Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling

Zhe Chen[4,1*†], **Weiyun Wang**[5,1*†], **Yue Cao**[4,1*†], **Yangzhou Liu**[4,1*†], **Zhangwei Gao**[7,1*†], **Erfei Cui**[7,1*†], **Jinguo Zhu**[1*], **Shenglong Ye**[1*], **Hao Tian**[2*], **Zhaoyang Liu**[1*†], **Lixin Gu**[1], **Xuehui Wang**[1†], **Qingyun Li**[1†], **Yimin Ren**[1†], **Zixuan Chen**[2], **Jiapeng Luo**[2], **Jiahao Wang**[2], **Tan Jiang**[2], **Bo Wang**[2], **Conghui He**[1], **Botian Shi**[1], **Xingcheng Zhang**[1], **Han Lv**[1], **Yi Wang**[1], **Wenqi Shao**[1], **Pei Chu**[1], **Zhongying Tu**[1], **Tong He**[1], **Zhiyong Wu**[1], **Huipeng Deng**[1], **Jiaye Ge**[1], **Kai Chen**[1], **Min Dou**[1], **Lewei Lu**[2], **Xizhou Zhu**[3,1], **Tong Lu**[4], **Dahua Lin**[6,1], **Yu Qiao**[1], **Jifeng Dai**[3,1✉], **Wenhai Wang**[6,1✉]

[1]Shanghai AI Laboratory [2]SenseTime Research [3]Tsinghua University [4]Nanjing University
[5]Fudan University [6]The Chinese University of Hong Kong [7]Shanghai Jiao Tong University

Code: `https://github.com/OpenGVLab/InternVL`
Model: `https://huggingface.co/OpenGVLab/InternVL2_5-78B`
HF Demo: `https://huggingface.co/spaces/OpenGVLab/InternVL`

Abstract

We introduce InternVL 2.5, an advanced multimodal large language model (MLLM) series that builds upon InternVL 2.0, maintaining its core model architecture while introducing significant enhancements in training and testing strategies as well as data quality. In this work, we delve into the relationship between model scaling and performance, systematically exploring the performance trends in vision encoders, language models, dataset sizes, and test-time configurations. Through extensive evaluations on a wide range of benchmarks, including multi-discipline reasoning, document understanding, multi-image / video understanding, real-world comprehension, multimodal hallucination detection, visual grounding, multilingual capabilities, and pure language processing, InternVL 2.5 exhibits competitive performance, rivaling leading commercial models such as GPT-4o and Claude-3.5-Sonnet. Notably, our model is the first open-source MLLMs to surpass 70% on the MMMU benchmark, achieving a 3.7-point improvement through Chain-of-Thought (CoT) reasoning and showcasing strong potential for test-time scaling. HuggingFace demo see `https://huggingface.co/spaces/OpenGVLab/InternVL`

1 Introduction

In recent years, multimodal large language models (MLLMs) [60, 137, 246, 36, 35, 248, 140, 228, 192, 275, 143, 54, 170] have emerged as a pivotal technology in artificial intelligence, capable of processing and understanding information from multiple modalities such as text, images, and videos. These models promise breakthroughs across fields like natural language processing, computer vision, and human-computer interaction. However, developing large-scale MLLMs remains a challenging task, requiring significant computational resources, sophisticated architectures, and the ability to effectively integrate diverse data types in a scalable manner.

Various attempts have been made to address these challenges, including enhancing model architectures [220, 232, 5, 172, 157, 210], scaling vision encoders [252, 66, 36, 293, 185] and language models [231, 235, 64, 19, 229, 221, 62], incorporating more diverse and high-quality datasets [124, 234, 25, 155], and refining the test-time scaling process [215, 249, 230] to boost performance. Notable commercial models, like GPT-4o [192] and Claude-3.5-Sonnet [8], have demonstrated exceptional performance, their closed nature limits transparency

* equal contribution; † interns at OpenGVLab, Shanghai AI Laboratory;
✉ corresponding authors (daijifeng@tsinghua.edu.cn, wangwenhai@pjlab.org.cn).

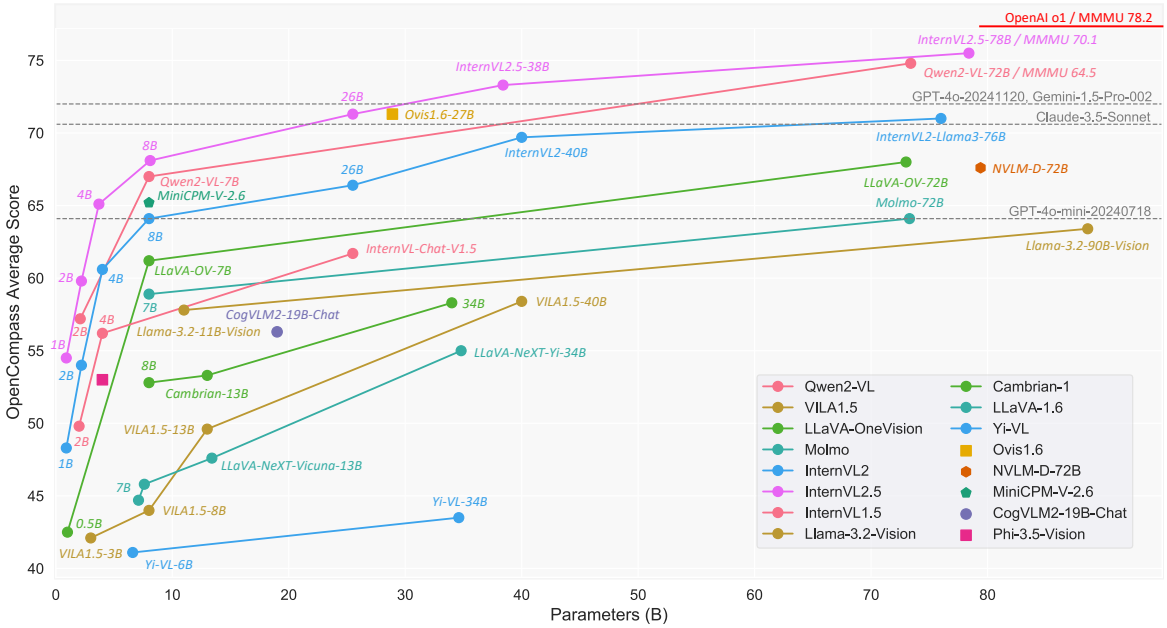


Figure 1: **Performance of various MLLMs on the OpenCompass leaderboard.** InternVL 2.5 showcases strong multimodal capabilities, rivaling closed-source models like GPT-4o [192] and Claude-3.5-Sonnet [8]. However, since the OpenCompass score is derived from 8 academic VQA benchmarks and covers only a subset of overall capabilities, we still need further effort to match the performance with closed-source models.

and accessibility, leaving gaps in the open-source community. While open-source multimodal models such as the InternVL series [36, 35, 71] and Qwen-VL series [13, 246] have provided high-performance, transparent alternatives, they still fall short in terms of achieving the desired levels of performance and efficiency.

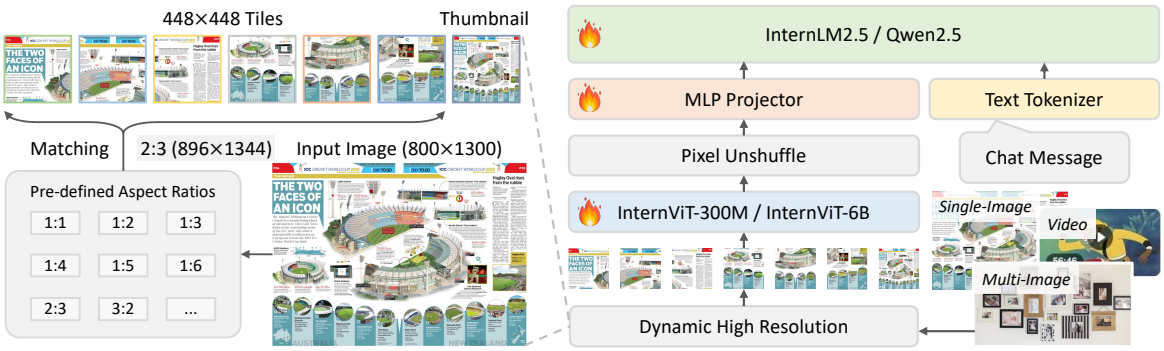
In this work, we introduce InternVL 2.5, an advanced large-scale MLLM series that builds upon the foundational architecture of InternVL 2.0. Continuing the objectives of the entire InternVL series, we aim to bridge the performance gap between commercial closed-source models and open-source multimodal models. In InternVL 2.5, we systematically explore various factors in MLLM, including how changes in vision encoders, language models, dataset sizes, and inference times affect the overall performance of the model, demonstrating the relationship between scaling and performance in multimodal models. Specifically, we have some interesting findings: (1) *Large vision encoders significantly reduce the dependency on training data when scaling up MLLMs.* As shown in Table 3, compared to Qwen2-VL-72B [246] equipped with a 600M vision encoder, our InternVL2.5-78B with a 6B vision encoder can achieve better performance using only 1/10 of the training tokens. This greatly reduces the exploration cost when scaling up MLLMs; (2) *Data quality matters.* Upgrading InternVL from 2.0 to 2.5 doubled the dataset size, but strict filtering greatly improved quality. For example, we carefully excluded the anomalous samples (*e.g.*, repetitive patterns), achieving substantial improvements in Chain-of-Thought (CoT) reasoning tasks such as MMMU [289] and complex challenges like the OlympiadBench [80]. Note that, most existing open-source MLLMs tend to underperform when using CoT [249]. (3) *Test-time scaling is beneficial for difficult multimodal QA.* For challenging tasks such as MMMU, the InternVL2.5-78B with CoT reaches 70.1%, which is 3.7 points higher than the direct response. Subsequently, we have successfully verified that CoT can be further combined with majority voting and bring additional improvements.

Our contributions can be summarized as follows:

(1) We release InternVL 2.5 to the open-source community, providing a powerful tool for the development and application of multimodal AI systems and encouraging further research in this domain.

(2) We investigate how scaling different components of the MLLMs such as vision encoders, language models, dataset sizes, and inference time affect performance.

(3) Through extensive evaluations on diverse benchmarks—including multi-discipline reasoning, document understanding, multi-image / video understanding, real-world comprehension, multimodal hallucination detection, visual grounding, multilingual capabilities, and pure language processing—InternVL 2.5 exhibits competitive performance, rivaling leading commercial models like GPT-4o [192] and Claude-3.5-Sonnet [8]. It is the



(a) Data Preprocessing (b) Model Architecture

Figure 2: **Overall architecture.** InternVL 2.5 retains the same model architecture as InternVL 1.5 [35] and InternVL 2.0, *i.e.* the widely-used "ViT-MLP-LLM" paradigm, which combines a pre-trained InternViT-300M or InternViT-6B with LLMs [19, 229] of various sizes via an MLP projector. Consistent with previous versions, we apply a pixel unshuffle operation to reduce the 1024 visual tokens produced by each 448×448 image tile to 256 tokens. Moreover, compared to InternVL 1.5, InternVL 2.0 and 2.5 introduced additional data types, incorporating multi-image and video data alongside the existing single-image and text-only data.

first open-source MLLM to surpass 70% on the MMMU validation set [289], setting a new benchmark and highlighting the potential of open-source solutions in advancing multimodal AI.

2 Model Architecture

2.1 Overall Architecture

As shown in Figure 2 and Table 2, InternVL 2.5 retains the same model architecture as its predecessors, InternVL 1.5 [35] and InternVL 2.0, following the "ViT-MLP-LLM" paradigm widely adopted in various MLLM studies [150, 151, 36, 316, 162, 246, 124, 256].

In this new version, our implementation of this architecture integrates a newly incrementally pre-trained InternViT-6B or InternViT-300M with various pre-trained LLMs of different sizes and types, including InternLM 2.5 [19] and Qwen 2.5 [229], using a randomly initialized 2-layer MLP projector. As in the previous version, to enhance scalability for high-resolution processing, we simply applied a pixel unshuffle operation, reducing the number of visual tokens to one-quarter of the original. Consequently, in our model, a 448×448 image tile is represented by 256 visual tokens.

In terms of input data preprocessing, we adopted a similar dynamic resolution strategy as InternVL 1.5, dividing images into tiles of 448×448 pixels based on the aspect ratio and resolution of the input images. The key difference, starting from InternVL 2.0, is that we additionally introduced support for multi-image and video data, as shown in Figure 2(b). Different data types correspond to different preprocessing configurations, which we will detail in Section 3.1.

2.2 Vision Encoder

InternVL employs InternViT [36] as the vision encoder. To better document the training progression of InternViT, we have provided detailed information in Table 1. InternViT currently has two different model sizes, including InternViT-6B and InternViT-300M.

InternViT-6B. InternViT-6B-224px was first introduced in our CVPR paper [36], and its structure follows the vanilla ViT [61], with minor adjustments incorporating QK-Norm [53] and RMSNorm [294]. It had 5.9B parameters, 48 layers, a hidden size of 3200, and 25 heads, and it was trained using a contrastive loss [195]. Due to the limited gains at that time, we adopted an incremental pre-training strategy to continuously refine its weights. Specifically, we connected InternViT-6B to an LLM via an MLP projector and, following a brief MLP warmup, jointly trained the InternViT-6B using a next token prediction loss (as shown in Figure 4(a)) to enhance its visual feature extraction capabilities. In the V1.0 and V1.2 versions, we used a fixed resolution of 448×448 for training, but in later versions, we switched to dynamic resolution training to improve high-resolution

Model Name	Train Res.	Width	Depth	MLP	#Heads	QK-Norm	Norm Type	Loss Type	#Param
InternViT-6B-224px	fixed 224	3200	48	12800	25	✓	RMS	CLIP	5.9B
InternViT-6B-448px-V1.0	fixed 448	3200	48	12800	25	✓	RMS	NTP	5.9B
InternViT-6B-448px-V1.2	fixed 448	3200	45	12800	25	✓	RMS	NTP	5.5B
InternViT-6B-448px-V1.5	dynamic 448	3200	45	12800	25	✓	RMS	NTP	5.5B
InternViT-6B-448px-V2.5	dynamic 448	3200	45	12800	25	✓	RMS	NTP	5.5B
InternViT-300M-448px-Distill	fixed 448	1024	24	4096	16	✗	LN	Cosine	0.3B
InternViT-300M-448px	dynamic 448	1024	24	4096	16	✗	LN	NTP	0.3B
InternViT-300M-448px-V2.5	dynamic 448	1024	24	4096	16	✗	LN	NTP	0.3B

Table 1: **Details of InternViT-6B and InternViT-300M models.** "fixed 224" refers to training images resized to 224×224, while "dynamic 448" means the model is trained with dynamic high resolution, with each image tile being 448×448. "CLIP" refers to the contrastive loss, "Cosine" represents the cosine distillation loss, while "NTP" indicates the next token prediction loss.

Model Name	#Param	Vision Encoder	Language Model	OpenCompass
InternVL-Chat-V1.5	25.5B	InternViT-6B-448px-V1.5	internlm2-chat-20b	61.7
InternVL2-1B	0.9B	InternViT-300M-448px	Qwen2-0.5B-Instruct	48.3
InternVL2-2B	2.2B	InternViT-300M-448px	internlm2-chat-1.8b	54.0
InternVL2-4B	4.2B	InternViT-300M-448px	Phi-3-mini-128k-instruct	60.6
InternVL2-8B	8.1B	InternViT-300M-448px	internlm2_5-7b-chat	64.1
InternVL2-26B	25.5B	InternViT-6B-448px-V1.5	internlm2-chat-20b	66.4
InternVL2-40B	40.1B	InternViT-6B-448px-V1.5	Nous-Hermes-2-Yi-34B	69.7
InternVL2-Llama3-76B	76.3B	InternViT-6B-448px-V1.5	Hermes-2-Theta-Llama-3-70B	71.0
InternVL2.5-1B	0.9B	InternViT-300M-448px-V2.5	Qwen2.5-0.5B-Instruct	54.5
InternVL2.5-2B	2.2B	InternViT-300M-448px-V2.5	internlm2_5-1_8b-chat	59.8
InternVL2.5-4B	3.7B	InternViT-300M-448px-V2.5	Qwen2.5-3B-Instruct	65.1
InternVL2.5-8B	8.1B	InternViT-300M-448px-V2.5	internlm2_5-7b-chat	68.1
InternVL2.5-26B	25.5B	InternViT-6B-448px-V2.5	internlm2_5-20b-chat	71.3
InternVL2.5-38B	38.4B	InternViT-6B-448px-V2.5	Qwen2.5-32B-Instruct	73.3
InternVL2.5-78B	78.4B	InternViT-6B-448px-V2.5	Qwen2.5-72B-Instruct	75.5
InternVL2.5-Pro	–	InternViT-6B-448px-V2.5	–	–

Table 2: **Pre-trained models used in the InternVL series.** In the InternVL 2.5 series, we upgraded both the vision encoder and the language model, resulting in improved performance. The OpenCompass scores for InternVL 1.5 and InternVL 2.0 were collected from the OpenCompass leaderboard, while the scores for InternVL 2.5 series were obtained through our local testing.

processing. As detailed in the InternVL 1.5 report [35], we removed the last three layers of InternViT-6B-448px-V1.2, reducing its depth from 48 to 45 layers, as these layers were more tuned to the CLIP loss objective, prioritizing global alignment over local information. As a result, all subsequent versions, including the latest InternViT-6B-448px-V2.5, have 45 layers and 5.5B parameters.

InternViT-300M. InternViT-300M-448px-Distill is a distilled variant of the teacher model, InternViT-6B-448px-V1.5, utilizing a cosine distillation loss. This model comprises 0.3B parameters, 24 layers, a hidden size of 1024, and 16 attention heads. Unlike the 6B version, the 0.3B variant employs standard LayerNorm [11] without QK-Norm [53]. To reduce distillation costs, we initialized this model using CLIP-ViT-Large-336px [195] where applicable, despite some architectural differences. After distillation, we integrated this model with an LLM and, following a similar procedure as described above, trained the vision encoder with dynamic high-resolution and the NTP loss. Then, we extracted the vision encoder and released it as InternViT-300M-448px. In this report, we further refined the InternViT-300M by incrementally pre-training the previous weights on a more diverse data mixture using the NTP loss, leading to the enhanced InternViT-300M-448px-V2.5.

2.3 Large Language Model

In Table 2, we provide an overview of the language models used across different versions of InternVL, including InternVL 1.5, InternVL 2.0, and the latest InternVL 2.5. As shown, earlier versions primarily built on language models such as InternLM 2 [19], Qwen 2 [268], Phi 3 [1], Yi [279], and Llama 3 [64]. To achieve better

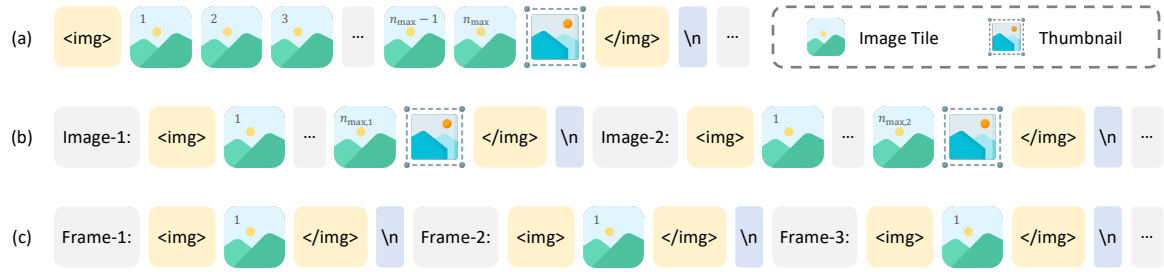


Figure 3: **Illustration of the data formats for various data types.** (a) For single-image datasets, the maximum number of tiles n_{max} is allocated to a single image, ensuring maximum resolution for the input. (b) For multi-image datasets, the total number of tiles n_{max} is distributed proportionally across all images within the sample. (c) For video datasets, the method simplifies the approach by setting $n_{\text{max}} = 1$, resizing individual frames to a fixed resolution of 448×448.

performance, in the InternVL 2.5 series, we have comprehensively upgraded the language backbones to the latest state-of-the-art models, including InternLM 2.5 [19] and Qwen 2.5 [229].

3 Training Strategy

3.1 Dynamic High-Resolution for Multimodal Data

In InternVL 2.0 and 2.5, we extend the dynamic high-resolution training approach introduced in InternVL 1.5 [35], enhancing its capabilities to handle multi-image and video datasets. The process mainly consists of the following steps:

Closest Aspect Ratio Matching. Given an input image I with dimensions $W \times H$, the aspect ratio is computed as $r = \frac{W}{H}$. The objective is to resize the image into tiles of size $S \times S$ (where $S = 448$) while selecting the closest aspect ratio that minimizes distortion. The number of tiles, n_{tiles}, is constrained within a predefined range $[n_{\min}, n_{\max}]$.

