romaric Besançon, José G Moreno, and Jesús Lovón Melgarejo. Viquae, a dataset for knowledge-based visual question answering about named entities. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3108–3120, 2022.
[45] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension, 2023.
[46] Bo Li, Peiyuan Zhang, Jingkang Yang, Yuanhan Zhang, Fanyi Pu, and Ziwei Liu. Otterhd: A high-resolution multi-modality model, 2023.
[47] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv.org*, 2023.
[48] 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.
[49] Zhuowan Li, Xingrui Wang, Elias Stengel-Eskin, Adam Kortylewski, Wufei Ma, Benjamin Van Durme, and Alan L Yuille. Super-clevr: A virtual benchmark to diagnose domain robustness in visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14963–14973, 2023.
[50] Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. *arXiv preprint arXiv:2311.06607*, 2023.
[51] Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023.
[52] Adam Dahlgren Lindström and Savitha Sam Abraham. Clevr-math: A dataset for compositional language, visual and mathematical reasoning. *arXiv preprint arXiv:2208.05358*, 2022.
[53] Fangyu Liu, Guy Edward Toh Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 2023.
[54] Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. Mmc: Advancing multimodal chart understanding with large-scale instruction tuning. *arXiv preprint arXiv:2311.10774*, 2023.
[55] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024.
[56] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv.org*, 2023.
[57] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhnag, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? *arXiv:2307.06281*, 2023.
[58] Yuliang Liu, Zhang Li, Biao Yang, Chunyuan Li, Xucheng Yin, Cheng lin Liu, Lianwen Jin, and Xiang Bai. On the hidden mystery of ocr in large multimodal models, 2024.
[59] Yuliang Liu, Biao Yang, Qiang Liu, Zhang Li, Zhiyin Ma, Shuo Zhang, and Xiang Bai. Textmonkey: An ocr-free large multimodal model for understanding document. *arXiv preprint arXiv:2403.04473*, 2024.
[60] Ziyu Liu, Zeyi Sun, Yuhang Zang, Wei Li, Pan Zhang, Xiaoyi Dong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. RAR: Retrieving and ranking augmented mllms for visual recognition. *arXiv preprint arXiv:2403.13805*, 2024.

 https://doi.org/10.52202/079017-1348

[61] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*, 2024.
[62] Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *The 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021.
[63] 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.
[64] Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *arXiv preprint arXiv:2209.14610*, 2022.
[65] Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv preprint arXiv:2110.13214*, 2021.
[66] Tengchao Lv, Yupan Huang, Jingye Chen, Lei Cui, Shuming Ma, Yaoyao Chang, Shaohan Huang, Wenhui Wang, Li Dong, Weiyao Luo, Shaoxiang Wu, Guoxin Wang, Cha Zhang, and Furu Wei. Kosmos-2.5: A multimodal literate model, 2023.
[67] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019.
[68] 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.
[69] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022.
[70] 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.
[71] Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, et al. Mm1: Methods, analysis & insights from multimodal llm pre-training. *arXiv preprint arXiv:2403.09611*, 2024.
[72] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*, 2019.
[73] OpenAI. Chatgpt. `https://openai.com/blog/chatgpt`, 2022.
[74] OpenAI. Gpt-4 technical report, 2023.
[75] Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. Im2text: Describing images using 1 million captioned photographs. In *Neural Information Processing Systems (NIPS)*, 2011.
[76] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:27730–27744, 2022.
[77] 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.org*, 2023.
[78] Qwen. Introducing qwen-7b: Open foundation and human-aligned models (of the state-of-the-arts), 2023.
[79] 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 *Proceedings of the International Conference on Machine learning (ICML)*, pages 8748–8763. PMLR, 2021.
[80] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.
[81] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pages 146–162. Springer, 2022.
[82] Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. Kvqa: Knowledge-aware visual question answering. In *Proceedings of the AAAI conference on artificial intelligence*, 2019.
[83] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.
[84] Baoguang Shi, Cong Yao, Minghui Liao, Mingkun Yang, Pei Xu, Linyan Cui, Serge Belongie, Shijian Lu, and Xiang Bai. Icdar2017 competition on reading chinese text in the wild (rctw-17). In *2017 14th iapr international conference on document analysis and recognition (ICDAR)*, volume 1, pages 1429–1434. IEEE, 2017.