To find the optimal aspect ratio for resizing, we define the set of target aspect ratios $\mathcal{R}$ as:

$$\mathcal{R} = \{i/j \mid 1 \leq i, j \leq n,\ i \times j \in [n_{\min}, n_{\max}]\}\,. \tag{1}$$

The closest aspect ratio r_{best} is selected by minimizing the difference between the original aspect ratio r and each target aspect ratio r_{target}:

$$r_{\text{best}} = \arg\min_{r_{\text{target}} \in \mathcal{R}} |r - r_{\text{target}}|\,. \tag{2}$$

In cases where multiple aspect ratios produce the same difference (*e.g.*, 1:2 and 2:4), we prioritize the aspect ratio that results in an area less than or equal to twice the original image size. This helps to some extent in preventing the excessive enlargement of low-resolution images.

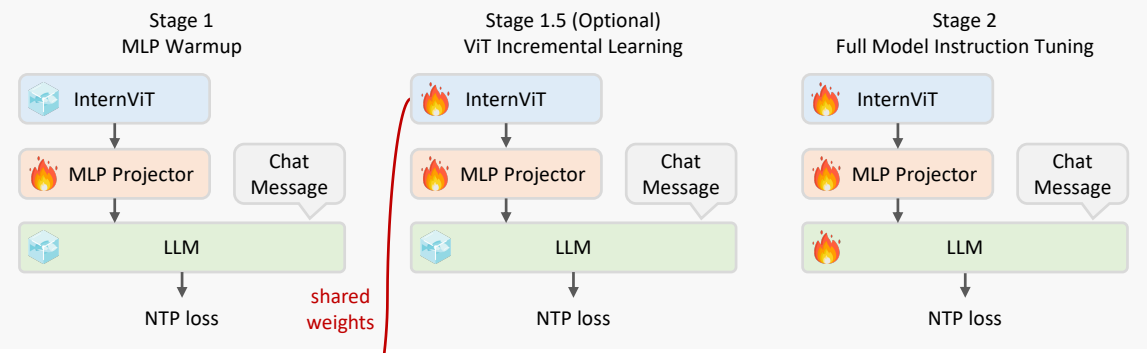
Image Resizing and Splitting. Once the best aspect ratio is determined, the image is resized to new dimensions $W_{\text{new}} \times H_{\text{new}}$, where i_{best} and j_{best} are the factors corresponding to r_{best}:

$$W_{\text{new}} = S \times i_{\text{best}}, \quad H_{\text{new}} = S \times j_{\text{best}}. \tag{3}$$

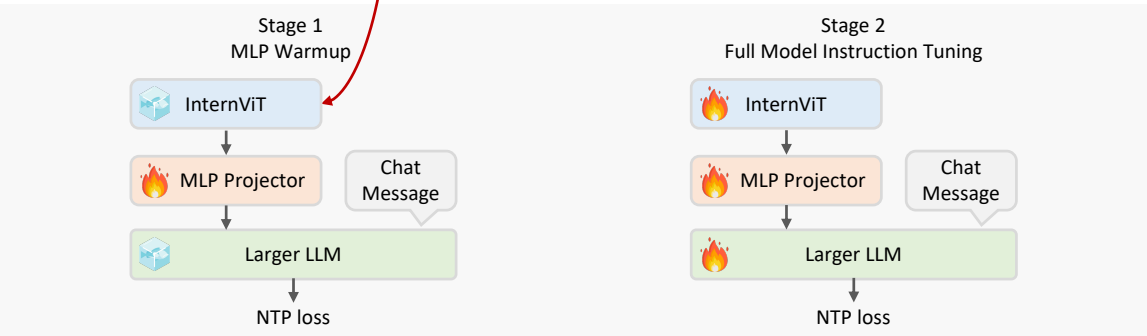
The image is then split into tiles of size $S \times S$, with the number of tiles calculated as $n_{\text{tiles}} = i_{\text{best}} \times j_{\text{best}}$. Each tile is cropped from the resized image to ensure consistent size.

Thumbnail Generation. Optionally, if the number of tiles $n_{\text{tiles}} > 1$, the original image I is resized to a square of dimensions $S \times S$ to generate an additional thumbnail I_{thumb}. This thumbnail is appended to the list of tiles, providing a global view alongside the localized tiles. In cases where $n_{\text{tiles}} = 1$, there is no thumbnail to append, and the mechanism naturally skips this step.

Data Formats for Different Data Types. As shown in Figure 3, the dynamic high-resolution method in InternVL 2.0 and 2.5 extends beyond single-image datasets to also support multi-image and video datasets.



(a) Single Model Training Pipeline



(b) Progressive Scaling Strategy

Figure 4: **Illustration of the training pipeline and progressive scaling strategy.** (a) Single model training pipeline. The training process is divided into three stages—Stage 1 (MLP warmup), optional Stage 1.5 (ViT incremental learning), and Stage 2 (full model instruction tuning). The multi-stage design progressively enhances vision-language alignment, stabilizes training, and prepares modules for integration with larger LLMs. (b) Progressive scaling strategy. The ViT module trained with a smaller LLM in earlier stages can be easily integrated with larger LLMs, enabling scalable model alignment with affordable resource overhead.

For single-image datasets, the maximum number of tiles n_{max} is allocated to a single image, ensuring that it is processed at the highest possible resolution. In this scenario, visual tokens are enclosed within `<img>` and `</img>` tags, with no additional auxiliary tags used.

In the case of multi-image datasets, the total number of tiles n_{max} is distributed across all images within one sample. Each image is identified by an auxiliary tag like `Image-1` to clearly label individual images. The images themselves are enclosed within `<img>` and `</img>` tags, denoting the start and end of the image data. The number of tiles assigned to each image I_i is proportional to the total number of images N_{image}, following the equation:

$$n_{\text{max, i}} = \max\left(1, \left\lfloor \frac{n_{\text{max}}}{N_{\text{image}}} \right\rfloor\right). \tag{4}$$

For video data, this approach is simplified by setting $n_{\text{max}} = 1$. Each video frame is resized to a fixed resolution of 448×448, eliminating the need for tiling. This is because, during training, a large number of frames (*e.g.*, 32 or 64) are typically extracted from a single video. For our model, even without high-resolution input, this still results in 8,192 or 16,384 visual tokens. Each video frame, labeled with tags like `Frame-1`, is enclosed within the `<img>` and `</img>` tags, similar to image data.

3.2 Single Model Training Pipeline

The training pipeline for a single model in InternVL 2.5 is structured across three stages, designed to enhance the model's visual perception and multimodal capabilities. Each stage progressively integrates vision and language modalities, balancing performance optimization with training efficiency.

Settings	InternVL2.5-1B		InternVL2.5-2B		InternVL2.5-4B		InternVL2.5-8B		
	Stage 1	Stage 2	Stage 1	Stage 2	Stage 1	Stage 2	Stage 1	Stage 1.5	Stage 2
Dataset	Pre-train Mixture	Fine-tune Mixture	Pre-train Mixture	Fine-tune Mixture	Pre-train Mixture	Fine-tune Mixture	Pre-train Mixture	Pre-train Mixture	Fine-tune Mixture
Trainable	MLP	Full Model	MLP	Full Model	MLP	Full Model	MLP	ViT+MLP	Full Model
Packed Batch Size	512	512	512	512	512	512	512	1024	512
Learning Rate	2e-4	4e-5	2e-5	4e-5	2e-5	4e-5	2e-4	1e-5	4e-5
Context Length	16384	16384	16384	16384	16384	16384	16384	16384	16384
Image Tile Threshold	48	48	48	48	48	48	48	48	48
ViT Drop Path	0.0	0.1	0.0	0.1	0.0	0.1	0.0	0.1	0.1
Weight Decay	0.01	0.01	0.01	0.01	0.01	0.01	0.05	0.05	0.05
Training Epochs	–	4	–	4	–	2	–	–	1
Training Tokens	∼191B	∼176B	∼277B	∼176B	∼164B	∼88B	∼22B	∼76B	∼44B

Settings	InternVL2.5-26B			InternVL2.5-38B		InternVL2.5-78B	
	Stage 1	Stage 1.5	Stage 2	Stage 1	Stage 2	Stage 1	Stage 2
Dataset	Pre-train Mixture	Pre-train Mixture	Fine-tune Mixture	Pre-train Mixture	Fine-tune Mixture	Pre-train Mixture	Fine-tune Mixture
Trainable	MLP	ViT+MLP	Full Model	MLP	Full Model	MLP	Full Model
Packed Batch Size	512	1024	512	512	512	512	512
Learning Rate	2e-4	1e-5	2e-5	2e-4	2e-5	2e-4	2e-5
Context Length	16384	16384	16384	16384	16384	16384	16384
Image Tile Threshold	48	48	48	48	48	48	48
ViT Drop Path	0.0	0.4	0.4	0.0	0.4	0.0	0.4
Weight Decay	0.05	0.05	0.05	0.05	0.05	0.05	0.05
Training Epochs	–	–	1	–	1	–	1
Training Tokens	∼31B	∼146B	∼44B	∼107B	∼44B	∼76B	∼44B

Table 3: **Training configurations and hyperparameters for InternVL 2.5.** This table presents the training setups for various scales of InternVL 2.5 models. The configurations are carefully optimized to ensure efficient scaling and performance across different parameter sizes and training stages. Notably, Qwen2-VL [246] processed a cumulative total of 1.4T tokens, while our InternVL2.5-78B is trained on just ∼120B tokens.

Stage 1: MLP Warmup. As shown in Figure 4(a), the training begins with warming up the MLP projector, which is the initial bridge between visual and language representations. In this stage, only the MLP projector is trained while both the vision encoder (*i.e.*, InternViT [36]) and language model are frozen. To achieve optimal performance, we begin using the dynamic high-resolution training strategy from this stage, even though it increases the training cost.

In this phase, we utilize the pre-training data mixture as outlined in Table 4. The data is formatted in a structured ChatML style and optimized with the NTP loss. Additionally, a higher learning rate is applied to accelerate convergence, allowing the MLP to quickly adapt to the LLM's input space and establish robust cross-modal alignment. The MLP warmup phase ensures the model is well-prepared to handle multimodal tasks before unlocking additional trainable components in later stages, thereby improving training stability.

Stage 1.5: ViT Incremental Learning (Optional). As shown in Figure 4(a), Stage 1.5 introduces incremental learning for the vision encoder. During this stage, both the vision encoder and MLP projector are trainable, and training is conducted using the same pre-training data mixture and NTP loss as in Stage 1. The aim of this stage is to enhance the vision encoder's ability to extract visual features, allowing it to capture more comprehensive information, especially for domains that are relatively rare in web-scale datasets (*e.g.*, LAION-5B [203]), such as multilingual OCR data and mathematical charts, among others.

As shown in Table 3, a lower learning rate is used in this stage to prevent catastrophic forgetting, ensuring the encoder doesn't lose previously learned capabilities. Additionally, the vision encoder only needs to be trained once unless new domain requirements or data are introduced. Once trained, it can be reused with different LLMs without retraining (see Figure 4(b) and Section 3.3), making Stage 1.5 optional. This is particularly beneficial when the encoder has already been optimized for some specific tasks, allowing it to integrate with LLMs of various sizes without significant additional costs.

Stage 2: Full Model Instruction Tuning. In the final stage, as illustrated in Figure 4(a), the entire model—comprising the ViT, MLP, and LLM—is trained on high-quality multimodal instruction datasets. Data quality is especially important here, as the LLM, responsible for generating the final user-facing output,

is now trainable. Even a small amount of noisy data (*e.g.*, a few thousand samples) can lead to abnormal model behavior, like repetitive output or specific erroneous results. To mitigate the degradation of the LLM, we implement strict data quality controls during this stage.

Additionally, the training hyperparameters in this stage are kept simple, with a unified learning rate applied to the entire model rather than different learning rates for various components. After completing this stage, InternVL 2.5's full training process is finished. Although further improvements could be made through Stage 3—post-training with higher-quality data or other training methods (*e.g.*, preference optimization)—we plan to leave this for the future.

3.3 Progressive Scaling Strategy

As shown in Figure 4, we propose a progressive scaling strategy to efficiently align the vision encoder (*e.g.*, InternViT) with LLMs. While we previously adopted a similar strategy in the training of InternVL 1.5 and 2.0, this is the first time the approach has been formalized into a clear methodology. This strategy adopts a staged training approach, starting with smaller, resource-efficient LLMs and progressively scaling up to larger LLMs. This approach stems from our observation that *even when the ViT and LLM are jointly trained using NTP loss, the resulting visual features are generalizable representations that can be easily understood by other LLMs*.

Specifically, in Stage 1.5, the InternViT is trained alongside a smaller LLM (*e.g.*, 20B), focusing on optimizing fundamental visual capabilities and cross-modal alignment. This phase avoids the high computational costs associated with training directly with a large LLM. Using a shared-weight mechanism, the trained InternViT can be easily transferred to a larger LLM (*e.g.*, 72B) without requiring retraining. Consequently, when training a larger model, Stage 1.5 can be skipped (see Table 3), as the optimized InternViT module from earlier stages is reused. This not only accelerates training but also ensures that the vision encoder's learned representations are preserved and effectively integrated into the larger model.

By employing this progressive scaling strategy, we achieve scalable model updates at a fraction of the cost typically associated with large-scale MLLM training. For example, Qwen2-VL [246] processes a cumulative total of 1.4 trillion tokens, whereas our InternVL2.5-78B is trained on only about 120 billion tokens—*less than one-tenth of Qwen2-VL*. This approach proves particularly advantageous in resource-constrained settings by maximizing the reuse of pre-trained components, minimizing redundant computations, and enabling the efficient training of models capable of addressing complex vision-language tasks.

3.4 Training Enhancements

To enhance the model's adaptability to real-world scenarios and overall performance, two key techniques are introduced. These optimizations are essential in improving the user experience and the model's benchmark performance.

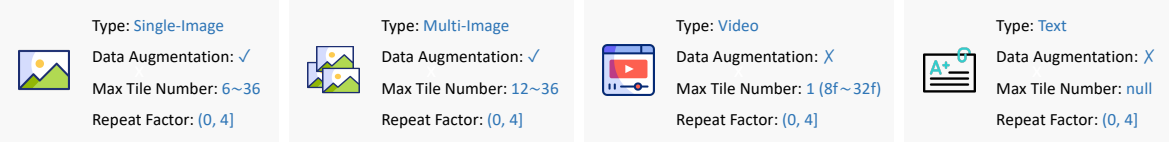
Random JPEG Compression. To avoid overfitting during training and enhance the model's real-world performance, we apply a data augmentation technique that preserves spatial information: JPEG compression. Specifically, random JPEG compression with quality levels between 75 and 100 is applied to simulate the degradation commonly found in internet-sourced images. This augmentation improves the model's robustness to noisy, compressed images and enhances the user experience by ensuring more consistent performance across varied image qualities.

Loss Reweighting. Token averaging and sample averaging are two widely applied strategies for weighting the NTP loss. Token averaging computes the average NTP loss across all tokens, whereas sample averaging first averages the NTP loss within each sample (across tokens) and then averages across the number of samples. These strategies can be expressed in a unified format:

$$\mathcal{L} = \frac{w_i}{\sum_j w_j} \cdot \mathcal{L}_i, \quad w_i = \begin{cases} \frac{1}{x^0}, & \text{for token averaging} \\ \frac{1}{x^1}, & \text{for sample averaging,} \end{cases} \tag{5}$$

where $\mathcal{L}_i$ and w_i denote the loss and weight of token i, respectively, and x denotes the number of tokens in the response to which token i belongs.

When using token averaging, each token contributes equally to the final loss, which can result in gradients biased toward responses with more tokens, leading to a drop in benchmark performance. In contrast, sample averaging ensures that each sample contributes equally, but it can cause the model to favor shorter responses, negatively impacting the user experience. To mitigate bias toward either longer or shorter responses during



(a) Single-Image Dataset (b) Multi-Image Dataset (c) Video Dataset (d) Text Dataset

Figure 5: **Dataset configuration.** In InternVL 2.0 and 2.5, data augmentation is applied selectively, enabled for image datasets and disabled for videos and text. The maximum tile number ($n_{\max}$) controls the resolution of inputs, with higher values for multi-image datasets and lower values for videos. The repeat factor (r) balances dataset sampling by adjusting the frequency of each dataset, ensuring robust and balanced training.

training, we apply a reweighting strategy where $w_i = \frac{1}{x^{0.5}}$. This approach, named **square averaging**, balances the contribution of responses with different lengths.

4 Data Organization

4.1 Dataset Configuration

In InternVL 2.0 and 2.5, the organization of the training data is controlled by several key parameters to optimize the balance and distribution of datasets during training, as shown in Figure 5.

Data Augmentation. Firstly, data augmentation (*i.e.*, JPEG compression introduced in Section 3.4) is applied conditionally, allowing for enhanced robustness by enabling or disabling augmentation techniques based on dataset characteristics. Specifically, we enable this augmentation for all image datasets, while disabling it for all video datasets, to ensure that different video frames have the same image quality.

Maximum Tile Number. The parameter $n_{\max}$ defines the maximum number of tiles allowed per dataset, effectively controlling the resolution of the image or video frame fed into the model. This ensures flexibility in handling datasets of varying complexity and type. For example, we can set $n_{\max} = 24$ or 36 for multi-image datasets, high-resolution documents, or infographics, use $n_{\max} = 6$ or 12 for most other low-resolution image datasets, and set $n_{\max} = 1$ for video datasets. This adjustment was first introduced in InternVL 2.0, whereas in InternVL 1.5, a uniform value of $n_{\max} = 12$ was applied across all datasets.

Repeat Factor. Finally, the repeat factor r determines the sampling frequency of each dataset. With $r \in (0, 4]$, this parameter enables down-sampling when $r < 1$, reducing the dataset's weight during training, or up-sampling when $r > 1$, effectively increasing the number of epochs for that dataset. This mechanism finely adjusts the relative proportions of datasets, ensuring a balanced distribution across training data. By adjusting r, especially in multi-task learning, the data from each domain or task receives appropriate training, preventing overfitting or underfitting of any single dataset, leading to more balanced model performance.

4.2 Multimodal Data Packing

In InternVL 2.0 and 2.5, we implement a data-packing strategy to enhance GPU utilization and improve training efficiency. This approach reduces padding by concatenating multiple samples into longer sequences, thereby maximizing the utilization of the model's input sequence capacity. Specifically, for multimodal models like InternVL, data packing should account for two dimensions: (a) *Sequence length for the LLM*, which corresponds to the standard input sequence length used in language models. This remains essential in multimodal tasks; (b) *Image tile number for the ViT*, which denotes the number of image tiles processed by the vision encoder. Efficient management of this dimension is crucial for optimizing training efficiency.

To handle these dimensions efficiently, our data-packing strategy comprises the following steps:

(1) **Select**: During the selection phase, the algorithm operates similarly to a standard dataset without data-packing, directly sampling independent data. Each sampled item is truncated into multiple smaller items and treated as separate samples. This ensures that the sequence length and image tile count of each sample are within the predefined thresholds $l_{\max}$ (context length) and $t_{\max}$ (image tile limit), respectively.

(2) **Search**: For a given independent sample, the algorithm searches for another sample from the buffer list to pack them together. The resulting packed sample must have a sequence length shorter than l_{max} and include fewer than t_{max} image tiles. If multiple buffers satisfy these requirements, the one with the longest sequence

Figure 6: **Visualization of abnormal samples in open-source datasets.** Abnormal samples are prevalent across various data types, including single-image, multi-image, video, and pure text datasets, with "repetitive patterns" being a prominent issue. We identify this as one of the most detrimental problems for test-time scaling, often leading models into loops in long-form outputs and CoT reasoning tasks. Thoroughly filtering the fine-tuning data mixture can mitigate this issue to some extent.

length and the maximum number of image tiles is selected. In practice, the buffer list is maintained in descending order and a binary search is performed to accelerate the search process.

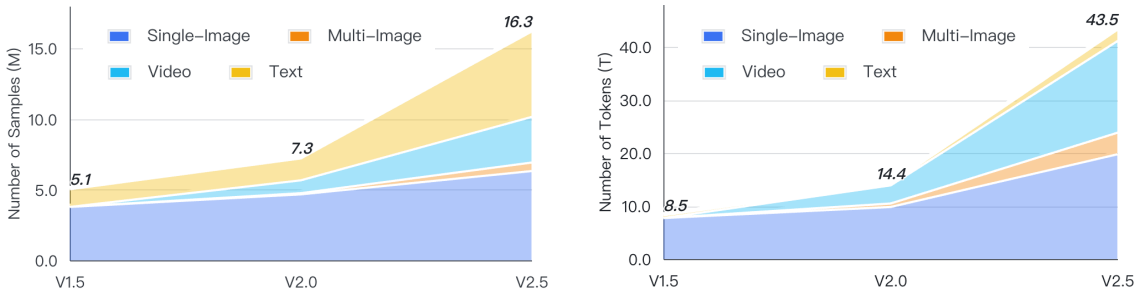
(3) **Pack**: The sampled data and the selected buffer are packed into a single sequence. If no buffer is selected in the previous step, the sample remains unchanged and proceeds directly to the next phase. Notably, tokens in the packed data can only attend to the context within their respective samples and cannot attend to tokens from other packed samples. Furthermore, the positional index of each sample is maintained independently.

(4) **Maintain**: In the maintenance phase, if a packed sample exceeds $l_{\max}$ or contains more than $t_{\max}$ image tiles, it is immediately yielded for training. Otherwise, the packed sample is inserted into the buffer list. If the buffer list exceeds its capacity, the sample with the longest sequence length and the highest number of image tiles is yielded to maintain buffer efficiency.

4.3 Data Filtering Pipeline

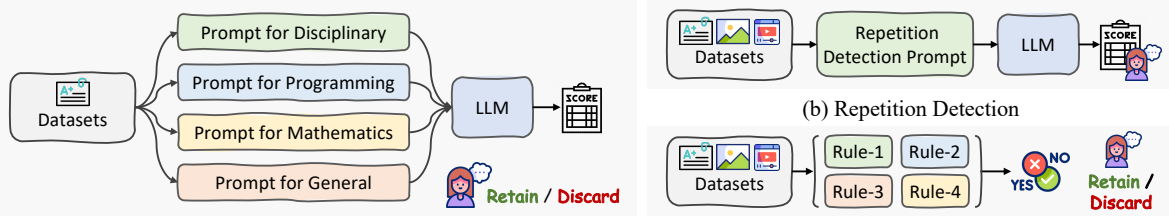
During model development, we observed that LLMs are significantly more sensitive to data noise than vision encoders. As shown in Figure 4, during Stage 2, when all model weights are fully trainable, even a small fraction of anomalous samples—such as outliers or repetitive data, numbering only a few thousand—can lead to aberrant model behavior during inference. While conventional wisdom assumes that minor noise in large-scale datasets can be ignored, our findings indicate otherwise: even a tiny fraction of noisy samples can degrade MLLM performance and user experience.

Among these anomalies, we identify *repetitive generation as one of the most detrimental issues*. In many open-source or synthetic datasets, a small number of samples with repetitive patterns—comprising merely thousands of examples in our fine-tuning data mixture—can cause the model to spiral into repetitive loops, particularly in long-form outputs or CoT reasoning tasks. This phenomenon undermines the effectiveness of test-time scaling strategies. To address this challenge and support future research, we designed an efficient data filtering pipeline to remove low-quality samples, thereby minimizing the risk of repetitive generation.



(a) Number of Samples (b) Number of Tokens

Figure 7: **Statistics of the fine-tuning data mixture.** The dataset shows consistent growth from InternVL 1.5 to 2.5 in terms of (a) the number of samples and (b) the number of tokens across multiple dataset types, including single-image, multi-image, video, and text. Note that the token count here refers to the total number of tokens in a specific modality dataset. For example, in the case of single-image datasets, the token count is the sum of the visual tokens and text tokens in these datasets. These statistics reflect iterative improvements in data scale and diversity, which enhance the model's multimodal understanding capabilities.



(a) LLM-Based Quality Scoring (b) Repetition Detection (c) Heuristic Rule-Based Filtering

Figure 8: **Dataset filtering pipeline.** For text data, we use three methods: (a) LLM-based quality scoring to assign domain-specific quality scores and filter low-quality samples; (b) Repetition detection to identify and remove data with repetitive patterns; and (c) heuristic rule-based filtering to detect anomalies using predefined rules. For multimodal data, only (b) repetition detection and (c) heuristic rule-based filtering are applied to mitigate repetitive patterns and ensure dataset integrity.

As shown in Figure 8, our data filtering pipeline consists of two modules. For pure-text data, we implemented three key strategies: (1) *LLM-Based Quality Scoring*: We begin by categorizing datasets into distinct domains (*e.g.*, disciplinary, programming, mathematics, general). Next, we assign a quality score, ranging from 0 to 10, to each sample using a pre-trained LLM [229] with a domain-specific prompt. Samples with scores below a specified threshold (*e.g.*, 7) are then removed to ensure data quality. (2) *Repetition Detection*: We use an LLM combined with a specialized prompt to identify repetitive patterns. These samples are then subjected to manual review, and those scoring below a threshold (*e.g.*, 3) are removed to maintain data quality. (3) *Heuristic Rule-Based Filtering*: We apply specific rules, such as filtering out sentences with abnormal lengths, excessively long sequences of zeros, text with an excessive number of duplicate lines, *etc*, to identify anomalies in the data. Although this approach may occasionally produce false positives, it improves the detection of anomalous samples. All flagged samples are manually reviewed before final removal.

For multimodal data, given the limitations of open-source MLLMs in scoring such data, we focused on mitigating repetitive patterns through two strategies: (1) *Repetition Detection*: We exempted high-quality academic datasets and used a specific prompt to identify repetitive patterns in the remaining data. These samples were removed following the same manual review process we applied to textual data. (2) *Heuristic Rule-Based Filtering*: Similar heuristic rules are applied, followed by manual verification to ensure dataset integrity.

This rigorous data-filtering pipeline significantly reduced the occurrence of anomalous behaviors, particularly repetitive generation, with notable improvements in CoT reasoning tasks. However, we recognize that data filtering alone cannot completely eliminate such issues. This may be due to the inherent noise introduced during the LLM's pre-training process, which our multimodal post-training efforts can only partially mitigate without fundamentally resolving the issue of repetitive outputs. Future work will explore preference optimization and other strategies to further suppress anomalies and enhance both model performance and user experience.

Task	Dataset
Type: Single/Multi-Image Datasets	
Captioning	FaceCaption [49], COCO-Caption [214], OpenImages-Caption [116], Objects365-Caption [208], TextCap [211], Laion-ZH [203], Laion-EN [203], Laion-COCO [204], LLaVAR [305], InternVL-SA-1B-Caption [113], MMInstruct [155], GRIT-Caption [194], ShareGPT4V [29], LVIS-Instruct-4V [244], ShareCaptioner [29], OmniCorpus [133], ShareGPT4o [35]
General QA	GQA [98], OKVQA [178], A-OKVQA [205], Visual7W [317], VisText [226], VSR [147], TallyQA [2], Objects365-YorN [208], IconQA [167], Stanford40 [273], VisDial [51], VQAv2 [74], Hateful-Memes [111]
Mathematics	MAVIS [300], GeomVerse [107], MetaMath-Rendered [281], MapQA [23], GeoQA+ [20], Geometry3K [164], UniGeo [26], GEOS [206], CLEVR-Math [144]
Chart	ChartQA [181], PlotQA [187], FigureQA [105], LRV-Instruction [148], ArxivQA [132], MMC-Inst [149], TabMWP [166], DVQA [104], UniChart [182], SimChart9K [263], Chart2Text [191], FinTabNet [312], SciTSR [39], Synthetic Chart2Markdown
OCR	LaionCOCO-OCR [204], Wukong-OCR [75], ParsynthOCR [89], SynthDoG-EN [112], SynthDoG-ZH [112], SynthDoG-RU [112], SynthDoG-JP [112], SynthDoG-KO [112], IAM [180], EST-VQA [253], ST-VQA [17], NAF [52], InfoVQA [183], HME100K [288], OCRVQA [188], SROIE [97], POIE [115], CTW [287], SynthText [79], ArT [40], LSVT [222], RCTW-17 [209], ReCTs [301], MTWI [82], TextVQA [212], CASIA [146], TextOCR [213], Chinese-OCR [14], EATEN [78], COCO-Text [238], Synthetic Arxiv OCR, Synthetic Image2Latex, Synthetic Handwritten OCR, Synthetic Infographic2Markdown
Knowledge	KVQA [207], A-OKVQA [205], ViQuAE [123], iNaturalist2018 [237], MovieNet [95], ART500K [176], KonIQ-10K [91], IconQA [167], VisualMRC [225], ChemVLM Data [129], ScienceQA [165], AI2D [109], TQA [110], Wikipedia-QA [81], Synthetic Multidisciplinary Knowledge / QA
Grounding	Objects365 [208], GRIT [278], RefCOCO [280], GPT4Gen-RD-BoxCoT [27], All-Seeing-V1 [251], All-Seeing-V2 [250], V3Det [243], TolokaVQA [236]
Document	DocReason25K [93], DocVQA [184], Docmatix [121], Synthetic Arxiv QA
Conversation	ALLaVA [25], SVIT [309], Cambrain-GPT4o [234], TextOCR-GPT4V [102], MMDU [159], Synthetic Real-World Conversations
Medical	PMC-VQA [303], VQA-RAD [120], ImageCLEF [72], SLAKE [145], Medical-Diff-VQA [94], PMC-CaseReport [260], GMAI-VL (subset) [134]
GUI	Screen2Words [240], WebSight [122]
Type: Video Datasets	
Captioning	Mementos [254], ShareGPT4Video [30], VideoGPT+ [174], ShareGPT4o-Video [35]
General QA	VideoChat2-IT [131], EgoTaskQA [99], NTU RGB+D [152], CLEVRER [276], STAR [259], LSMDC [201]

Table 4: **Summary of the pre-training data mixture of InternVL 2.5.** Notably, we exclusively use conversaiton-format instruction data, and at this stage, only the MLP or both MLP and ViT parameters are trainable, allowing the incorporation of both low-quality and high-quality data.

4.4 Pre-training Data Mixture

To comprehensively enhance the model's performance and strengthen its ability to handle complex tasks in real-world scenarios, we collect a broader range of domain-specific data compared to the training corpus of InternVL 1.5 and 2.0. As shown in Table 4, our training corpus is sourced from captioning, general QA, mathematics, charts, OCR, knowledge, grounding, documents, conversation, medical, and GUI tasks.

Notably, during the development of our models, we utilized conversation-format instruction data. For non-conversational datasets, such as image captioning, OCR, and object detection datasets, we construct questions to transform the data into a conversational format. At this stage, since only the parameters of MLP (*i.e.*, Stage 1) or MLP and ViT (*i.e.*, Stage 1.5) are trainable, both low-quality and high-quality data are incorporated. The goal is to enrich the model's world knowledge as much as possible by exposing it to diverse domain data, thereby improving its generalization capabilities.

In our view, the ideal scenario is for the fine-tuning data mixture to be a subset of the pre-training data mixture. This ensures that the data in this subset can be adequately trained within the vision encoder. However, in practice, due to the high training costs of Stage 1.5, achieving this is often difficult. Therefore, in the training of InternVL 2.5, only a subset of the datasets from the fine-tuning data mixture was included in the pre-training data mixture.

4.5 Fine-tuning Data Mixture

As shown in Figure 7, from InternVL 1.5 to 2.0 and then to 2.5, the dataset has undergone iterative improvements in scale, quality, and diversity. In terms of data scale, the number of samples grows from 5.1M in InternVL 1.5 to 7.3M in InternVL 2.0, and further doubles to 16.3M in InternVL 2.5. For diversity, our training data spans multiple domains, including general QA, charts, documents, OCR, science, medical, GUI, code, mathematics, *et al.*, while covering multiple modalities such as single-image, multi-image, video, and text.

Task	Dataset
Type: Single-Image Datasets	
Captioning	TextCaps (en) [211], ShareGPT4o (en & zh) [35], InternVL-SA-1B-Caption (en & zh) [36], NewYorkerCaptionContest (en) [88], MMInstruct (en & zh) [155]
General QA	VQAv2 (en) [74], GQA (en) [98], OKVQA (en) [178], Visual7W (en) [317], MMInstruct (en & zh) [155], VSR (en) [147], FSC147 (en) [197], Objects365-YorN (en) [208], Hateful-Memes (en) [111]
Mathematics	GeoQA+ (en) [20], CLEVR-Math (en) [144], Super-CLEVR (en) [141], MapQA (en) [23], MAVIS (en) [300], Geometry3K (en) [164], TallyQA (en) [2], MetaMath (en) [281], GEOS (en) [206], UniGeo (en) [26], GeomVerse (en) [107], CMM-Math (zh) [154]
Chart	ChartQA (en) [181], MMTab (en) [310], PlotQA (en) [187], FigureQA (en) [105], VisText (en) [226], LRV-Instruction (en) [148], ArxivQA (en) [132], TabMWP (en) [166], MMC-Inst (en) [149], DVQA (en) [104], UniChart (en) [182], SimChart9K (en) [263], Chart2Text (en) [191], FinTabNet (zh) [312], SciTSR (zh) [39], Synthetic Chart2Markdown (en)
OCR	OCRVQA (en) [188], InfoVQA (en) [183], TextVQA (en) [212], ArT (en & zh) [40], HME100K (en) [288], COCO-Text (en) [238], CTW (zh) [287], LSVT (zh) [222], RCTW-17 (zh) [209], VCR (en & zh) [302], EST-VQA (en & zh) [253], ST-VQA (en) [17], EATEN (zh) [78], LLaVAR (en) [305], CASIA (zh) [146], Chinese-OCR (zh) [14], CyrillicHandwriting (ru) [239], IAM (en) [180], NAF (en) [52], POIE (en) [115], ReCTs (zh) [301], MTWI (zh) [82], TextOCR (en) [213], SROIE (en) [97], Synthetic Arxiv OCR (en), MTVQA (ko & ja & it & ru & de & fr & th & ar & vi) [227], Synthetic Image2Latex (en), Synthetic Handwritten OCR (zh), Synthetic Infographic2Markdown (en & zh)
Knowledge	KVQA (en) [207], A-OKVQA (en) [205], ViQuAE (en) [123], iNaturalist2018 (en) [237], MovieNet (en) [95], ART500K (en) [176], KonIQ-10K (en) [91], Synthetic Multidisciplinary Knowledge / QA (en & zh)
Document	DocVQA (en) [42], Docmatix (en) [121], DocReason25K (en) [93], Sujet-Finance-QA-Vision (en) [217]
Grounding	RefCOCO/+/g (en) [280, 177], GPT4Gen-RD-BoxCoT (en) [27], All-Seeing-V2 (en) [250], V3Det (en & zh) [243], DsLMF (en) [272], COCO-ReM (en & zh) [214], TolokaVQA (en) [236]
Science	AI2D (en) [109], ScienceQA (en) [165], TQA (en) [110], ChemVLM Data (en & zh) [129]
Conversation	ALLaVA (en & zh) [25], Viet-ShareGPT4o (vi) [59], Cambrain-GPT4o (en) [234] , RLAIF-V (en) [282], Laion-GPT4V (en) [119], TextOCR-GPT4V (en) [102], WildVision-GPT4o (en) [171], Synthetic Real-World Conversations (en & zh)
Medical	PMC-VQA (en) [303], VQA-RAD (en) [120], ImageCLEF (en) [72], PMC (en) [261], SLAKE (en & zh) [145], GMAI-VL (en & zh) [134], VQA-Med (en) [15], Medical-Diff-VQA (en) [94], PathVQA (en) [83], PMC-CaseReport (en) [260]
GUI	Screen2Words (en) [240], WebSight (en) [122], Widget-Caption (en) [136], RICOSCA (en) [55], Seeclick (en) [37], ScreenQA (en) [92], AMEX (en) [22], AITW (en) [198], Odyssey (en) [168], UIBert (en) [12], AndroidControl (en) [135], Mind2Web (en) [57], OmniACT (en) [106], WaveUI (en) [4]
Type: Multi-Image Datasets	
General QA	Img-Diff (en) [101], Birds-to-Words (en) [100], Spot-the-Diff (en) [100], MultiVQA (en) [100], NLVR2 (en) [216], ContrastiveCaption (en) [100], DreamSim (en) [100], InternVL-SA-1B-Caption (en & zh) [36]
Document	MP-DocVQA (en) [233], MP-Docmatix (en) [121]
Type: Video Datasets	
Captioning	Vript (en & zh) [269], OpenVid (en) [190], Mementos (en) [254], ShareGPT4o-Video (en & zh) [35], ShareGPT4Video (en & zh) [30], VideoGPT+ (en) [174]
General QA	VideoChat2-IT (en & zh) [130, 131], EgoTaskQA (en) [99], NTU RGB+D (en) [152], CLEVRER (en) [276], LLaVA-Video (en) [307], FineVideo (en) [67], PerceptionTest (en) [193], HiREST (en) [291], STAR (en) [259], EgoSchema (en) [175], ScanQA (en) [10], LSMDC (en) [201]
GUI	GUI-World (en) [24]
Type: Text Datasets	
General QA	UltraFeedback (en) [48], UltraChat (en) [58], Unnatural-Instructions (en) [90], NoRobots (en) [196], MOSS (en) [221], LIMA (en) [314], SlimOrca (en) [142], WizardLM-Evol-Instruct-70K (en) [265], Llama-3-Magpie-Pro (en) [266], Magpie-Qwen2-Pro (en & zh) [266], KOpen-HQ-Hermes-2.5-60K (ko) [179], Firefly (zh) [270], Dolly (en) [44], OpenAI-Summarize-TLDR (en) [21], Know-Saraswati-CoT (en) [114], FLAN (en) [258], FLANv2 (en & zh) [41]
Code	Code-Feedback (en) [311], Glaive-Code-Assistant (en) [73], XCoder-80K (en) [255], LeetCode (en & zh), Evol-Instruct-Code (en) [173], InternLM2-Code (en & zh) [19]
Long Context	Long-Instruction-with-Paraphrasing (en & zh) [286], LongCite (en & zh) [298], LongQLoRA (en) [271], LongAlpaca (en) [34]
Mathematics	GSM8K-Socratic (en) [43], NuminaMath-CoT/TIR (en) [128], Orca-Math (en) [189], MathQA (en) [6], InfinityMATH (en) [295], InternLM2-Math (en & zh) [19]
Knowledge	Synthetic Multidisciplinary Knowledge / QA (en)

Table 5: **Summary of the fine-tuning data mixture of InternVL 2.5.** We expanded our fine-tuning data mixture through extensive collection of open-source datasets and self-synthesized data. This mixture is predominantly in English (en) and Chinese (zh), with smaller portions in other languages, including Korean (ko), Japanese (ja), Italian (it), Russian (ru), German (de), French (fr), Thai (th), Arabic (ar), and Vietnamese (vi).