[85] Chenglei Si, Yanzhe Zhang, Zhengyuan Yang, Ruibo Liu, and Diyi Yang. Design2code: How far are we from automating front-end engineering?, 2024.
[86] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 742–758. Springer, 2020.
[87] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019.
[88] Yipeng Sun, Zihan Ni, Chee-Kheng Chng, Yuliang Liu, Canjie Luo, Chun Chet Ng, Junyu Han, Errui Ding, Jingtuo Liu, Dimosthenis Karatzas, et al. Icdar 2019 competition on large-scale street view text with partial labeling-rrc-lsvt. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1557–1562. IEEE, 2019.
[89] Zeyi Sun, Ye Fang, Tong Wu, Pan Zhang, Yuhang Zang, Shu Kong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Alpha-CLIP: A clip model focusing on wherever you want. *arXiv preprint arXiv:2312.03818*, 2023.
[90] Gemini Team. Gemini: A family of highly capable multimodal models, 2023.
[91] InternLM Team. Internlm: A multilingual language model with progressively enhanced capabilities. `https://github.com/InternLM/InternLM`, 2023.
[92] 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.org*, 2023.
[93] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models, 2023.
[94] 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.
[95] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models, 2023.
[96] Yonghui Wang, Wengang Zhou, Hao Feng, Keyi Zhou, and Houqiang Li. Towards improving document understanding: An exploration on text-grounding via mllms. *arXiv preprint arXiv:2311.13194*, 2023.
[97] 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 models. *arXiv preprint arXiv:2312.06109*, 2023.
[98] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, et al. Q-bench: A benchmark for general-purpose foundation models on low-level vision. *arXiv preprint arXiv:2309.14181*, 2023.
[99] Ruyi Xu, Yuan Yao, Zonghao Guo, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, Maosong Sun, and Gao Huang. Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. *arXiv preprint arXiv:2403.11703*, 2024.
[100] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, et al. mPLUG-DocOwl: Modularized multimodal large language model for document understanding. *arXiv preprint arXiv:2307.02499*, 2023.
[101] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, et al. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. *arXiv preprint arXiv:2310.05126*, 2023.
[102] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv.org*, 2023.
[103] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.
[104] Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*, 2023.
[105] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.
[106] Tai-Ling Yuan, Zhe Zhu, Kun Xu, Cheng-Jun Li, Tai-Jiang Mu, and Shi-Min Hu. A large chinese text dataset in the wild. *Journal of Computer Science and Technology*, 34(3):509–521, 2019.
[107] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*, 2023.
[108] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan

 https://doi.org/10.52202/079017-1348

Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. GLM-130b: An open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

[109] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-CLIP: Unlocking the long-text capability of clip. *arXiv preprint arXiv:2403.15378*, 2024.

[110] Pan Zhang, Xiaoyi Dong Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Hang Yan, et al. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*, 2023.

[111] Rui Zhang, Yongsheng Zhou, Qianyi Jiang, Qi Song, Nan Li, Kai Zhou, Lei Wang, Dong Wang, Minghui Liao, Mingkun Yang, et al. Icdar 2019 robust reading challenge on reading chinese text on signboard. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 1577–1581. IEEE, 2019.

[112] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. LLaVAR: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023.

[113] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv.org*, 2023.