Model Name	MMMU (val)	MMMU (test)	MMMU-Pro (std10 / vision / overall)	MathVista (mini)	MATH-Vision (mini / full)	MathVerse (mini)	Olympiad Bench
LLaVA-OneVision-0.5B [124]	31.4	–	–	34.8	–	17.9	–
InternVL2-1B [35]	36.7	32.8	16.0 / 13.6 / 14.8	37.7	12.2 / 11.1	18.4	0.3
InternVL2.5-1B	40.9	35.8	23.3 / 15.5 / 19.4	43.2	16.8 / 14.4	28.0	1.7
Qwen2-VL-2B [246]	41.1	–	25.3 / 17.2 / 21.2	43.0	19.7 / 12.4	21.0	–
Aquila-VL-2B [76]	47.4	–	–	59.0	21.1 / 18.4	26.2	–
InternVL2-2B [35]	36.3	34.7	21.6 / 14.9 / 18.2	46.3	15.8 / 12.1	25.3	0.4
InternVL2.5-2B	43.6	38.2	27.3 / 20.1 / 23.7	51.3	13.5 / 14.7	30.6	2.0
Phi-3.5-Vision-4B [1]	43.0	–	26.3 / 13.1 / 19.7	43.9	17.4 / 15.5	24.1	–
InternVL2-4B [35]	47.9	41.4	28.2 / 21.3 / 24.7	58.6	17.8 / 16.5	32.0	1.1
InternVL2.5-4B	52.3	46.3	36.4 / 29.0 / 32.7	60.5	21.7 / 20.9	37.1	3.0
Ovis1.6-Gemma2-9B [169]	55.0	–	–	67.2	– / 18.8	–	–
MiniCPM-V2.6 [274]	49.8	–	30.2 / 24.2 / 27.2	60.6	16.1 / 17.5	25.7	–
Qwen2-VL-7B [246]	54.1	–	34.1 / 27.0 / 30.5	58.2	22.0 / 16.3	31.9	–
InternVL2-8B [35]	52.6	44.3	32.5 / 25.4 / 29.0	58.3	20.4 / 18.4	37.0	1.9
InternVL2.5-8B	56.0	48.9	38.2 / 30.4 / 34.3	64.4	22.0 / 19.7	39.5	4.9
InternVL-Chat-V1.5 [35]	46.8	41.0	29.5 / 19.9 / 24.7	53.5	15.8 / 15.0	28.4	0.6
InternVL2-26B [35]	51.2	43.8	32.8 / 27.1 / 30.0	59.4	23.4 / 17.0	31.1	3.5
InternVL2.5-26B	60.0	51.8	41.6 / 32.6 / 37.1	67.7	28.0 / 23.1	40.1	8.8
Cambrian-34B [234]	49.7	–	–	53.2	–	–	–
VILA-1.5-40B [143]	55.1	46.9	35.9 / 14.1 / 25.0	49.5	–	–	–
InternVL2-40B [35]	55.2	49.3	36.3 / 32.1 / 34.2	63.7	21.4 / 16.9	36.3	3.9
InternVL2.5-38B	63.9	57.6	48.0 / 44.0 / 46.0	71.9	32.2 / 31.8	49.4	12.1
GPT-4V [192]	63.1	–	–	58.1	– / 24.0	32.8	18.0
GPT-4o-20240513 [192]	69.1	–	54.0 / 49.7 / 51.9	63.8	– / 30.4	50.2	25.9
Claude-3.5-Sonnet [8]	68.3	–	55.0 / 48.0 / 51.5	67.7	–	–	–
Gemini-1.5-Pro [200]	62.2	–	49.4 / 44.4 / 46.9	63.9	– / 19.2	–	–
LLaVA-OneVision-72B [124]	56.8	–	38.0 / 24.0 / 31.0	67.5	–	39.1	–
NVLM-D-72B [50]	59.7	54.6	–	66.6	–	–	–
Molmo-72B [54]	54.1	–	–	58.6	–	–	–
Qwen2-VL-72B [246]	64.5	–	49.2 / 43.3 / 46.2	70.5	– / 25.9	–	–
InternVL2-Llama3-76B [35]	62.7	55.1	41.9 / 38.0 / 40.0	65.5	23.7 / 23.6	42.8	5.5
InternVL2.5-78B	70.1	61.8	51.4 / 45.9 / 48.6	72.3	34.9 / 32.2	51.7	11.6

Table 6: **Comparison of multimodal reasoning and mathematical performance.** MMMU [289] and MMMU-Pro [290] are multidisciplinary reasoning benchmarks, while MathVista [163], MATH-Vision [245], MathVerse [299], and OlympiadBench [80] are mathematics benchmarks. Part of results are collected from [54, 8, 290, 245, 299, 80] and the OpenCompass leaderboard [46].

In InternVL 2.5, single-image data constituted the majority with 45.92% of tokens, while multi-image data accounted for 9.37%, video data contributed 39.79%, and pure-text data made up 4.92%. Compared to earlier versions, multi-image and video data achieved the most notable increases, leading to the enhanced multi-image and long video comprehension abilities of InternVL 2.5. Quality improvements were achieved through unifying conversation templates, using language models to score and refine data, removing repetitive patterns, applying heuristic rules to filter low-quality samples, and rewriting short responses into high-quality and longer interactions. This ensured a robust dataset for model training.

5 Evaluation on Multimodal Capability

To comprehensively evaluate InternVL's performance on multimodal tasks, we employ a diverse set of benchmarks, including both well-established classic datasets and newly introduced ones provided by VLMEvalKit [63]. These benchmarks span a wide range of categories, aiming to provide a thorough and balanced assessment of InternVL's capabilities across various multimodal tasks.

5.1 Multimodal Reasoning and Mathematics

5.1.1 Benchmarks

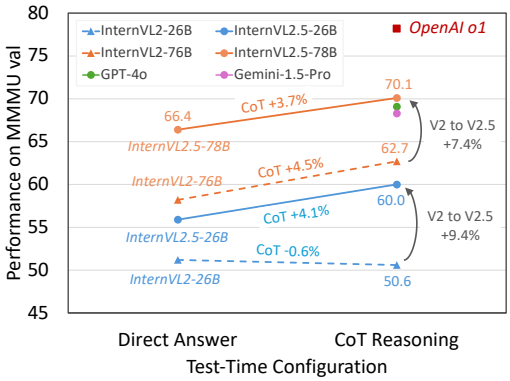
We evaluate InternVL's multimodal mathematical and reasoning capabilities through a comprehensive assessment across various discipline-related benchmarks.

Answer the preceding multiple choice question. The last line of your response should follow this format: 'Answer: \boxed{$LETTER}' (without quotes), where LETTER is one of the options. If you are uncertain or the problem is too complex, make a reasoned guess based on the information provided. Avoid repeating steps indefinitely—provide your best guess even if unsure. Think step by step logically, considering all relevant information before answering.

(a) CoT prompt for multiple-choice questions

Answer the preceding question. The last line of your response should follow this format: 'Answer: \boxed{$FINAL_ANSWER}' (without quotes), where 'FINAL_ANSWER' is your conclusion based on the reasoning provided. If you are uncertain or the problem is too complex, make a reasoned guess based on the information provided. Avoid repeating steps indefinitely—provide your best guess even if unsure. Think step by step logically, considering all relevant information before answering.

(b) CoT prompt for open-ended questions



(c) Model performance on MMMU val

Figure 9: **CoT prompts used in our model testing.** By leveraging these prompts for CoT reasoning, we can scale up testing time, significantly enhancing the performance of InternVL 2.5 models on MMMU [289].

MMMU [289]: MMMU is a benchmark evaluating MLLMs on college-level tasks across six disciplines, testing expert-level reasoning and advanced perception in specific fields. We report the maximum accuracy achieved across both direct-answer and CoT reasoning approaches on the MMMU validation and test sets.

MMMU-Pro [290]: MMMU-Pro is an upgraded version of the MMMU benchmark, designed to more accurately and rigorously evaluate the multimodal understanding and reasoning capabilities of models in a wide range of academic disciplines. We report three metrics: standard (10 options), vision, and overall (the average of standard and vision). Here, "standard" and "vision" are the maximum scores from the CoT and direct-answer settings, consistent with the original paper.

MathVista [163]: MathVista is a benchmark for evaluating MLLMs' mathematical reasoning in visual contexts, encompassing reasoning types such as algebra, geometry, and statistics. We report the scores on the testmini set.

MATH-Vision [245]: MATH-Vision is a high-quality dataset of 3,040 visually contextualized math problems sourced from real competitions. We report performance on both the testmini and full sets.

MathVerse [299]: MathVerse is a visual math benchmark for evaluating MLLMs in solving diagram-based math problems. It comprises 2,612 high-quality, multi-subject math problems, each transformed into six distinct versions with varying degrees of visual and textual information. We report performance on the testmini set.

OlympiadBench [80]: OlympiadBench is a bilingual, multimodal benchmark with high-difficulty math and physics problems from Olympiad competitions and Gaokao. Each problem is annotated with expert-level step-by-step reasoning, enabling detailed assessment of logical deduction and problem-solving abilities. This benchmark is challenging, and a well-defined CoT prompt can significantly improve performance.

5.1.2 Evaluation Results

Multidisciplinary reasoning ability reflects a model's capacity to comprehend, process, and manipulate abstract concepts, which is crucial for complex problem-solving and decision-making tasks. In the left section of Table 6, we provide a comparison of InternVL 2.5's performance on multidisciplinary reasoning-related benchmarks, including MMMU [289] and MMMU-Pro [290].

Here, we test both direct-answer and CoT reasoning performance, reporting the higher score. The results suggest that our model achieves encouraging improvements over existing open-source models, such as LLaVA-OneVision [124], NVLM [50], VILA 1.5 [143], and Qwen2-VL [246], as well as notable progress compared to earlier versions of the InternVL2 series. Specifically, InternVL2.5-78B achieves a score exceeding 70 on the MMMU validation set, representing a 7.4-point improvement over InternVL2-Llama3-76B. These results indicate that our model's performance is moving closer to that of some advanced closed-source models, such as GPT-4o [192], Claude-3.5-Sonnet [8], and Gemini-1.5-Pro [200]. Additionally, through majority voting, the score of InternVL2-Llama3-76B on the MMMU benchmark is improved from 62.7 to 65.3 when using CoT. We observe a similar phenomenon in InternVL 2.5 as well, which demonstrates that test-time scaling can improve the CoT reasoning of MLLMs.

Model Name	AI2D (w / wo M)	ChartQA (test avg)	TextVQA (val)	DocVQA (test)	InfoVQA (test)	OCR Bench	SEED-2 Plus	CharXiv (RQ / DQ)	VCR-EN-Easy (EM / Jaccard)
LLaVA-OneVision-0.5B [124]	57.1 / –	61.4	–	70.0	41.8	565	–	–	–
InternVL2-1B [35]	64.1 / 70.5	72.9	70.5	81.7	50.9	754	54.3	18.1 / 30.7	21.5 / 48.4
InternVL2.5-1B	69.3 / 77.8	75.9	72.0	84.8	56.0	785	59.0	19.0 / 38.4	91.5 / 97.0
Qwen2-VL-2B [246]	74.7 / 84.6	73.5	79.7	90.1	65.5	809	62.4	–	81.5 / –
Aquila-VL-2B [76]	75.0 / –	76.5	76.4	85.0	58.3	772	63.0	–	70.0 / –
InternVL2-2B [35]	74.1 / 82.3	76.2	73.4	86.9	58.9	784	60.0	21.0 / 40.6	32.9 / 59.2
InternVL2.5-2B	74.9 / 83.5	79.2	74.3	88.7	60.9	804	60.9	21.3 / 49.7	93.2 / 97.6
Phi-3.5-Vision-4B [1]	77.8 / 87.6	81.8	72.0	69.3	36.6	599	62.2	–	39.3 / 60.4
InternVL2-4B [35]	78.9 / 87.8	81.5	74.4	89.2	67.0	788	63.9	24.5 / 48.0	33.7 / 61.1
InternVL2.5-4B	81.4 / 90.5	84.0	76.8	91.6	72.1	828	66.9	24.9 / 61.7	93.7 / 97.8
Ovis1.6-Gemma2-9B [169]	84.4 / –	–	–	–	–	830	–	–	–
MiniCPM-V2.6 [274]	82.1 / –	82.4	80.1	90.8	–	852	65.7	31.0 / 57.1	73.9 / 85.7
Molmo-7B-D [54]	– / 93.2	84.1	81.7	92.2	72.6	694	–	–	–
Qwen2-VL-7B [246]	83.0 / 92.1	83.0	84.3	94.5	76.5	866	69.0	–	89.7 / 93.8
InternVL2-8B [35]	83.8 / 91.7	83.3	77.4	91.6	74.8	794	67.5	31.2 / 56.1	37.9 / 61.5
InternVL2.5-8B	84.5 / 92.8	84.8	79.1	93.0	77.6	822	69.7	32.9 / 68.6	92.6 / 97.4
InternVL-Chat-V1.5 [35]	80.7 / 89.8	83.8	80.6	90.9	72.5	724	66.3	29.2 / 58.5	14.7 / 51.4
InternVL2-26B [35]	84.5 / 92.5	84.9	82.3	92.9	75.9	825	67.6	33.4 / 62.4	74.5 / 86.7
InternVL2.5-26B	86.4 / 94.4	87.2	82.4	94.0	79.8	852	70.8	35.9 / 73.5	94.4 / 98.0
Cambrian-34B [234]	79.5 / –	75.6	76.7	75.5	46.0	600	–	27.3 / 59.7	79.7 / 89.3
VILA-1.5-40B [143]	69.9 / –	67.2	73.6	–	–	460	–	24.0 / 38.7	–
InternVL2-40B [35]	86.6 / 94.5	86.2	83.0	93.9	78.7	837	69.2	32.3 / 66.0	84.7 / 92.6
InternVL2.5-38B	87.6 / 95.1	88.2	82.7	95.3	83.6	842	71.2	42.4 / 79.6	94.7 / 98.2
GPT-4V [192]	78.2 / 89.4	78.5	78.0	88.4	75.1	645	53.8	37.1 / 79.9	52.0 / 65.4
GPT-4o-20240513 [192]	84.6 / 94.2	85.7	77.4	92.8	79.2	736	72.0	47.1 / 84.5	91.6 / 96.4
Claude-3-Opus [8]	70.6 / 88.1	80.8	67.5	89.3	55.6	694	44.2	30.2 / 71.6	62.0 / 77.7
Claude-3.5-Sonnet [8]	81.2 / 94.7	90.8	74.1	95.2	74.3	788	71.7	60.2 / 84.3	63.9 / 74.7
Gemini-1.5-Pro [200]	79.1 / 94.4	87.2	78.8	93.1	81.0	754	–	43.3 / 72.0	62.7 / 77.7
LLaVA-OneVision-72B [124]	85.6 / –	83.7	80.5	91.3	74.9	741	–	–	–
NVLM-D-72B [50]	85.2 / 94.2	86.0	82.1	92.6	–	853	–	–	–
Molmo-72B [54]	– / 96.3	87.3	83.1	93.5	81.9	–	–	–	–
Qwen2-VL-72B [246]	88.1 / –	88.3	85.5	96.5	84.5	877	–	–	91.3 / 94.6
InternVL2-Llama3-76B [35]	87.6 / 94.8	88.4	84.4	94.1	82.0	839	69.7	38.9 / 75.2	83.2 / 91.3
InternVL2.5-78B	89.1 / 95.7	88.3	83.4	95.1	84.1	854	71.3	42.4 / 82.3	95.7 / 94.5

Table 7: **Comparison of OCR, chart, and document understanding performance.** We evaluate OCR-related capabilities across 9 benchmarks, including AI2D [109], ChartQA [181], TextVQA [212], DocVQA [184], InfoVQA [183], OCRBench [158], SEED-2-Plus [125], CharXiv [257], and VCR [302]. Part of results are collected from [64, 54, 8, 257, 302] and the OpenCompass leaderboard [46].

Mathematical reasoning reflects a higher-level reasoning capability and enhances the potential of MLLMs in scientific and engineering applications. In the right-hand section of Table 6, we present InternVL 2.5's performance across four multimodal mathematical benchmarks. These results demonstrate significant progress over InternVL 2.0. Notably, InternVL2.5-78B achieved an accuracy of 72.3% on the MathVista test-mini set [163]. Additionally, on the challenging OlympiadBench [80], the InternVL 2.5 series showed an overall improvement compared to the 2.0 series. We attribute part of this advancement to our data filtering pipeline. Specifically, we observed that the 2.0 models frequently encountered deadlocks during CoT reasoning, failing to reach correct final answers, while this issue has been mitigated in the 2.5 series.

5.2 OCR, Chart, and Document Understanding

5.2.1 Benchmarks

We assess InternVL's OCR, chart, and document understanding capabilities through a comprehensive evaluation on a variety of OCR-related datasets.

AI2D [109]: AI2D is a dataset of over 5,000 elementary school science diagrams, each with detailed annotations and corresponding multiple-choice questions. For a fair comparison, we report results for both "mask" and "no mask" settings on the test set.

ChartQA [181]: ChartQA is a dataset focused on assessing models' abilities to interpret and reason with data visualizations such as charts and graphs. Our evaluation metric is the average relaxed accuracy across both human and augmented test sets in ChartQA.

TextVQA [212]: TextVQA is a dataset designed to benchmark visual reasoning based on text within images. It requires models to read and interpret text in images to accurately answer related questions. We report the VQA accuracy on the TextVQA validation set.

DocVQA [42]: DocVQA is a dataset aimed at evaluating models' ability to comprehend and retrieve information from text within document images. Performance is reported on the test set using the ANLS metric, which captures answer accuracy by measuring text similarity.

InfoVQA [183]: InfographicVQA is a dataset aimed at evaluating models' ability to interpret and reason with complex infographics that combine text, graphics, and visual elements. Performance is measured using the ANLS metric on the test set.

OCRBench [158]: OCRBench evaluates the OCR capabilities of MLLMs across five tasks: text recognition, scene text VQA, document VQA, key information extraction, and handwritten math expression recognition, with a maximum score of 1000.

SEEDBench-2-Plus [125]: SEED-Bench-2-Plus evaluates MLLMs on text-rich visual tasks, with 2,300 human-annotated questions across charts, maps, and webs. We report the average accuracy on this dataset.

CharXiv [257]: CharXiv is a comprehensive evaluation suite featuring 2,323 charts from scientific papers. It includes two types of questions: reasoning questions (RQ) requiring synthesis of complex visual information, and descriptive questions (DQ) assessing basic chart element understanding.

VCR [302]: Visual Caption Restoration (VCR) is a task that involves restoring partially hidden text within images by understanding both the visual content and the text. We report the Exact Match (EM) score and Jaccard similarity on the VCR-EN-Easy subset.

5.2.2 Evaluation Results

Table 7 provides a detailed comparison of InternVL 2.5 with its predecessor InternVL 2.0, other representative open-source models (*e.g.*, Qwen2-VL [246], LLaVA-OneVision [124]), and closed-source models (*e.g.*, GPT-4o [192], Claude-3.5-Sonnet [8]) on OCR-related tasks. Across most benchmarks, InternVL 2.5 achieves significant improvements over InternVL 2.0 at all model scales and demonstrates performance comparable to the current state-of-the-art model, Qwen2-VL-72B [246], reflecting the effectiveness of the improvements in training strategies and data quality.

However, at the 2B scale, InternVL2.5-2B underperforms compared to Qwen2-VL-2B on benchmarks such as TextVQA [212], DocVQA [184], and InfoVQA [183]. We suspect that, in addition to differences in data and training strategies, model architecture may also play a significant role. Specifically, Qwen2-VL-2B features a 600M vision encoder and a 1.5B language model, whereas InternVL2.5-2B employs a smaller 300M vision encoder paired with a 1.8B language model. It appears that, for a smaller-scale MLLM (*e.g.*, 2B), the size of the vision encoder plays a relatively important role in OCR performance, given the same total parameter budget.

Additionally, InternVL 2.5 demonstrates exceptional performance on the visual caption restoration (VCR) task [302]. The 2.5 series achieves a significant improvement over InternVL 2.0 on this task, with the 2B model reaching EM/Jaccard scores of 93.2/97.6, far surpassing the previous generation's 32.9/59.2. This improvement can be attributed to the introduction of a small portion of the VCR training set (approximately 22K samples). We find that the model's poor performance on VCR tasks was not due to inadequate OCR capabilities but rather to its insufficient instruction-following ability for task-specific directives. By leveraging these few but focused samples, InternVL 2.5 exhibits a remarkable enhancement in its instruction-following ability for the VCR task, resulting in a substantial performance boost.

5.3 Multi-Image Understanding

5.3.1 Benchmarks

We assess InternVL's capabilities in multi-image relation perception and understanding across various multi-image benchmarks.

BLINK [70]: The BLINK benchmark evaluates the core visual perception capabilities of MLLMs through 14 tasks inspired by classic computer vision challenges. Over half of the questions involve multiple images. Our results are reported on the validation set.

Model Name	BLINK (val)	Mantis Eval	MMIU	Muir Bench	MMT (val)	MIRB (avg)	RealWorld QA	MME-RW (EN)	WildVision (win rate)	R-Bench (dis)
LLaVA-OneVision-0.5B [124]	52.1	39.6	–	25.5	–	–	55.6	–	–	–
InternVL2-1B [35]	38.6	46.1	37.3	29.3	49.5	31.5	50.3	40.2	17.8	55.6
InternVL2.5-1B	42.0	51.2	38.5	29.9	50.3	35.6	57.5	44.2	43.4	59.0
Qwen2-VL-2B [246]	44.4	–	–	–	55.1	–	62.6	–	–	–
InternVL2-2B [35]	43.8	48.4	39.8	32.5	50.4	32.1	57.3	47.3	31.8	56.8
InternVL2.5-2B	44.0	54.8	43.5	40.6	54.5	36.4	60.1	48.8	44.2	62.2
Phi-3.5-Vision-4B [1]	58.3	–	–	–	53.6	–	53.6	–	–	55.5
InternVL2-4B [35]	46.1	61.3	43.3	40.5	55.7	39.9	60.7	52.1	44.2	64.5
InternVL2.5-4B	50.8	62.7	43.8	45.2	62.4	51.7	64.3	55.3	49.4	66.1
Qwen2-VL-7B [246]	53.2	–	–	–	64.0	–	70.1	56.5	–	64.0
MiniCPM-V2.6 [274]	53.0	69.0	–	–	60.8	–	65.0	–	–	–
InternVL2-8B [35]	50.9	65.4	42.0	48.7	60.0	50.0	64.4	53.5	54.4	67.9
InternVL2.5-8B	54.8	67.7	46.7	51.1	62.3	52.5	70.1	59.1	62.0	70.1
InternVL-Chat-V1.5 [35]	46.6	66.8	37.4	38.5	58.0	50.3	66.0	49.4	56.6	67.9
InternVL2-26B [35]	56.2	69.6	42.6	50.6	60.6	53.7	68.3	58.7	62.2	70.1
InternVL2.5-26B	61.8	75.6	49.4	61.1	66.9	55.7	74.5	61.8	65.2	72.9
Cambrian-34B [234]	–	–	–	–	–	–	67.8	44.1	–	–
InternVL2-40B [35]	57.2	71.4	47.9	54.4	66.2	55.2	71.8	61.8	63.2	73.3
InternVL2.5-38B	63.2	78.3	55.3	62.7	70.0	61.2	73.5	64.0	66.4	72.1
GPT-4V [192]	54.6	62.7	–	62.3	64.3	53.1	61.4	–	71.8	65.6
GPT-4o-20240513 [192]	68.0	–	55.7	68.0	65.4	–	75.4	45.2	80.6	77.7
Claude-3.5-Sonnet [8]	–	–	53.4	–	–	–	60.1	51.6	–	–
Gemini-1.5-Pro [200]	–	–	53.4	–	64.5	–	67.5	38.2	–	–
LLaVA-OneVision-72B [124]	55.4	77.6	–	54.8	–	–	71.9	–	–	–
Qwen2-VL-72B [246]	–	–	–	–	71.8	–	77.8	–	–	–
InternVL2-Llama3-76B [35]	56.8	73.7	44.2	51.2	67.4	58.2	72.2	63.0	65.8	74.1
InternVL2.5-78B	63.8	77.0	55.8	63.5	70.8	61.1	78.7	62.9	71.4	77.2

Table 8: **Comparison of multi-image and real-world understanding performance.** Multi-image benchmarks include BLINK [70], Mantis-Eval [100], MMIU [186], MuirBench [241], MMT-Bench [277], and MIRB [308]. Real-world benchmarks encompass RealWorldQA [47], MME-RealWorld [306], WildVision [171], and R-Bench [126]. Part of the results are sourced from the benchmark papers and the OpenCompass leaderboard [46].

Mantis-Eval [100]: Mantis-Eval is a meticulously curated small-scale benchmark for evaluating MLLMs' reasoning capabilities across multiple images. It comprises 217 challenging, human-annotated problems covering topics such as size perception and weight comparison.

MMIU [186]: MMIU is an extensive benchmark suite developed to rigorously assess the performance of MLLMs in multi-image tasks. It encompasses 7 distinct types of multi-image relationships and spans 52 diverse tasks, providing a comprehensive framework for evaluation.

MuirBench [241]: MuirBench is a comprehensive benchmark for evaluating MLLMs capabilities in multi-image understanding. It spans 12 tasks and 10 types of multi-image relations and enhances model assessment with unanswerable instance variants.

MMT-Bench [277]: MMT-Bench evaluates MLLMs on multimodal tasks like driving and navigation, focusing on recognition, reasoning, and planning, with many sub-tasks requiring multi-image understanding. To speed up testing, results are reported on the validation set.

MIRB [308]: MIRB is a benchmark designed to evaluate the ability of MLLMs to understand and reason across multiple images. It contains four task categories: perception, visual world knowledge, reasoning, and multi-hop reasoning. The reported performance is the average score across these four categories.

5.3.2 Evaluation Results

As multi-image content becomes an increasingly common form of information exchange on the internet, it is essential for models to possess the ability to simultaneously understand and analyze relationships between multiple images. In the left part of Table 8, we evaluate the multi-image understanding capabilities of InternVL 2.5 across six diverse benchmarks: BLINK [70], Mantis-Eval [100], MMIU [186], MuirBench [241], MMT-Bench [277], and MIRB [308]. These benchmarks test a range of skills, including reasoning across images, integrating information, and addressing task-specific requirements.

InternVL 2.5 achieves consistent improvements over InternVL 2.0 across all model scales, reflecting enhanced reasoning ability and better integration of multi-image information. For instance, at the 2B scale, InternVL2.5-2B delivers significant gains on Mantis-Eval (54.8 *vs.* 48.4) and MuirBench (40.6 *vs.* 32.5). These advancements can be largely attributed to the inclusion of additional multi-image datasets, as detailed in Section 4.5. These datasets, which were carefully curated and of high quality, played a critical role in improving the model's ability to understand and reason across multiple visual inputs.

At larger scales, InternVL 2.5 demonstrates substantial progress and achieves competitive performance with advanced closed-source models. For example, InternVL2.5-78B scores 55.8 on MMIU, closely matching GPT-4o's 55.7, and achieves a score of 70.8 on MMT-Bench, surpassing GPT-4o's 65.4. These results highlight the importance of scaling model size and incorporating high-quality training data specifically tailored for multi-image tasks. However, on BLINK and MuirBench, our model still exhibits a performance gap of around 5 points compared to GPT-4o [192], suggesting that further improvements are needed, potentially through the inclusion of additional high-quality multi-image training data.

5.4 Real-World Comprehension

5.4.1 Benchmarks

We assess InternVL's performance on a suite of real-world benchmarks designed to evaluate its capabilities on realistic and complex tasks.

RealWorldQA [47]: RealWorldQA is a benchmark designed to evaluate the real-world spatial understanding capabilities of MLLMs. It contains more than 700 images, each accompanied by a question and a verifiable answer, from various real-world scenarios.

MME-RealWorld [306]: MME-RealWorld is a benchmark for evaluating MLLMs on complex, high-resolution image tasks across 43 real-world scenarios in 5 domains. Here, we test the English full set of the dataset.

WildVision [171]: WildVision-Bench is a benchmark designed to evaluate MLLMs in the wild with human preferences. It comprises 500 high-quality samples meticulously curated from real-world user QA interactions. The benchmark uses a win rate metric to quantify the performance of models, providing insights into their ability to meet human expectations in practical applications.

R-Bench [126]: R-Bench is a benchmark designed to evaluate the robustness of MLLMs against real-world image distortions, measuring their resilience in handling corrupted images in practical scenarios. We report the absolute robustness overall score for the MCQ task, which is the average score across low, mid, and high difficulty levels, corresponding to "R-Bench-Dis" in VLMEvalKit.

5.4.2 Evaluation Results

Given the complexity and dynamic nature of real-world environments, models must be robust enough to handle a wide range of challenging conditions. As shown in the right part of Table 8, InternVL 2.5 achieves leading performance across four real-world understanding benchmarks, including RealWorldQA [47], MME-RealWorld [306], WildVision [171], and R-Bench [126], and significantly outperforms the previous version, InternVL 2.0. This indicates that InternVL 2.5 has a stronger potential for practical application in complex and ever-changing real-world scenarios.

In benchmarks like RealWorldQA, MME-RealWorld, and R-Bench, which involve multiple-choice questions, InternVL 2.5 demonstrates strong real-world perceptual and understanding abilities. Differently, the WildVision benchmark uses GPT-4o [192] as the judge model to evaluate the performance of various MLLM against the reference model, Claude-3-Sonnet [8]. In this benchmark, the model's output quality and user experience are key metrics. Although InternVL2.5-78B performs well in providing concise answers, it still shows a gap when generating longer responses to match human preferences. Specifically, InternVL2.5-78B scores 71.4, while GPT-4o scores 80.6, indicating a notable difference in user experience.

These results indicate that, while InternVL 2.5 delivers accurate and concise responses across most tasks, there is potential for improvement in generating more personalized and detailed answers. Future work will focus on enhancing the model's performance in open-ended tasks and complex interactions, aiming to better align with human preferences, bridge the gap in user experience with GPT-4o.

5.5 Comprehensive Multimodal Evaluation

5.5.1 Benchmarks

We evaluate InternVL's comprehensive multimodal capabilities through a range of benchmarks, including:

MME [68]: MME is the first comprehensive evaluation benchmark designed for MLLMs. It assesses models' perception and cognitive abilities across 14 subtasks, including object presence, counting, position, color recognition, as well as commonsense reasoning, numerical computation, text translation, and code reasoning. We report the overall score across all tasks.

MMBench [156]: MMBench evaluates the multimodal understanding of MLLMs through nearly 3,000 multiple-choice questions spanning 20 dimensions. It supports both English and Chinese versions, and we present the model's performance scores on the test set.

MMBench v1.1 [156]: Compared to MMBench, MMBench v1.1 features a refined dataset with a small number of noisy or low-quality questions removed, resulting in a subtle improvement in overall data quality. We report the model's performance on the English version of the test set.

MMVet [283]: MMVet is a benchmark designed to assess the integrated capabilities of MLLMs on complex tasks. It evaluates six core competencies: recognition, knowledge, spatial awareness, language generation, OCR, and mathematics, across 16 integrated tasks. Note that VLMEvalKit uses GPT-4-Turbo as the scoring model for this benchmark, which yields slightly lower scores compared to the official evaluation server.

MMVet v2 [284]: Expanding on MMVet, MMVet v2 introduces an enhanced benchmark with a new capability: image-text sequence understanding, allowing for the assessment of models' ability to process interleaved content. Here, we utilize the official evaluation server for scoring, which employs GPT-4-0613 as the scoring model.

MMStar [28]: MMStar is a benchmark for evaluating the multimodal capabilities of MLLMs. It includes 1,500 carefully curated samples focusing on advanced visual and language understanding, minimizing data leakage, and emphasizing visual dependency.

5.5.2 Evaluation Results

Comprehensive multimodal evaluation benchmarks, such as MME [68], the MMBench series [156], the MMVet series [283, 284], and MMStar [28], provide valuable and widely adopted frameworks for assessing model performance across a diverse set of multimodal tasks.

As shown in the left section of Table 9, the InternVL 2.5 models consistently outperform the previous InternVL 2.0 series across various model sizes, especially for smaller models with 1B-8B parameters. For example, in the MMBench v1.0 benchmark, which evaluates tasks in both English and Chinese, the InternVL 2.5 models show significant improvements. The InternVL2.5-4B achieves a score of 81.1/79.3, surpassing the InternVL2-4B's 78.6/73.9, while the InternVL2.5-8B reaches 84.6/82.6, outperforming the InternVL2-8B's 81.7/81.2.

It is also noteworthy that, while we have significantly improved the performance of smaller models on the MMVet series benchmarks, our largest model, InternVL2.5-78B, still does not surpass the Qwen2-VL-72B [246]. Currently, the state-of-the-art models on MMVet v2 remain closed-source models like GPT-4o [192] and Claude-3.5-Sonnet [8]. This highlights the gap between open-source models and closed-source ones in multimodal integrated capability. We recognize this as an important direction for future development.

5.6 Multimodal Hallucination Evaluation

5.6.1 Benchmarks

We evaluate InternVL's tendency toward hallucinations across four different benchmarks, including:

HallusionBench [77]: HallusionBench is a benchmark for evaluating image-context reasoning in MLLMs through a Yes/No judgment question format, focusing on challenges such as language hallucination and visual illusion. We report performance using the average scores of its three metrics: aAcc, fAcc, and qAcc.

MMHal-Bench [223]: MMHal-Bench is a benchmark designed to evaluate hallucinations in MLLMs. It includes 96 challenging questions derived from images in the OpenImages dataset, along with their corresponding ground-truth answers and image content. Scoring is conducted using GPT-4o, with scores ranging from 0 to 6.

Model Name	MME (sum)	MMB (EN / CN)	MMBv1.1 (EN)	MMVet (turbo)	MMVetv2 (0613)	MMStar	HallBench (avg)	MMHal (score)	CRPE (relation)	POPE (avg)
LLaVA-OneVision-0.5B [124]	1438.0	61.6 / 55.5	59.6	32.2	–	37.7	27.9	–	–	–
InternVL2-1B [35]	1794.4	65.4 / 60.7	61.6	32.7	36.1	45.7	34.0	2.25	57.5	87.3
InternVL2.5-1B	1950.5	70.7 / 66.3	68.4	48.8	43.2	50.1	39.0	2.49	60.9	89.9
Qwen2-VL-2B [246]	1872.0	74.9 / 73.5	72.2	49.5	–	48.0	41.7	–	–	–
InternVL2-2B [35]	1876.8	73.2 / 70.9	70.2	39.5	39.6	50.1	37.9	2.52	66.3	88.3
InternVL2.5-2B	2138.2	74.7 / 71.9	72.2	60.8	52.3	53.7	42.6	2.94	70.2	90.6
Phi-3.5-Vision-4B [1]	–	76.0 / 66.1	72.1	43.2	–	47.5	40.5	–	–	–
InternVL2-4B [35]	2059.8	78.6 / 73.9	75.8	51.0	46.6	54.3	41.9	2.75	71.1	87.2
InternVL2.5-4B	2337.5	81.1 / 79.3	79.3	60.6	55.4	58.3	46.3	3.31	75.5	90.9
Qwen2-VL-7B [246]	2326.8	83.0 / 80.5	80.7	62.0	–	60.7	50.6	3.40	74.4	88.1
MiniCPM-V2.6 [274]	2348.4	81.5 / 79.3	78.0	60.0	–	57.5	48.1	3.60	75.2	87.3
InternVL2-8B [35]	2210.3	81.7 / 81.2	79.5	54.2	52.3	62.0	45.2	3.33	75.8	86.9
InternVL2.5-8B	2344.1	84.6 / 82.6	83.2	62.8	58.1	62.8	50.1	3.65	78.4	90.6
InternVL-Chat-V1.5 [35]	2194.2	82.2 / 82.0	80.3	61.5	51.5	57.3	50.3	3.11	75.4	88.4
InternVL2-26B [35]	2260.7	83.4 / 82.0	81.5	62.1	57.2	61.2	50.7	3.55	75.6	88.0
InternVL2.5-26B	2373.3	85.4 / 85.5	84.2	65.0	60.8	66.5	55.0	3.70	79.1	90.6
Cambrian-34B [234]	–	80.4 / 79.2	78.3	53.2	–	54.2	41.6	–	–	–
InternVL2-40B [35]	2307.5	86.8 / 86.5	85.1	65.5	63.8	65.4	56.9	3.75	77.6	88.4
InternVL2.5-38B	2455.8	86.5 / 86.3	85.5	68.8	62.1	67.9	56.8	3.71	78.3	90.7
GPT-4V [192]	1926.6	81.0 / 80.2	80.0	67.5	66.3	56.0	46.5	–	–	–
GPT-4o-20240513 [192]	–	83.4 / 82.1	83.1	69.1	71.0	64.7	55.0	4.00	76.6	86.9
Claude-3-Opus [8]	1586.8	63.3 / 59.2	60.1	51.7	55.8	45.7	37.8	–	–	–
Claude-3.5-Sonnet [8]	–	82.6 / 83.5	80.9	70.1	71.8	65.1	55.5	–	–	–
Gemini-1.5-Pro [200]	–	73.9 / 73.8	74.6	64.0	66.9	59.1	45.6	–	–	–
LLaVA-OneVision-72B [124]	2261.0	85.8 / 85.3	85.0	60.6	–	65.8	49.0	–	–	–
Qwen2-VL-72B [246]	2482.7	86.5 / 86.6	85.9	74.0	66.9	68.3	58.1	–	–	–
InternVL2-Llama3-76B [35]	2414.7	86.5 / 86.3	85.5	65.7	68.4	67.4	55.2	3.83	77.6	89.0
InternVL2.5-78B	2494.5	88.3 / 88.5	87.4	72.3	65.5	69.5	57.4	3.89	78.8	90.8

Table 9: **Comparison of comprehensive multimodal understanding and hallucination performance.** Comprehensive multimodal benchmarks include MME [68], MMBench series [156], MMVet series [283, 284], and MMStar [28]. Hallucination benchmarks encompass HallusionBench [77], MMHal [223], CRPE [250], and POPE [139]. Part of the results are sourced from the benchmark papers and the OpenCompass leaderboard [46].

CRPE [250]: CRPE is a benchmark that measures the hallucination level of the relation between objects using multiple-choice questions. We report accuracy on the relation subset for this benchmark.

POPE [139]: POPE is a benchmark for evaluating object hallucination in MLLMs, utilizing binary questions to quantify and analyze hallucination tendencies. We report the average F1 score across three categories: random, popular, and adversarial.

5.6.2 Evaluation Results

We evaluate the performance of InternVL on four key hallucination evaluation benchmarks: HallusionBench [77], MMHal [223], CRPE [250], and POPE [139]. These benchmarks assess the frequency of hallucinations, or factual inaccuracies, across multimodal tasks, providing a measure of model reliability in handling complex inputs like text and images.

The InternVL 2.5 models show significant progress over the InternVL 2.0 series, particularly in smaller models (*e.g.*, 1B-8B parameters). For instance, InternVL2.5-1B and InternVL2.5-2B demonstrate improved scores on all hallucination benchmarks, with the 1B model achieving a 39.0 score on HallusionBench, up from 34.0 in the earlier version. Similarly, the 2B model improved to 42.6, outperforming the previous 2B model by nearly 5 points. These results indicate substantial gains in reducing hallucinations while handling multimodal data.

The largest model, InternVL2.5-78B, also shows improvements, reducing hallucinations compared to both prior versions and other leading models. It scores 57.4 on HallusionBench, competing with top models like Qwen2-VL-72B (58.1) and GPT-4o (55.0). Although InternVL2.5-78B demonstrates relatively low hallucination rates on these hallucination evaluation benchmarks, some hallucinations are still inevitably present when generating long responses in practical use. This is a challenge we plan to tackle in future work.

Model Name	RefCOCO			RefCOCO+			RefCOCOg		avg.
	val	test-A	test-B	val	test-A	test-B	val	test	
Grounding-DINO-L [153]	90.6	93.2	88.2	82.8	89.0	75.9	86.1	87.0	86.6
UNINEXT-H [267]	92.6	94.3	91.5	85.2	89.6	79.8	88.7	89.4	88.9
ONE-PEACE [247]	92.6	94.2	89.3	88.8	92.2	83.2	89.2	89.3	89.8
Shikra-7B [27]	87.0	90.6	80.2	81.6	87.4	72.1	82.3	82.2	82.9
Ferret-v2-13B [297]	92.6	95.0	88.9	87.4	92.1	81.4	89.4	90.0	89.6
CogVLM-Grounding-17B [248]	92.8	94.8	89.0	88.7	92.9	83.4	89.8	90.8	90.3
MM1.5 [296]	–	92.5	86.7	–	88.7	77.8	–	87.1	–
Qwen2-VL-7B [246]	91.7	93.6	87.3	85.8	90.5	79.5	87.3	87.8	87.9
TextHawk2 [285]	91.9	93.0	87.6	86.2	90.0	80.4	88.2	88.1	88.2
InternVL2-8B [35]	87.1	91.1	80.7	79.8	87.9	71.4	82.7	82.7	82.9
InternVL2.5-8B	90.3	94.5	85.9	85.2	91.5	78.8	86.7	87.6	87.6
Qwen2-VL-72B [246]	93.2	95.3	90.7	90.1	93.8	85.6	89.9	90.4	91.1
InternVL2-Llama3-76B [35]	92.2	94.8	88.4	88.8	93.1	82.8	89.5	90.3	90.0
InternVL2.5-78B	93.7	95.6	92.5	90.4	94.7	86.9	92.7	92.2	92.3

Table 10: **Comparison of visual grounding performance.** We evaluate InternVL's visual grounding capability on RefCOCO, RefCOCO+, and RefCOCOg datasets [108, 177]. Parts of the results are collected from [246].

5.7 Visual Grounding

5.7.1 Benchmarks

We evaluate InternVL's visual grounding capability via referring expression comprehension (REC) on the RefCOCO, RefCOCO+, and RefCOCOg datasets, where the model identifies target objects in images from given descriptions.

RefCOCO [108]: Built on COCO, this dataset contains 19,994 images with 142,210 referring expressions for 50,000 objects, split into subsets like test A (people-focused) and test B (other objects) for REC tasks.

RefCOCO+ [108]: Similar to RefCOCO but emphasizing attribute-based descriptions without absolute location cues. It includes 19,992 images and 141,564 expressions, requiring models to focus on descriptive attributes.

RefCOCOg [177]: With 25,799 images and 95,010 expressions, this dataset features longer, more complex expressions, and challenging models to manage intricate language in REC tasks.