A More examples chat with our InternLM-XComposer2-4KHD

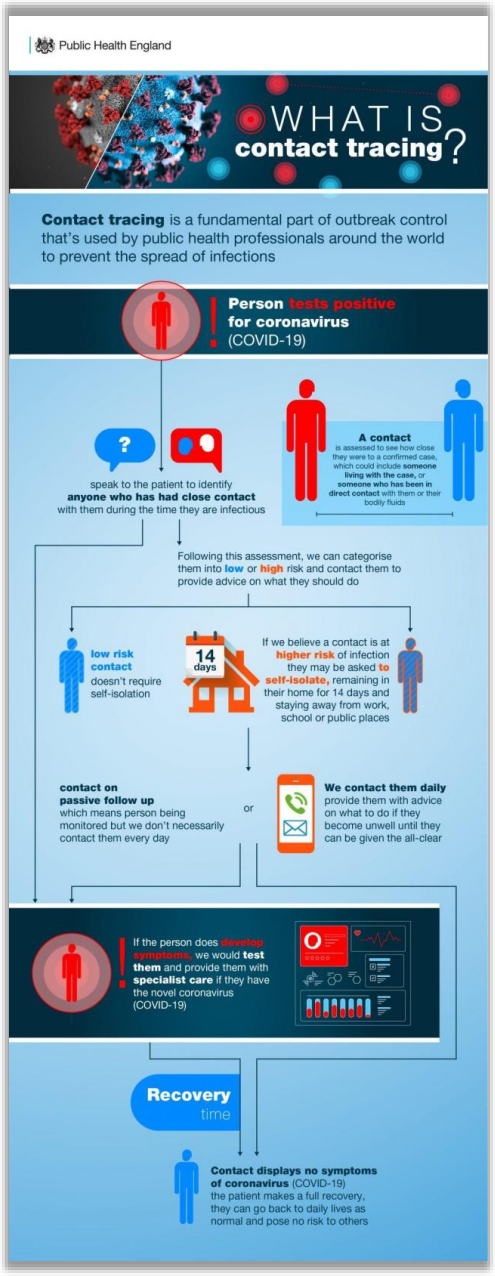


Image size: 986 x 2560

Give an in-depth description of the visual.

The image is an infographic from Public Health England that explains the concept of contact tracing in the context of preventing the spread of coronavirus (COVID-19). Contact tracing is a fundamental part of outbreak control used by public health professionals around the world to prevent the spread of infections.

The infographic is divided into several sections, each providing a different aspect of the contact tracing process. The first section explains that a contact is assessed to see how close they were to a confirmed case, which could include someone living with the case, or someone who has been in direct contact with them or their bodily fluids. Following this assessment, the contacts are categorised into low or high risk and provided with advice on what they should do.

The second section explains that if a contact is at high risk of infection, they may be asked to self-isolate, remaining in their home for 14 days and staying away from work, school or public places. If a contact is at low risk, they don't require self-isolation.

The third section explains that if the person develops symptoms, they would be tested and provided with specialist care if they have the novel coronavirus (COVID-19).

Finally, the infographic emphasizes the importance of recovery time, as contact displays no symptoms of coronavirus (COVID-19) the patient makes a full recovery, they can go back to daily lives as normal and pose no risk to others.

Figure 5: **Chat with InternLM-XComposer2-4KHD**.

https://doi.org/10.52202/079017-1348

The HD image (816 x 3813)

Detect all the word in the given image

American Heart Association. VennGAGE.

How to Develop Healthy Eating Habits. And Reduce the Risk of Heart Disease.

Your best defense is a healthy lifestyle filled with physical activity and nutritious meals, while cutting down or avoiding drinking and smoking.

The Current Trend

75% of the US population is low in its fruit and vegetables consumption. 1/10 adults meets the recommendation for fruit and vegetable consumption. Men and young adults in general eat the fewest fruits and vegetables compared to any other group.

3 Steps to Eating Healthy 1. Add produce of various colors to your diet to diversify your produce consumption. 2. Use hacks to incorporate more veggies and fruits to your meal, and throughout the day. 3. Take steps to improve access to fruits and vegetables for all.

1. Adding Color to Your Diet

Eating 2 to 2% cups (4 to 5 servings) daily of fruits and vegetables may reduce your risk for cardiovascular disease, including heart attack and stroke.

Examples of a single serving of fruit: 1 medium apple, 1 small banana, 1 medium avocado, 4 large strawberries. Examples of a single serving of vegetables: 5 - 8 broccoli florets, 2/3 large sweet potato, 6 baby carrots, 1 cup of raw spinach.