5.7.2 Evaluation Results

Visual grounding is critical for connecting textual descriptions with visual content, enabling accurate multimodal interaction. Table 10 compares InternVL 2.5 with its predecessor, InternVL 2.0, at the 8B and 78B scales, alongside other leading MLLMs (*e.g.*, CogVLM-Grounding-17B [248], Qwen2-VL [246]) and specialized grounding models (*e.g.*, Grounding-DINO-L [153], UNINEXT-H [267], ONE-PEACE [247]), evaluated on the RefCOCO [108], RefCOCO+ [108], and RefCOCOg [177] datasets.

InternVL2.5-8B improves its predecessor's performance, with the average score rising from 82.9 to 87.6, achieving comparable results to Qwen2-VL-7B (87.6 *vs.* 87.9), though slightly behind Ferret-v2-13B [297] and CogVLM-Grounding-17B [248], which benefit from fine-tuning for grounding and larger model sizes. At the larger scale, InternVL2.5-78B achieves state-of-the-art performance with an average score of 92.3, a 2.3-point improvement over InternVL2-Llama3-76B, surpassing Qwen2-VL-72B [246]. These gains highlight the effectiveness of our data and training optimizations, significantly enhancing localization capabilities.

5.8 Multimodal Multilingual Understanding

5.8.1 Benchmarks

We assess InternVL's multimodal multilingual understanding capabilities using three representative benchmarks:

MMMB and Multilingual MMBench [218]: MMMB is a large-scale multilingual multimodal benchmark with 6 languages, 15 categories, and 12,000 questions. The languages evaluated are English (en), Chinese (zh), Portuguese (pt), Arabic (ar), Turkish (tr), and Russian (ru). Multilingual MMBench extends MMBench [156] to these 6 languages using GPT-4 translation for multilingual understanding evaluation.

Model Name	MMMB						Multilingual MMBench						MTVQA
	en	zh	pt	ar	tr	ru	en	zh	pt	ar	tr	ru	(avg)
InternVL2-1B [35]	73.2	67.4	55.5	53.5	43.8	55.2	67.9	61.2	50.8	43.3	31.8	52.7	12.6
InternVL2.5-1B	78.8	70.2	61.5	55.0	45.3	61.1	72.5	64.7	57.0	43.0	37.8	53.2	21.4
Qwen2-VL-2B [246]	78.3	74.2	72.6	68.3	61.8	72.8	72.1	71.1	69.9	61.1	54.4	69.3	20.0
InternVL2-2B [35]	79.4	71.6	54.0	43.5	46.4	48.1	73.8	69.6	51.4	29.8	31.3	42.3	10.9
InternVL2.5-2B	81.4	74.4	58.2	48.3	46.4	53.2	76.5	71.6	55.9	37.3	33.9	44.8	21.8
InternVL2-4B [35]	82.0	76.1	75.6	54.3	51.2	67.4	77.3	72.4	72.6	43.6	46.5	61.2	15.3
InternVL2.5-4B	83.7	81.0	79.7	76.0	70.5	79.9	82.3	81.1	78.9	73.4	68.1	76.2	28.4
mPLUG-Owl2 [275]	67.3	61.0	59.7	45.8	45.4	62.6	66.2	59.4	58.2	37.9	47.7	60.4	–
Qwen2-VL-7B [246]	83.9	82.4	81.2	79.0	74.7	82.4	81.8	81.6	79.1	75.6	74.5	79.3	25.6
InternVL2-8B [35]	83.4	81.5	76.1	66.3	69.2	75.7	82.9	81.8	76.0	60.5	66.0	74.4	20.9
InternVL2.5-8B	84.3	83.1	78.6	69.3	71.5	79.5	83.8	83.2	79.4	64.3	67.8	77.3	27.6
InternVL-Chat-V1.5 [35]	82.6	80.8	76.3	65.2	68.6	74.0	81.1	80.2	76.9	56.2	66.7	71.0	20.5
InternVL2-26B [35]	83.8	81.7	78.0	68.8	69.3	76.3	82.7	81.8	77.8	61.9	69.6	74.4	17.7
InternVL2.5-26B	86.2	83.8	81.6	73.3	73.7	82.8	86.1	85.5	80.7	67.5	75.0	79.6	28.5
InternVL2-40B [35]	85.3	84.1	81.1	70.3	74.2	81.4	86.2	85.8	82.8	64.0	74.2	81.8	20.6
InternVL2.5-38B	86.4	85.1	84.1	84.3	82.8	84.9	87.5	88.6	85.3	84.5	84.0	85.9	31.7
GPT-4V [192]	75.0	74.2	71.5	73.5	69.0	73.1	77.6	74.4	72.5	72.3	70.5	74.8	22.0
GPT-4o [192]	–	–	–	–	–	–	–	–	–	–	–	–	27.8
Gemini-1.0-Pro [228]	75.0	71.9	70.6	69.9	69.6	72.7	73.6	72.1	70.3	61.1	69.8	70.5	–
Qwen2-VL-72B [246]	86.8	85.3	85.2	84.8	84.2	85.3	86.9	87.2	85.8	83.5	84.4	85.3	30.9
InternVL2-Llama3-76B [35]	85.3	85.1	82.8	82.8	83.0	83.7	87.8	87.3	85.9	83.1	85.0	85.7	22.0
InternVL2.5-78B	86.3	85.6	85.1	84.8	83.1	85.4	90.0	89.7	87.4	83.3	84.9	86.3	31.9

Table 11: **Comparison of multimodal multilingual performance.** We evaluate multilingual capabilities across 3 benchmarks, including MMMB [218], Multilingual MMBench [218] and MTVQA [227]. The languages evaluated are English (en), Chinese (zh), Portuguese (pt), Arabic (ar), Turkish (tr), and Russian (ru).

MTVQA [227]: MTVQA is a multilingual benchmark tailored for text-centric visual question answering. It includes high-quality, expert human annotations across nine languages, specifically addressing the "visual-text misalignment" challenge in multilingual contexts. We report the average score of MTVQA.

5.8.2 Evaluation Results

Multilingual ability is critical for MLLMs as it expands their application and improves cross-language communication. For global deployment, MLLMs must effectively handle both high-resource and low-resource languages. As shown in Table 11, we evaluated our model's performance on three multilingual benchmarks: MMMB [218], Multilingual MMBench [218], and MTVQA [227].

A comparison between InternVL2.5-78B and Qwen2-VL-72B [246] reveals that, despite differences in training data, model architecture, and training strategies, their multilingual performance is quite similar. This suggests that the multilingual capabilities of MLLMs are largely inherited from the underlying language model. Both models share the same LLM, indicating that a strong multilingual LLM forms the foundation for effective multilingual performance in MLLMs.

5.9 Video Understanding

5.9.1 Benchmarks

Video-MME [69]: Video-MME is a benchmark for evaluating MLLMs in full-spectrum video analysis. It features a wide variety of video types across multiple domains and durations, with multimodal inputs including video, subtitles, and audio. For this benchmark, we test with four settings: 16, 32, 48, and 64 frames, and report the maximum results. We report results for both "with subtitle" and "without subtitle" settings.

MVBench [131]: MVBench is a video understanding benchmark designed to comprehensively evaluate the temporal awareness of MLLMs in the open world. It covers 20 challenging video tasks, ranging from perception to cognition, which cannot be effectively solved using a single frame. We test this benchmark using 16 frames.

MMBench-Video [65]: MMBench-Video is a quantitative benchmark for evaluating MLLMs' video understanding and temporal reasoning skills, covering diverse domains, multi-shot long videos, and features like hallucination, commonsense reasoning, and temporal reasoning. For this benchmark, we test with four different settings: 16, 32, 48, and 64 frames, and report the maximum scores.

Model Name	Video-MME (wo / w sub)	MVBench	MMBench-Video (val)	MLVU (M-Avg)	LongVideoBench (val total)	CG-Bench (long / clue acc.)
InternVL2-1B [35]	42.9 / 45.4	57.5	1.14	51.6	43.3	–
InternVL2.5-1B	50.3 / 52.3	64.3	1.36	57.3	47.9	–
Qwen2-VL-2B [246]	55.6 / 60.4	63.2	–	–	–	–
InternVL2-2B [35]	46.2 / 49.1	60.2	1.30	54.3	46.0	–
InternVL2.5-2B	51.9 / 54.1	68.8	1.44	61.4	52.0	–
InternVL2-4B [35]	53.9 / 57.0	64.0	1.45	59.9	53.0	–
InternVL2.5-4B	62.3 / 63.6	71.6	1.73	68.3	55.2	–
VideoChat2-HD [130]	45.3 / 55.7	62.3	1.22	47.9	–	–
MiniCPM-V-2.6 [274]	60.9 / 63.6	–	1.70	–	54.9	–
LLaVA-OneVision-7B [124]	58.2 / –	56.7	–	–	–	–
Qwen2-VL-7B [246]	63.3 / 69.0	67.0	1.44	–	55.6	–
InternVL2-8B [35]	56.3 / 59.3	65.8	1.57	64.0	54.6	–
InternVL2.5-8B	64.2 / 66.9	72.0	1.68	68.9	60.0	–
InternVL2-26B [35]	57.0 / 60.2	67.5	1.67	64.2	56.1	–
InternVL2.5-26B	66.9 / 69.2	75.2	1.86	72.3	59.9	–
Oryx-1.5-32B [160]	67.3 / 74.9	70.1	1.52	72.3	–	–
VILA-1.5-40B [143]	60.1 / 61.1	–	1.61	56.7	–	–
InternVL2-40B [35]	66.1 / 68.6	72.0	1.78	71.0	60.6	–
InternVL2.5-38B	70.7 / 73.1	74.4	1.82	75.3	63.3	–
GPT-4V/4T [3]	59.9 / 63.3	43.7	1.53	49.2	59.1	–
GPT-4o-20240513 [192]	71.9 / 77.2	–	1.63	64.6	66.7	–
GPT-4o-20240806 [192]	–	–	1.87	–	–	41.8 / 58.3
Gemini-1.5-Pro [200]	75.0 / 81.3	–	1.30	–	64.0	40.1 / 56.4
VideoLLaMA2-72B [38]	61.4 / 63.1	62.0	–	–	–	–
LLaVA-OneVision-72B [124]	66.2 / 69.5	59.4	–	66.4	61.3	–
Qwen2-VL-72B [246]	71.2 / 77.8	73.6	1.70	–	–	41.3 / 56.2
InternVL2-Llama3-76B [35]	64.7 / 67.8	69.6	1.71	69.9	61.1	–
InternVL2.5-78B	72.1 / 74.0	76.4	1.97	75.7	63.6	42.2 / 58.5

Table 12: **Comparison of video understanding performance.** We evaluate InternVL's video understanding capabilities across 6 benchmarks. For Video-MME [69], MMBench-Video [65], MLVU [315], and LongVideoBench [262], we test with four different settings: 16, 32, 48, and 64 frames, and report the maximum results. For MVBench [131], we conduct testing using 16 frames. For CG-Bench [7], we use 32 frames.

MLVU [315]: MLVU is a comprehensive benchmark designed to evaluate MLLMs in long video understanding tasks, featuring videos ranging from 3 minutes to 2 hours. It includes nine different evaluation tasks divided into three categories: holistic understanding, single-detail understanding, and multi-detail understanding. We evaluate four settings: 16, 32, 48, and 64 frames, and report the highest "M-Avg" results.

LongVideoBench [262]: LongVideoBench focuses on referring reasoning tasks that involve long-frame inputs, requiring the model to accurately retrieve and reason about detailed multimodal information based on referring queries. We test four settings—16, 32, 48, and 64 frames—and report the best results on the validation set.

CG-Bench [7]: CG-Bench is a benchmark for evaluating long video understanding in MLLMs. Unlike existing benchmarks, it focuses on models' ability to retrieve relevant clues for answering questions. It includes 1,219 curated videos and over 12,000 question-answer pairs. Two novel clue-based evaluation methods are introduced to assess genuine video understanding. We test this benchmark using 32 frames.

5.9.2 Evaluation Results

Video understanding is vital for assessing MLLMs' ability to process temporal and multimodal information. To evaluate this comprehensively, we tested six benchmarks: Video-MME [69], MVBench [131], MMBench-Video [65], MLVU [315], LongVideoBench [262], and CG-Bench [7], covering diverse tasks from short video comprehension to long video reasoning.

As shown in Table 12, InternVL 2.5 achieves consistent improvements over InternVL 2.0 across all benchmarks. For example, our smallest model, InternVL2.5-1B improves Video-MME scores from 42.9/45.4 to 50.3/52.3 and MVBench from 57.5 to 64.3. Moreover, we find that *InternVL 2.5 demonstrates better scalability when handling increasing input frames compared to its predecessor*, as shown in Figure 10. We attribute these improvements to two key enhancements: (1) The inclusion of more high-quality video data, which has significantly enhanced the model's video understanding capabilities. (2) Adjusting the training frame sampling strategy from 4–24 to 8–32 frames (as shown in Figure 5(c)) enhanced the model's ability to process richer video information.

Dataset	Settings	InternLM2-1.8B-Chat	InternVL2-2B	InternLM2.5-1.8B-Chat	InternVL2.5-2B	InternLM2.5-7B-Chat	InternVL2-8B	InternVL2.5-8B	InternLM2-20B-Chat	InternVL2-26B	InternLM2.5-20B-Chat	InternVL2.5-26B
MMLU	5-shot	47.3	46.4	50.5	52.6	72.8	73.2	74.6	66.5	68.2	73.3	76.6
CMMLU	5-shot	46.1	47.1	62.7	57.0	78.2	79.2	78.7	64.7	68.1	79.4	81.9
C-Eval	5-shot	48.6	48.6	60.4	56.2	77.9	80.1	79.7	61.8	67.7	80.2	83.8
GAOKAO	0-shot	33.1	32.3	54.7	52.6	78.7	75.0	77.3	63.5	62.3	81.0	86.9
TriviaQA	0-shot	37.3	31.5	32.3	31.2	64.0	62.0	63.4	61.8	61.8	67.3	69.0
NaturalQuestions	0-shot	15.3	13.2	10.1	11.8	21.1	28.1	29.4	23.6	28.8	21.3	36.1
C3	0-shot	75.8	76.9	61.4	78.0	88.1	94.2	94.7	92.2	93.2	94.0	95.8
RACE-High	0-shot	74.0	72.6	78.5	77.4	90.5	90.8	90.8	86.2	86.5	91.3	92.2
WinoGrande	0-shot	56.5	58.7	56.9	59.1	84.9	85.9	83.5	76.4	79.9	86.4	87.9
HellaSwag	0-shot	57.9	53.7	76.2	68.2	94.8	94.9	94.1	85.3	87.5	95.9	95.8
BBH	0-shot	37.9	36.3	43.4	40.9	73.1	72.7	73.4	70.1	69.8	78.4	78.9
GSM8K	4-shot	42.7	40.7	53.3	55.1	85.1	75.6	77.8	80.7	80.0	88.5	82.9
MATH	4-shot	11.0	7.0	39.5	33.5	60.6	39.5	49.9	34.9	35.5	54.7	53.7
TheoremQA	0-shot	13.9	12.3	11.4	12.0	23.4	15.6	23.8	22.1	15.3	23.9	15.4
HumanEval	4-shot	34.8	32.3	41.5	52.4	74.4	69.5	75.0	71.3	67.1	69.5	68.9
MBPP	3-shot	40.9	33.1	42.8	50.6	63.0	58.8	68.5	70.8	66.2	70.0	72.0
MBPP-CN	0-shot	28.2	23.4	33.8	34.2	51.6	48.2	55.2	55.8	54.2	61.0	61.6
Average	–	41.3	39.2	47.6	48.4	69.5	67.2	70.0	64.0	64.2	71.5	72.9
Gain	–	–	(-2.1)	–	(+0.8)	–	(-2.3)	(+0.5)	–	(+0.2)	–	(+1.4)

Table 13: **Comparison of language capabilities across multiple benchmarks.** These results were obtained using the OpenCompass toolkit for testing. Training InternVL 2.0 models led to a decline in pure language capabilities. InternVL 2.5 addresses this by collecting more high-quality open-source data and filtering out low-quality data, achieving better preservation of pure language performance.

Consequently, while InternVL 2.0 models typically perform best at 16 or 32 frames but degrade with more input frames, InternVL 2.5 could benefit from increasing input frames, showing better scalability for long video understanding.

Our largest model, InternVL2.5-78B, achieves leading performance among open-source models and approaches the performance of closed-source systems. Compared to open-source models, InternVL2.5-78B surpasses Qwen2-VL-72B on MVBench (76.4 *vs.* 73.6) and MMBench-Video (1.97 *vs.* 1.70), though its Video-MME score with subtitles is slightly lower (74.0 *vs.* 77.8). Against closed-source models like GPT-4o [192] and Gemini-1.5-Pro [138], InternVL2.5-78B demonstrates competitive performance. On Video-MME, it scores 72.1/74.0, closely matching GPT-4o (71.9/77.2) and Gemini-1.5-Pro (75.0/81.3). However, on LongVideoBench, it achieves 63.6, slightly trailing Gemini-1.5-Pro (64.0) and GPT-4o (66.7). This highlights the remaining challenges in long video understanding for open-source models, indicating room for further improvement.

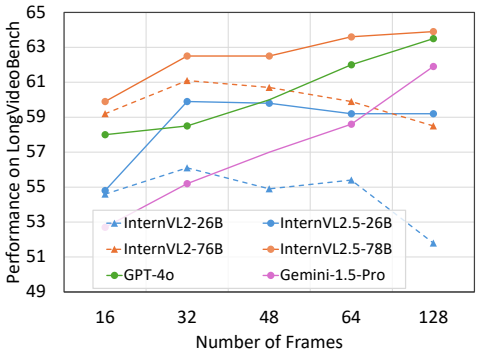


Figure 10: Performance on LongVideoBench with varying input video frames.

6 Evaluation on Language Capability

To thoroughly assess the language capabilities of LLMs and MLLMs, we evaluate their performance across five core dimensions using a diverse set of datasets. These benchmarks encompass tasks like comprehensive examination, language and knowledge, reasoning, mathematics, and coding.

6.1 Benchmarks

Comprehensive Examination. We conduct a thorough evaluation of LLMs and MLLMs using various exam-related datasets: (1) **MMLU** [85] includes 57 subtasks covering diverse topics such as humanities, social sciences, and STEM, evaluated with a 5-shot approach. (2) **CMMLU** [127], focused on a Chinese context, features 67 subtasks spanning general and Chinese-specific domains, also tested in a 5-shot setting. (3) **C-Eval** [96] contains 52 subtasks across four difficulty levels, evaluated in a 5-shot setting. (4) **GAOKAO-Bench** [304], derived from Chinese college entrance exams, offers comprehensive coverage of both subjective and objective question types, with objective questions evaluated in a 0-shot setting.

Language and Knowledge. For language and knowledge-based assessments, we use a range of datasets designed to test the capabilities: (1) **TriviaQA** [103], which includes both reading comprehension and open-domain QA tasks with multiple answers per question, evaluated in a 0-shot setting. (2) **NaturalQuestions** [117], featuring user-generated questions validated by experts, also evaluated in a 0-shot manner. (3) **C3** [219], a free-form multiple-choice Chinese machine reading comprehension dataset, with 0-shot results reported. (4) **RACE** [118], a reading comprehension dataset containing English exam questions for Chinese middle and high school students aged 12 to 18, with results reported for the high school subset in a 0-shot setting.

Reasoning. To measure reasoning capabilities, we use datasets like (1) **WinoGrande** [202], which tests commonsense reasoning through 44,000 multiple-choice questions requiring pronoun disambiguation, evaluated in a 0-shot setting. (2) **HellaSwag** [292] challenges models with natural language inference scenarios and four outcome options, demanding selection of the most logical conclusion, also evaluated in a 0-shot manner. (3) **BigBench Hard (BBH)** [224] comprises 23 tasks specifically chosen for their difficulty in surpassing human performance, further evaluating reasoning depth, with 0-shot results reported.

Mathematics. In the domain of mathematics, (1) **GSM8K-Test** [43] offers approximately 1,300 elementary-level situational problems, evaluated in a 4-shot setting. (2) **MATH** [86] presents 12,500 high school competition-level problems across subjects like algebra and calculus, each with detailed solutions, also evaluated in a 4-shot manner. (3) **TheoremQA** [33] introduces 800 STEM-focused problems requiring theorem application in fields like mathematics, physics, and finance, with 0-shot results reported.

Coding. To evaluate coding capabilities, we employ the following benchmarks: (1) **HumanEval** [31]: This benchmark includes 164 Python programming tasks, each paired with detailed specifications, serving as a standard for assessing coding performance. It is evaluated in a 4-shot setting. (2) **MBPP** [9]: Comprising 974 entry-level programming tasks, MBPP covers a wide range of challenges, from simple arithmetic problems to more complex sequence definitions, evaluated in a 3-shot setting. (3) **MBPP-CN** [9]: A Chinese adaptation of MBPP designed to assess multilingual programming capabilities. This extension broadens the evaluation scope to include linguistic and contextual diversity, with 0-shot results reported.

6.2 Evaluation Results

In the development of MLLMs, maintaining strong pure language capabilities remains critically important. Following the approach of InternLM2 [19], we conducted a comprehensive evaluation of our models' performance across 17 pure language benchmarks using the OpenCompass toolkit [46]. These benchmarks are categorized into five major groups, providing a thorough assessment of the models' pure language abilities.

The results show that InternVL 2.0 demonstrates a slight decline in pure language performance compared to its foundational LLM counterparts. For example, InternVL2-2B achieved an average score of 39.2, a decrease of 2.1 points compared to InternLM2-1.8B-Chat. Similarly, InternVL2-8B scored an average of 67.2, 2.3 points lower than InternLM2.5-7B-Chat.

To address this issue, we curated a large collection of high-quality open-source pure language instruction data and applied rigorous filtering pipelines to eliminate low-quality samples, thereby enhancing the overall data quality. These improvements in InternVL 2.5 have effectively mitigated the decline in language performance, enabling the model to match or even surpass the original LLM in several tasks. This demonstrates that supplementing and optimizing with high-quality language data can not only preserve MLLM's pure language capabilities but also establish a stronger foundation for multimodal tasks.

7 Evaluation on Vision Capability

In this section, we present a comprehensive evaluation of the vision encoder's performance across various domains and tasks. The evaluation is divided into two key categories: (1) *image classification*, representing

global-view semantic quality, and (2) *semantic segmentation*, capturing local-view semantic quality. This approach allows us to assess the representation quality of InternViT across its successive version updates.

7.1 Image Classification

7.1.1 Benchmarks

We assess the global-view semantic quality of InternViT through a comprehensive evaluation on diverse image classification datasets.

ImageNet-1K [56]: A widely-used large-scale dataset containing over 1 million images across 1,000 classes, commonly used for benchmarking image classification models.

ImageNet-ReaL [16]: A re-labeled version of ImageNet's validation set, providing multi-label annotations that are more accurate and robust, following an enhanced labeling protocol.

ImageNet-V2 [199]: A dataset designed to evaluate the robustness of models trained on ImageNet-1K, featuring new test images collected using the original ImageNet methodology.

ImageNet-A [87]: A challenging dataset of naturally occurring, unmodified images that are often misclassified by ResNet models. It highlights the limitations of models when exposed to adversarially difficult examples in real-world settings.

ImageNet-R [84]: A rendition dataset with 30K images across 200 ImageNet classes, composed of art, sketches, toys, sculptures, and other creative representations. It assesses the robustness of models in recognizing abstract renditions of common objects.

ImageNet-Sketch [242]: This dataset contains 51K sketch images, with approximately 50 sketches per ImageNet class. It is constructed via Google Image queries using the class name followed by "sketch of," testing a model's ability to generalize to abstract, hand-drawn representations.

7.1.2 Settings

In this study, two evaluation methods, linear probing [32] and attention pooling probing, are employed to assess the performance of the InternViT models:

- **Linear Probing** [32]: This method involves freezing the pre-trained model and training only a linear classifier on top. It evaluates the quality of the learned features without updating the backbone, providing insights into how effectively the pre-trained model captures semantic information usable by a simple linear classifier in downstream tasks like image classification.
- **Attention Pooling Probing**: In contrast, attention pooling probing evaluates the model by adding an attention pooling layer on top of the frozen features. This approach allows the vision encoder to retain richer information in the final layer, as attention pooling can dynamically select task-relevant features for classification without interference from unrelated information.

For both experiments, we use ImageNet-1K [56] as the training set and evaluate the models on the ImageNet-1K validation set along with several ImageNet variants (*i.e.*, ImageNet-ReaL [16], ImageNet-V2 [199], ImageNet-A [87], ImageNet-R [84], and ImageNet-Sketch [242]) to benchmark their domain generalization capabilities.

The models are trained using SGD as the optimizer, with a peak learning rate of 0.2, a momentum of 0.9, and no weight decay. A cosine learning rate decay schedule is applied over 10 training epochs, with 1 warmup epoch. We use input resolutions of 448×448, with a patch size of 14 and a total batch size of 1024. Data augmentation techniques, such as random resized cropping and horizontal flipping, are employed during training. The code and logs of these classification experiments will be released on our GitHub repository[1].

7.1.3 Evaluation Results

As shown in Table 14, the results reveal an interesting trend across the version updates of InternViT: as the model progresses, the performance of linear probing declines substantially, with all versions showing an average below the gray baseline. In contrast, attention pooling probing consistently outperforms the gray baseline despite some fluctuations. This results in a growing trend in the average score difference (from 3.5 to 6.7), denoted as Δ, across successive InternViT versions.

[1] https://github.com/OpenGVLab/InternVL/tree/main/classification

Model Name	res.	Linear Probing							Attention Pooling Probing							Δ
		IN-1K	IN-ReaL	IN-V2	IN-A	IN-R	IN-Ske	avg.	IN-1K	IN-ReaL	IN-V2	IN-A	IN-R	IN-Ske	avg.	
InternViT-6B-224px	224	88.2	90.4	79.9	77.5	89.8	69.1	82.5	89.2	91.1	82.3	84.7	93.1	72.7	85.5	3.0
InternViT-6B-224px	448	87.8	90.2	79.8	77.2	87.1	65.8	81.3	88.8	91.0	82.0	85.4	91.3	70.5	84.8	3.5
InternViT-6B-448px-V1.0	448	87.0	90.0	78.8	77.2	85.5	65.1	80.6	88.7	91.0	82.0	88.7	92.8	72.0	85.9	5.3
InternViT-6B-448px-V1.2	448	87.0	89.9	78.5	77.1	83.9	59.7	79.4	88.6	91.1	82.0	88.7	92.7	71.6	85.8	6.4
InternViT-6B-448px-V1.5	448	86.5	89.9	78.1	69.8	82.9	60.1	77.9	88.4	91.2	81.6	86.0	92.2	70.9	85.1	7.2
InternViT-6B-448px-V2.5	448	86.6	90.1	77.8	73.7	82.7	60.0	78.5	88.3	91.2	81.3	86.9	92.4	70.8	85.2	6.7

Table 14: **Image classification performance across different versions of InternViT.** We use IN-1K [56] for training and evaluate on the IN-1K validation set as well as multiple ImageNet variants, including IN-ReaL [16], IN-V2 [199], IN-A [87], IN-R [84], and IN-Sketch [242]. Results are reported for both linear probing and attention pooling probing methods, with average accuracy for each method. Δ represents the performance gap between attention pooling probing and linear probing, where a larger Δ suggests a shift from learning simple linear features to capturing more complex, nonlinear semantic representations.

This suggests that features in the model's final layer become less linearly separable, likely as representations evolve to capture more complex, open-ended semantic information. The attention pooling mechanism effectively selects relevant features from this enriched representation space, offsetting challenges from reduced linear separability. Additionally, these findings imply that InternViT maintains key pre-training attributes through iterative updates without catastrophic forgetting. With each version, its representations grow more diverse, capturing open-set semantics and enhancing generalization—an advantage particularly valuable for MLLMs requiring high abstraction for real-world tasks.

7.2 Semantic Segmentation

7.2.1 Benchmarks

We evaluate the local-view semantic quality of InternViT using two representative semantic segmentation datasets, ADE20K and COCO-Stuff-164K.

ADE20K [313]: A comprehensive dataset containing over 20,000 images with annotations across 150 object and background categories, widely used for scene parsing. It provides detailed pixel-level labels for both objects and parts, facilitating a range of fine-grained segmentation tasks.

COCO-Stuff-164K [18]: An extension of the original COCO images with pixel-level annotations, adding 91 "stuff" classes (like grass and sky) to 80 "thing" categories (like people and cars), covering a total of 172 classes. With these comprehensive labels, the dataset supports tasks in scene parsing and semantic segmentation, enabling richer context understanding in image analysis.

7.2.2 Settings

In this study, three evaluation methods—linear probing, head tuning, and full tuning—are employed to assess the performance of the InternViT models on semantic segmentation tasks:

- **Linear Probing**: Linear probing applies a frozen backbone with a linear segmentation head, offering insight into the linear separability of learned features. This method provides a baseline for evaluating pixel-level semantic information with minimal adaptation, though it may not fully capture the encoder's capacity for complex features.
- **Head Tuning**: In head tuning, the InternViT is frozen while the UperNet [264] head remains trainable, allowing the model to utilize a stronger head to reduce its dependence on linear separability. This setup mitigates the decline in linear separability caused by the complex, open-ended features, enabling a more precise evaluation of the vision encoder's capabilities.
- **Full Tuning**: Full tuning involves making both the InternViT backbone and the UperNet [264] segmentation head trainable, allowing the model to adapt all layers for the target task and minimizing reliance on pre-existing linear separability. This setup provides an alternative perspective for evaluating the vision encoder's capacity to extract visual features.

We use AdamW [161] with a peak learning rate of 4e-5 and a polynomial decay schedule. Layer-wise learning rate decay (0.95) is applied in full tuning. Weight decay is set to 0.05 for both head and full tuning, and none for

Model Name	Linear Probing			Head Tuning (UperNet)			Full Tuning (UperNet)			Δ_1	Δ_2
	ADE20K	COCO	avg.	ADE20K	COCO	avg.	ADE20K	COCO	avg.		
InternViT-6B-224px	47.2	42.8	45.0	54.9	48.9	51.9	58.9	51.6	55.3	6.9	10.2
InternViT-6B-448px-V1.0	43.6	38.5	41.0	55.4	49.4	52.4	58.1	51.7	54.9	11.3	13.9
InternViT-6B-448px-V1.2	40.7	36.1	38.4	55.2	48.8	52.0	58.8	51.7	55.2	13.6	16.8
InternViT-6B-448px-V1.5	40.9	36.3	38.6	55.0	49.1	52.0	58.8	51.5	55.2	13.4	16.6
InternViT-6B-448px-V2.5	39.4	35.6	37.5	55.4	49.7	52.6	58.6	51.8	55.2	15.1	17.7

Table 15: **Semantic segmentation performance across different versions of InternViT.** The models are evaluated on ADE20K [313] and COCO-Stuff-164K [18] using three configurations: linear probing, head tuning, and full tuning. The table shows the mIoU scores for each configuration and their averages. Δ_1 represents the gap between head tuning and linear probing, while Δ_2 shows the gap between full tuning and linear probing. A larger Δ value indicates a shift from simple linear features to more complex, nonlinear representations.

linear probing. The input resolution is 504×504, with a patch size of 14 and a batch size of 16. Training consists of 1.5K warmup iterations and 80K total iterations. A drop path rate of 0.4 is applied in full tuning. We utilize default data augmentation from MMSegmentation [45]. All the code and logs related to these experiments will be released on GitHub[2].

7.2.3 Evaluation Results

As shown in Table 15, the semantic segmentation performance of InternViT models is evaluated across three configurations—linear probing, head tuning, and full tuning—on ADE20K [313] and COCO-Stuff-164K [18]. The results reveal distinct trends in how the models' feature representations evolve across version updates.

Linear probing results show a decline in mIoU scores as the model versions progress, with average scores dropping from 45.0 in InternViT-6B-224px to 37.5 in InternViT-6B-448px-V2.5. This indicates that as InternViT updates, the features become less linearly separable, reflecting a shift toward capturing more complex and open-ended information.

In head tuning, the models display a different trend compared to linear probing. All other versions of InternViT surpass the baseline InternViT-6B-224px's mIoU score of 51.9, showing no performance decline. This leads to increasing Δ_1 values, growing from 6.9 in InternViT-6B-224px to 15.1 in InternViT-6B-448px-V2.5. The rise in Δ_1 suggests that while the features become less linearly separable, their quality remains intact, effectively capturing complex information. Similarly, full tuning yields consistent results, as seen in the Δ_2 values. The increase in Δ_2 from 10.2 in InternViT-6B-224px to 17.7 in InternViT-6B-448px-V2.5 further supports this trend.

Overall, the increasing values of Δ_1 and Δ_2 across model versions highlight the shift from simple, linearly separable features to more complex, nonlinear representations. This evolution aligns with InternViT's growing capability to extract visual information as its versions progress within the development of InternVL. It demonstrates the effectiveness of our ViT incremental learning strategy in enhancing the vision encoder's ability to extract open-ended features.

8 Conclusion

In this work, we introduce InternVL 2.5, an advanced open-source multimodal large language model (MLLM) series that builds upon the architecture of InternVL 2.0 with significant improvements in training, testing strategies, and data quality. We systematically explore the relationship between model scaling and performance, analyzing vision encoders, language models, dataset sizes, and test-time configurations. Extensive evaluations on diverse benchmarks demonstrate that InternVL 2.5 achieves competitive performance across tasks such as multi-discipline reasoning, document understanding, video understanding, multilingual processing, *etc*. Notably, it is the first open-source MLLM to surpass 70% on the MMMU benchmark, narrowing the gap between open-source and commercial models like OpenAI o1. By sharing InternVL 2.5 with the community, we hope to contribute a powerful tool for advancing multimodal AI research and applications, and we look forward to seeing future developments building upon this work.

[2]https://github.com/OpenGVLab/InternVL-MMDetSeg

Acknowledgement

This work is supported by the National Key R&D Program of China (No. 2022ZD0160102, 2022ZD0161300), the National Natural Science Foundation of China (No. 62376134, 62372223, U24A20330), the China Mobile Zijin Innovation Institute (No. NR2310J7M), and the Youth PhD Student Research Project under the National Natural Science Foundation (No. 623B2050).

References

[1] Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. 4, 14, 16, 18, 21

[2] Manoj Acharya, Kushal Kafle, and Christopher Kanan. Tallyqa: Answering complex counting questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8076–8084, 2019. 12, 13

[3] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 24

[4] AgentSea. Wave-ui. `https://huggingface.co/datasets/agentsea/wave-ui`, 2024. 13

[5] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 1

[6] Aida Amini, Saadia Gabriel, Peter Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. *arXiv preprint arXiv:1905.13319*, 2019. 13

[7] Anonymous. CG-bench: Clue-grounded question answering benchmark for long video understanding. In *Submitted to The Thirteenth International Conference on Learning Representations*, 2024. under review. 24

[8] Anthropic. The claude 3 model family: Opus, sonnet, haiku. `https://www.anthropic.com`, 2024. 1, 2, 14, 15, 16, 17, 18, 19, 20, 21

[9] Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021. 26

[10] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19129–19139, 2022. 13

[11] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 4

[12] Chongyang Bai, Xiaoxue Zang, Ying Xu, Srinivas Sunkara, Abhinav Rastogi, Jindong Chen, et al. Uibert: Learning generic multimodal representations for ui understanding. *arXiv preprint arXiv:2107.13731*, 2021. 13

[13] 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 preprint arXiv:2308.12966*, 2023. 2

[14] Ltd. Beijing Anjie Zhihe Technology Co. Chinese-ocr. `https://huggingface.co/datasets/longmaodata/Chinese-OCR`, 2024. 12, 13

[15] Asma Ben Abacha, Sadid A Hasan, Vivek V Datla, Dina Demner-Fushman, and Henning Müller. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In *Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes*, 2019. 13

[16] Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020. 27, 28

[17] 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. 12, 13

[18] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1209–1218, 2018. 28, 29

[19] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*, 2024. 1, 3, 4, 5, 13, 26

[20] Jie Cao and Jing Xiao. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1511–1520, 2022. 12, 13

[21] CarperAI. openai summarize tldr dataset. `https://huggingface.co/datasets/CarperAI/openai_summarize_tldr`, 2023. 13

[22] Yuxiang Chai, Siyuan Huang, Yazhe Niu, Han Xiao, Liang Liu, Dingyu Zhang, Peng Gao, Shuai Ren, and Hongsheng Li. Amex: Android multi-annotation expo dataset for mobile gui agents. *arXiv preprint arXiv:2407.17490*, 2024. 13

[23] Shuaichen Chang, David Palzer, Jialin Li, Eric Fosler-Lussier, and Ningchuan Xiao. Mapqa: A dataset for question answering on choropleth maps. *arXiv preprint arXiv:2211.08545*, 2022. 12, 13

[24] Dongping Chen, Yue Huang, Siyuan Wu, Jingyu Tang, Liuyi Chen, Yilin Bai, Zhigang He, Chenlong Wang, Huichi Zhou, Yiqiang Li, et al. Gui-world: A dataset for gui-oriented multimodal llm-based agents. *arXiv preprint arXiv:2406.10819*, 2024. 13

[25] Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. *arXiv preprint arXiv:2402.11684*, 2024. 1, 12, 13

[26] Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. Unigeo: Unifying geometry logical reasoning via reformulating mathematical expression. *arXiv preprint arXiv:2212.02746*, 2022. 12, 13

[27] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 12, 13, 22

[28] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024. 20, 21

[29] 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. 12

[30] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. Sharegpt4video: Improving video understanding and generation with better captions. *arXiv preprint arXiv:2406.04325*, 2024. 12, 13

[31] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021. 26

[32] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *The International Conference on Learning Representations*, pages 1597–1607. PMLR, 2020. 27

[33] Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. Theoremqa: A theorem-driven question answering dataset. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 7889–7901. Association for Computational Linguistics, 2023. 26

[34] Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. Longlora: Efficient fine-tuning of long-context large language models. *arXiv preprint arXiv:2309.12307*, 2023. 13

[35] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 1, 2, 3, 4, 5, 12, 13, 14, 16, 18, 21, 22, 23, 24

[36] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 1, 2, 3, 7, 13

[37] Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Yantao Li, Jianbing Zhang, and Zhiyong Wu. Seeclick: Harnessing gui grounding for advanced visual gui agents. *arXiv preprint arXiv:2401.10935*, 2024. 13

[38] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 24

[39] Zewen Chi, Heyan Huang, Heng-Da Xu, Houjin Yu, Wanxuan Yin, and Xian-Ling Mao. Complicated table structure recognition. *arXiv preprint arXiv:1908.04729*, 2019. 12, 13

[40] 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 *International Conference on Document Analysis and Recognition*, pages 1571–1576, 2019. 12, 13

[41] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024. 13

[42] Christopher Clark and Matt Gardner. Simple and effective multi-paragraph reading comprehension. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 845–855, 2018. 13, 17

[43] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. 13, 26

[44] Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world's first truly open instruction-tuned llm. *Company Blog of Databricks*, 2023. 13

[45] MMSegmentation Contributors. Mmsegmentation: Openmmlab semantic segmentation toolbox and benchmark. `https://github.com/open-mmlab/mmsegmentation`, 2020. 29

[46] OpenCompass Contributors. Opencompass: A universal evaluation platform for foundation models. `https://github.com/open-compass/opencompass`, 2023. 14, 16, 18, 21, 26

[47] X.AI Corp. Grok-1.5 vision preview: Connecting the digital and physical worlds with our first multimodal model. `https://x.ai/blog/grok-1.5v`, 2024. 18, 19

[48] Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with high-quality feedback. *arXiv preprint arXiv:2310.01377*, 2023. 13

[49] Dawei Dai, YuTang Li, YingGe Liu, Mingming Jia, Zhang YuanHui, and Guoyin Wang. 15m multimodal facial image-text dataset. *arXiv preprint arXiv:2407.08515*, 2024. 12

[50] Wenliang Dai, Nayeon Lee, Boxin Wang, Zhuolin Yang, Zihan Liu, Jon Barker, Tuomas Rintamaki, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nvlm: Open frontier-class multimodal llms. *arXiv preprint arXiv:2409.11402*, 2024. 14, 15, 16

[51] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 326–335, 2017. 12

[52] Brian Davis, Bryan Morse, Scott Cohen, Brian Price, and Chris Tensmeyer. Deep visual template-free form parsing. In *International Conference on Document Analysis and Recognition*, pages 134–141, 2019. 12, 13

[53] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *International Conference on Machine Learning*, pages 7480–7512, 2023. 3, 4

[54] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024. 1, 14, 16

[55] Biplab Deka, Zifeng Huang, Chad Franzen, Joshua Hibschman, Daniel Afergan, Yang Li, Jeffrey Nichols, and Ranjitha Kumar. Rico: A mobile app dataset for building data-driven design applications. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*, pages 845–854, 2017. 13

[56] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 27, 28

[57] Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems*, 36, 2024. 13

[58] Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*, 2023. 13

[59] Khang T Doan, Bao G Huynh, Dung T Hoang, Thuc D Pham, Nhat H Pham, Quan Nguyen, Bang Q Vo, and Suong N Hoang. Vintern-1b: An efficient multimodal large language model for vietnamese. *arXiv preprint arXiv:2408.12480*, 2024. 13