Adding a variety of produce to your diet has a number of health benefits. It helps you: Manage your weight, Control your blood pressure, Support healthy digestion, Reduce the risk of some cancers, such as colon cancer, Reduce the risk of chronic health problems, such as diabetes.

2. Use Healthy Eating Hacks

1. Fresh, frozen, canned and dried all count toward your daily servings, so you have plenty of options. 2. Always reserve half of your plate for fruits and vegetables. 3. Choose whole fruits and vegetables. 4. Try to eat a variety of vegetables, instead of the same thing all the time. 5. Look for fruit packed in its own fruit juice and 100% vegetable juice, with no added sugar. 6. Look for low/no-sodium options for canned vegetables, and 100% vegetable juice.

3. Take Action for Better Access

In many communities, fruits and vegetables are limited/expensive and only available in corner stores, convenience stores, bodegas and gas stations.

Lack of access can profoundly impact diet. Here's how: Contact your city and state leaders to let them know what food access is in your community.

Find ways to spread the word about nutrition assistance programs, such as SNAP, WIC and school meals.

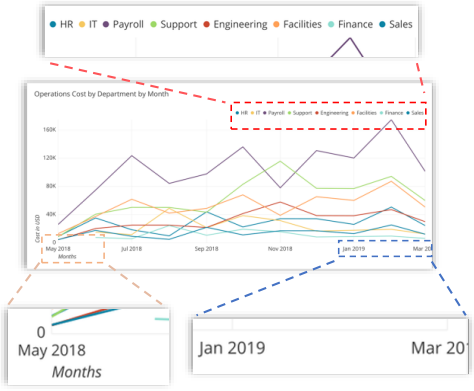
Create a petition for more variety, improve affordability and advocate for better signage/placement.

Meet with an after-school or daycare program representative to discuss serving more fruits and vegetables for snacks. Organize a letter-writing campaign and set up a meeting with state leaders.

For example, ask for funding to host a farmers market in an underserved community. Sign up for "You're the Cure" and send a note to your Congressperson advocating for healthier meals at school.

It's American Heart Month! Share this infographic with your family, friends and network today. You can carry out all of these actions or just a few. But whichever actions you choose, they pave the way for greater access to nutritious food for your community.

Figure 6: **Chat with InternLM-XComposer2-4KHD**.



The HD image (1475 x 675)

Conduct an in-depth analysis of the chart.

The line chart presents the operations cost by department in US dollars for various categories from May 2018 to Mar 2019. The categories include Human Resources (HR), Information Technology (IT), Payroll, Support, Engineering, Facilities, Finance, and Sales. Each category is represented by a different color and line on the graph. The y-axis represents the cost in thousands of US dollars, while the x-axis represents the months from May 2018 to Mar 2019. The graph shows fluctuations in the cost for each category throughout the months, with some categories experiencing peaks and troughs at different times.

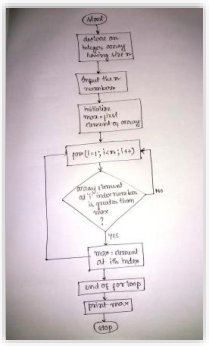
Which department has the highest cost in most cases?

Payroll has the highest cost in most cases, as indicated by the purple line on the graph.

The HD image (1045x1428)

analyze this flow chart step by step

This flow chart outlines the process of ordering a book from Amazon. Here's a step-by-step analysis:
1. Start: The process begins with the user initiating the order.
2. Enter the website from the browser: The user navigates to the Amazon website.
3. Registered?: The system checks if the user is already registered. If not, they are prompted to register.
4. Sign in: If the user is already registered, they are prompted to sign in.
5. On home page search for book of your choice: The user searches for the desired book on the home page.
6. Select your preferred book: The user selects their preferred book from the search results.
7. Buy new?: The system asks if the user wants to buy a new book. If the answer is "Yes", the process continues to step
8. If the answer is "No", the process continues to step 9.8. Select a new book: The user selects a new book.
9. Select an old book: The user selects an old book.
10. Add to cart: The user adds the selected book to their cart.
11. Review order: The user reviews their order.
12. Change information: The user has the option to change their information. If they choose not to, the process continues to step 13. If they choose to, the process returns to step 11.
13. Enter shipping information: The user enters their shipping information.
14. Select payment method: The user selects their preferred payment method.
15. Place order: The user places their order.16. End: The process concludes with the order being placed



The HD image (1888 x 3172)

What is the algorithm of this flow chart?

The flow chart in the image is an algorithm for finding the maximum element in an array. Here is a detailed explanation of the algorithm:

1. Declare an integer array having size m.
2. Input the m numbers into the array.
3. Initialize max as the first element of the array.
4. For each element in the array, starting from the second element:
   If the current element is greater than max, set max to the current element.
5. At the end of the loop, max will hold the maximum element in the array.
6. Output the maximum element.

This algorithm iterates through the array, comparing each element to the current maximum. If a larger element is found, it becomes the new maximum. At the end of the loop, the maximum element is stored in the variable max, and it is then outputted

Figure 7: **Chat with InternLM-XComposer2-4KHD**.

https://doi.org/10.52202/079017-1348

B Experiment Details

B.1 Data sampling strategy.

We maintain different dataloaders for image-text data and pure text data, sampling them in a weighted manner. The pure language dataloader is sampled with a fixed weight of 0.1 and the image-text dataloader is sampled with a weight of 0.9.

Within each dataloader, the training data comes from multiple sources, and we sample them in a weighted manner. In detail, for K datasets and the data number of the k_{th} dataset is n_k, its weight w_k is $min(100, n_k//1000)$ and the normalized weight is $\hat{w}_k = w_k / \sum w_k$. For each get_data operation within the dataloader, we use weighted sampling to choose a dataset and randomly choose a training sample from it.

C Broader Impacts

On the positive side, our model introduces a novel solution that significantly enhances the ability of large language models to comprehend high-resolution images. This innovative approach is expected to be highly beneficial for the research community. It paves the way for new explorations and discoveries in the field of MLLM and image processing. The potential applications of this model are vast.

Moreover, we have plans to make our model open-source. By sharing our work with the public, we aim to foster an environment of shared learning and progress. Users, researchers, and developers can utilize our model, adapt it to their needs, and even contribute to its further development. This open-source approach will accelerate the pace of innovation and bring about more rapid advancements in the field.

On the negative side, like any powerful tool, there is a potential for misuse of our model. There is a risk that individuals with malicious intent may exploit the capabilities of the model for unethical or harmful purposes. This is a challenge that we, as researchers, must acknowledge and address. However, it is also crucial for users and the wider community to use our model responsibly and ethically. We believe that through collective effort, we can mitigate these risks and ensure that the benefits of our model are realized while minimizing potential harm.

D Limitation

In this paper, we study the influence of training resolution in a wide range, from HD-9 to 4KHD (HD-25 + HD-55). Our results show that high-resolution OCR-related tasks rely on the training resolution heavily, and get significant performance gains with increased resolution. Till our largest setting 4KHD, its gain is not saturated. If we keep increasing the resolution, it may get better results. However, due to computational constraints, we failed to fully explore the potential improvements from further increasing training resolution, and we have to leave it as future work.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We write the paper carefully and confirm that the main claim in the abstract and introduction accurately reflects the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

https://doi.org/10.52202/079017-1348

Justification: We mentioned the limitation of not fully exploring the upper bound of improvement by increasing training resolution, due to the computational constraints, and we would leave it to future work.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: We do not have theoretical results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The training details and datasets are provided in the paper, we also plan to open-source our model.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We plan to open-source the code and data after the final decision of the paper.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

 https://doi.org/10.52202/079017-1348

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We have provided the details in Sec.3

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: It is computationally expensive to report the error bars.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have provided it in Sec 3.4

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have discussed it in the appendix.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

https://doi.org/10.52202/079017-1348

Justification: Our work focus on high-resolution image understanding and does not has such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All the assets are properly cited in the paper.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We provide a new model with the capability to understand high-resolution images, and we have provided the details of the model in the paper.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-1348