[60] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Songyang Zhang, Haodong Duan, Wenwei Zhang, Yining Li, et al. Internlm-xcomposer2-4khd: A pioneering large vision-language model handling resolutions from 336 pixels to 4k hd. *arXiv preprint arXiv:2404.06512*, 2024. 1

[61] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *The International Conference on Learning Representations*, 2020. 3

[62] 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 Annual Meeting of the Association for Computational Linguistics*, pages 320–335, 2022. 1

[63] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11198–11201, 2024. 14

[64] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 1, 4, 16

[65] Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *arXiv preprint arXiv:2406.14515*, 2024. 23, 24

[66] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369, 2023. 1

[67] Miquel Farré, Andi Marafioti, Lewis Tunstall, Leandro Von Werra, and Thomas Wolf. Finevideo. `https://huggingface.co/datasets/HuggingFaceFV/finevideo`, 2024. 13

[68] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 20, 21

[69] Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024. 23, 24

[70] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. *arXiv preprint arXiv:2404.12390*, 2024. 17, 18

[71] Zhangwei Gao, Zhe Chen, Erfei Cui, Yiming Ren, Weiyun Wang, Jinguo Zhu, Hao Tian, Shenglong Ye, Junjun He, Xizhou Zhu, et al. Mini-internvl: A flexible-transfer pocket multimodal model with 5% parameters and 90% performance. *arXiv preprint arXiv:2410.16261*, 2024. 2

[72] Alba Garcia Seco De Herrera, Henning Müller, and Stefano Bromuri. Overview of the imageclef 2015 medical classification task. In *Working Notes of CLEF 2015–Cross Language Evaluation Forum, CEUR*, volume 1391. CEUR Workshop Proceedings, 2015. 12, 13

[73] GlaiveAI. Glaive code assistant v3 dataset. `https://huggingface.co/datasets/glaiveai/glaive-code-assistant-v3`, 2024. 13

[74] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017. 12, 13

[75] Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Niu Minzhe, Xiaodan Liang, Lewei Yao, Runhui Huang, Wei Zhang, Xin Jiang, et al. Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. *Advances in Neural Information Processing Systems*, 35:26418–26431, 2022. 12

[76] Shuhao Gu, Jialing Zhang, Siyuan Zhou, Kevin Yu, Zhaohu Xing, Liangdong Wang, Zhou Cao, Jintao Jia, Zhuoyi Zhang, Yixuan Wang, et al. Infinity-mm: Scaling multimodal performance with large-scale and high-quality instruction data. *arXiv preprint arXiv:2410.18558*, 2024. 14, 16

[77] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: An advanced diagnostic suite for entangled language hallucination & visual illusion in large vision-language models. *arXiv preprint arXiv:2310.14566*, 2023. 20, 21

[78] He Guo, Xiameng Qin, Jiaming Liu, Junyu Han, Jingtuo Liu, and Errui Ding. Eaten: Entity-aware attention for single shot visual text extraction. In *International Conference on Document Analysis and Recognition*, pages 254–259, 2019. 12, 13

[79] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data for text localisation in natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2315–2324, 2016. 12

[80] Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*, 2024. 2, 14, 15, 16

[81] Conghui He, Zhenjiang Jin, Chao Xu, Jiantao Qiu, Bin Wang, Wei Li, Hang Yan, Jiaqi Wang, and Dahua Lin. Wanjuan: A comprehensive multimodal dataset for advancing english and chinese large models. *arXiv preprint arXiv:2308.10755*, 2023. 12

[82] Mengchao He, Yuliang Liu, Zhibo Yang, Sheng Zhang, Canjie Luo, Feiyu Gao, Qi Zheng, Yongpan Wang, Xin Zhang, and Lianwen Jin. Icpr2018 contest on robust reading for multi-type web images. In *International Conference on Pattern Recognition*, pages 7–12, 2018. 12, 13

[83] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020. 13

[84] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021. 27, 28

[85] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *The International Conference on Learning Representations*, 2020. 26

[86] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In Joaquin Vanschoren and Sai-Kit Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021. 26

[87] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021. 27, 28

[88] Jack Hessel, Ana Marasović, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. Do androids laugh at electric sheep? Humor "understanding" benchmarks from The New Yorker Caption Contest. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2023. 13

[89] Hezarai. Parsynth-ocr-200k. `https://huggingface.co/datasets/hezarai/parsynth-ocr-200k`, 2024. 12

[90] Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. Unnatural instructions: Tuning language models with (almost) no human labor. *arXiv preprint arXiv:2212.09689*, 2022. 13

[91] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, 2020. 12, 13

[92] Yu-Chung Hsiao, Fedir Zubach, Gilles Baechler, Victor Carbune, Jason Lin, Maria Wang, Srinivas Sunkara, Yun Zhu, and Jindong Chen. Screenqa: Large-scale question-answer pairs over mobile app screenshots. *arXiv preprint arXiv:2209.08199*, 2022. 13

[93] 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. 12, 13

[94] Xinyue Hu, L Gu, Q An, M Zhang, L Liu, K Kobayashi, T Harada, R Summers, and Y Zhu. Medical-diff-vqa: a large-scale medical dataset for difference visual question answering on chest x-ray images. *PhysioNet*, 2023. 12, 13

[95] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In *European Conference on Computer Vision*, 2020. 12, 13

[96] Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, et al. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36, 2024. 26

[97] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. Icdar2019 competition on scanned receipt ocr and information extraction. In *International Conference on Document Analysis and Recognition*, pages 1516–1520, 2019. 12, 13

[98] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6700–6709, 2019. 12, 13

[99] Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. Egotaskqa: Understanding human tasks in egocentric videos. *Advances in Neural Information Processing Systems*, 35:3343–3360, 2022. 12, 13

[100] Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*, 2024. 13, 18

[101] Qirui Jiao, Daoyuan Chen, Yilun Huang, Yaliang Li, and Ying Shen. Img-diff: Contrastive data synthesis for multimodal large language models. *arXiv preprint arXiv:2408.04594*, 2024. 13

[102] Jimmycarter. Textocr gpt-4v dataset. `https://huggingface.co/datasets/jimmycarter/textocr-gpt4v`, 2023. 12, 13

[103] Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017. 26

[104] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5648–5656, 2018. 12, 13

[105] Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*, 2017. 12, 13

[106] Raghav Kapoor, Yash Parag Butala, Melisa Russak, Jing Yu Koh, Kiran Kamble, Waseem AlShikh, and Ruslan Salakhutdinov. Omniact: A dataset and benchmark for enabling multimodal generalist autonomous agents for desktop and web. In *European Conference on Computer Vision*, pages 161–178. Springer, 2025. 13

[107] Mehran Kazemi, Hamidreza Alvari, Ankit Anand, Jialin Wu, Xi Chen, and Radu Soricut. Geomverse: A systematic evaluation of large models for geometric reasoning. *arXiv preprint arXiv:2312.12241*, 2023. 12, 13

[108] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 787–798, 2014. 22

[109] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European Conference on Computer Vision*, pages 235–251, 2016. 12, 13, 16

[110] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 4999–5007, 2017. 12, 13

[111] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in Neural Information Processing Systems*, 33:2611–2624, 2020. 12, 13

[112] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pages 498–517. Springer, 2022. 12

[113] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 12

[114] knowrohit07. know saraswati cot dataset. `https://huggingface.co/datasets/knowrohit07/know-saraswati-cot`, 2023. 13

[115] Jianfeng Kuang, Wei Hua, Dingkang Liang, Mingkun Yang, Deqiang Jiang, Bo Ren, and Xiang Bai. Visual information extraction in the wild: practical dataset and end-to-end solution. In *International Conference on Document Analysis and Recognition*, pages 36–53. Springer, 2023. 12, 13

[116] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 128(7):1956–1981, 2020. 12

[117] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019. 26

[118] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*, 2017. 26

[119] LAION. Gpt-4v dataset. `https://huggingface.co/datasets/laion/gpt4v-dataset`, 2023. 13

[120] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5:1–10, 2018. 12, 13

[121] Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions. *arXiv preprint arXiv:2408.12637*, 2024. 12, 13

[122] Hugo Laurençon, Léo Tronchon, and Victor Sanh. Unlocking the conversion of web screenshots into html code with the websight dataset. *arXiv preprint arXiv:2403.09029*, 2024. 12, 13

[123] 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. 12, 13

[124] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 1, 3, 14, 15, 16, 17, 18, 21, 24

[125] Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. *arXiv preprint arXiv:2404.16790*, 2024. 16, 17

[126] Chunyi Li, Jianbo Zhang, Zicheng Zhang, Haoning Wu, Yuan Tian, Wei Sun, Guo Lu, Xiaohong Liu, Xiongkuo Min, Weisi Lin, et al. R-bench: Are your large multimodal model robust to real-world corruptions? *arXiv preprint arXiv:2410.05474*, 2024. 18, 19

[127] Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. Cmmlu: Measuring massive multitask language understanding in chinese. *arXiv preprint arXiv:2306.09212*, 2023. 26

[128] Jia Li, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Huang, Kashif Rasul, Longhui Yu, Albert Q Jiang, Ziju Shen, et al. Numinamath: The largest public dataset in ai4maths with 860k pairs of competition math problems and solutions. *Hugging Face repository*, 2024. 13

[129] Junxian Li, Di Zhang, Xunzhi Wang, Zeying Hao, Jingdi Lei, Qian Tan, Cai Zhou, Wei Liu, Yaotian Yang, Xinrui Xiong, et al. Chemvlm: Exploring the power of multimodal large language models in chemistry area. *arXiv preprint arXiv:2408.07246*, 2024. 12, 13

[130] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 13, 24

[131] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024. 12, 13, 23, 24

[132] Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. *arXiv preprint arXiv:2403.00231*, 2024. 12, 13

[133] Qingyun Li, Zhe Chen, Weiyun Wang, Wenhai Wang, Shenglong Ye, Zhenjiang Jin, Guanzhou Chen, Yinan He, Zhangwei Gao, Erfei Cui, et al. Omnicorpus: An unified multimodal corpus of 10 billion-level images interleaved with text. *arXiv preprint arXiv:2406.08418*, 2024. 12

[134] Tianbin Li, Yanzhou Su, Wei Li, Bin Fu, Zhe Chen, Ziyan Huang, Guoan Wang, Chenglong Ma, Ying Chen, Ming Hu, et al. Gmai-vl & gmai-vl-5.5 m: A large vision-language model and a comprehensive multimodal dataset towards general medical ai. *arXiv preprint arXiv:2411.14522*, 2024. 12, 13

[135] Wei Li, William E Bishop, Alice Li, Christopher Rawles, Folawiyo Campbell-Ajala, Divya Tyamagundlu, and Oriana Riva. On the effects of data scale on ui control agents. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. 13

[136] Yang Li, Gang Li, Luheng He, Jingjie Zheng, Hong Li, and Zhiwei Guan. Widget captioning: Generating natural language description for mobile user interface elements. *arXiv preprint arXiv:2010.04295*, 2020. 13

[137] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4804–4814, 2022. 1

[138] 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. 25

[139] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *The Conference on Empirical Methods in Natural Language Processing*, pages 292–305, 2023. 21

[140] 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. 1

[141] 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. 13

[142] Wing Lian, Guan Wang, Bleys Goodson, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknium". Slimorca: An open dataset of gpt-4 augmented flan reasoning traces, with verification. `https://https://huggingface.co/Open-Orca/SlimOrca`, 2023. 13

[143] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26689–26699, 2024. 1, 14, 15, 16, 24

[144] Adam Dahlgren Lindström and Savitha Sam Abraham. Clevr-math: A dataset for compositional language, visual and mathematical reasoning. *arXiv preprint arXiv:2208.05358*, 2022. 12, 13

[145] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging*, pages 1650–1654. IEEE, 2021. 12, 13

[146] Brian Liu, Xianchao Xu, and Yu Zhang. Offline handwritten chinese text recognition with convolutional neural networks. *arXiv preprint arXiv:2006.15619*, 2020. 12, 13

[147] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023. 12, 13

[148] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023. 12, 13

[149] 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. 12, 13

[150] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 3

[151] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge. https://llava-vl.github.io/blog/2024-01-30-llava-next/, January 2024. 3

[152] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(10):2684–2701, 2020. 12, 13

[153] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2025. 22

[154] Wentao Liu, Qianjun Pan, Yi Zhang, Zhuo Liu, Ji Wu, Jie Zhou, Aimin Zhou, Qin Chen, Bo Jiang, and Liang He. Cmm-math: A chinese multimodal math dataset to evaluate and enhance the mathematics reasoning of large multimodal models. *arXiv preprint arXiv:2409.02834*, 2024. 13

[155] Yangzhou Liu, Yue Cao, Zhangwei Gao, Weiyun Wang, Zhe Chen, Wenhai Wang, Hao Tian, Lewei Lu, Xizhou Zhu, Tong Lu, et al. Mminstruct: A high-quality multi-modal instruction tuning dataset with extensive diversity. *arXiv preprint arXiv:2407.15838*, 2024. 1, 12, 13

[156] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023. 20, 21, 22

[157] Yuan Liu, Zhongyin Zhao, Ziyuan Zhuang, Le Tian, Xiao Zhou, and Jie Zhou. Points: Improving your vision-language model with affordable strategies. *arXiv preprint arXiv:2409.04828*, 2024. 1

[158] Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezhi Peng, Mingyu Liu, Mingrui Chen, Chunyuan Li, Lianwen Jin, et al. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023. 16, 17

[159] Ziyu Liu, Tao Chu, Yuhang Zang, Xilin Wei, Xiaoyi Dong, Pan Zhang, Zijian Liang, Yuanjun Xiong, Yu Qiao, Dahua Lin, et al. Mmdu: A multi-turn multi-image dialog understanding benchmark and instruction-tuning dataset for lvlms. *arXiv preprint arXiv:2406.11833*, 2024. 12

[160] Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution. *arXiv preprint arXiv:2409.12961*, 2024. 24

[161] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 28

[162] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, et al. Deepseek-vl: Towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024. 3

[163] 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. *arXiv preprint arXiv:2310.02255*, 2023. 14, 15, 16

[164] 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. *arXiv preprint arXiv:2105.04165*, 2021. 12, 13

[165] 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. 12, 13

[166] 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. 12, 13

[167] 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. 12

[168] Quanfeng Lu, Wenqi Shao, Zitao Liu, Fanqing Meng, Boxuan Li, Botong Chen, Siyuan Huang, Kaipeng Zhang, Yu Qiao, and Ping Luo. Gui odyssey: A comprehensive dataset for cross-app gui navigation on mobile devices. *arXiv preprint arXiv:2406.08451*, 2024. 13

[169] Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. Ovis: Structural embedding alignment for multimodal large language model. *arXiv preprint arXiv:2405.20797*, 2024. 14, 16

[170] Xudong Lu, Yinghao Chen, Cheng Chen, Hui Tan, Boheng Chen, Yina Xie, Rui Hu, Guanxin Tan, Renshou Wu, Yan Hu, et al. Bluelm-v-3b: Algorithm and system co-design for multimodal large language models on mobile devices. *arXiv preprint arXiv:2411.10640*, 2024. 1

[171] Yujie Lu, Dongfu Jiang, Wenhu Chen, William Yang Wang, Yejin Choi, and Bill Yuchen Lin. Wildvision: Evaluating vision-language models in the wild with human preferences. *arXiv preprint arXiv:2406.11069*, 2024. 13, 18, 19

[172] Gen Luo, Xue Yang, Wenhan Dou, Zhaokai Wang, Jifeng Dai, Yu Qiao, and Xizhou Zhu. Mono-internvl: Pushing the boundaries of monolithic multimodal large language models with endogenous visual pre-training. *arXiv preprint arXiv:2410.08202*, 2024. 1

[173] Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. Wizardcoder: Empowering code large language models with evol-instruct. *arXiv preprint arXiv:2306.08568*, 2023. 13

[174] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Videogpt+: Integrating image and video encoders for enhanced video understanding. *arXiv preprint arXiv:2406.09418*, 2024. 12, 13

[175] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023. 13

[176] Hui Mao, Ming Cheung, and James She. Deepart: Learning joint representations of visual arts. In *Proceedings of the ACM International Conference on Multimedia*, pages 1183–1191, 2017. 12, 13

[177] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11–20, 2016. 13, 22

[178] 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. 12, 13

[179] MarkrAI. Kopen-hq-hermes-2.5-60k dataset. `https://huggingface.co/datasets/MarkrAI/KOpen-HQ-Hermes-2.5-60K`, 2023. 13

[180] U-V Marti and Horst Bunke. The iam-database: an english sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition*, 5:39–46, 2002. 12, 13

[181] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 2263–2279, 2022. 12, 13, 16

[182] Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. Unichart: A universal vision-language pretrained model for chart comprehension and reasoning. *arXiv preprint arXiv:2305.14761*, 2023. 12, 13

[183] 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. 12, 13, 16, 17

[184] 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. 12, 16, 17

[185] 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. 1

[186] Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, Ping Luo, et al. Mmiu: Multimodal multi-image understanding for evaluating large vision-language models. *arXiv preprint arXiv:2408.02718*, 2024. 18

[187] Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1527–1536, 2020. 12, 13

[188] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *International Conference on Document Analysis and Recognition*, pages 947–952, 2019. 12, 13

[189] Arindam Mitra, Hamed Khanpour, Corby Rosset, and Ahmed Awadallah. Orca-math: Unlocking the potential of slms in grade school math. *arXiv preprint arXiv:2402.14830*, 2024. 13

[190] Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. *arXiv preprint arXiv:2407.02371*, 2024. 13

[191] Jason Obeid and Enamul Hoque. Chart-to-text: Generating natural language descriptions for charts by adapting the transformer model. *arXiv preprint arXiv:2010.09142*, 2020. 12, 13

[192] OpenAI. Gpt-4v(ision) system card. `https://cdn.openai.com/papers/GPTV_System_Card.pdf`, 2023. 1, 2, 14, 15, 16, 17, 18, 19, 20, 21, 23, 24, 25

[193] Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36, 2024. 13

[194] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 12

[195] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021. 3, 4

[196] Nazneen Rajani, Lewis Tunstall, Edward Beeching, Nathan Lambert, Alexander M. Rush, and Thomas Wolf. No robots. Hugging Face repository, `https://huggingface.co/datasets/HuggingFaceH4/no_robots`, 2023. 13

[197] Viresh Ranjan, Udbhav Sharma, Thu Nguyen, and Minh Hoai. Learning to count everything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3394–3403, 2021. 13

[198] Christopher Rawles, Alice Li, Daniel Rodriguez, Oriana Riva, and Timothy Lillicrap. Androidinthewild: A large-scale dataset for android device control. *Advances in Neural Information Processing Systems*, 36, 2024. 13

[199] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400, 2019. 27, 28

[200] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 14, 15, 16, 18, 21, 24

[201] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3202–3212, 2015. 12, 13

[202] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8732–8740, 2020. 26

[203] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 7, 12

[204] Christoph Schuhmann, Andreas Köpf, Richard Vencu, Theo Coombes, and Romain Beaumont. Laion coco: 600m synthetic captions from laion2b-en. `https://laion.ai/blog/laion-coco/`, 2022. 12

[205] 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, 2022. 12, 13

[206] Minjoon Seo, Hannaneh Hajishirzi, Ali Farhadi, Oren Etzioni, and Clint Malcolm. Solving geometry problems: Combining text and diagram interpretation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1466–1476, 2015. 12, 13

[207] 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*, volume 33, pages 8876–8884, 2019. 12, 13

[208] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8430–8439, 2019. 12, 13

[209] 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 *International Conference on Document Analysis and Recognition*, volume 1, pages 1429–1434, 2017. 12, 13

[210] Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, De-An Huang, Hongxu Yin, Karan Sapra, Yaser Yacoob, Humphrey Shi, et al. Eagle: Exploring the design space for multimodal llms with mixture of encoders. *arXiv preprint arXiv:2408.15998*, 2024. 1

[211] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *European Conference on Computer Vision*, pages 742–758, 2020. 12, 13

[212] 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. 12, 13, 16, 17

[213] Amanpreet Singh, Guan Pang, Mandy Toh, Jing Huang, Wojciech Galuba, and Tal Hassner. Textocr: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8802–8812, 2021. 12, 13

[214] Shweta Singh, Aayan Yadav, Jitesh Jain, Humphrey Shi, Justin Johnson, and Karan Desai. Benchmarking object detectors with coco: A new path forward. In *European Conference on Computer Vision*, pages 279–295. Springer, 2025. 12, 13

[215] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024. 1

[216] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*, 2018. 13

[217] Hamed Rahimi Sujet AI, Allaa Boutaleb. Sujet-finance-qa-vision-100k: A large-scale dataset for financial document vqa. `https://huggingface.co/datasets/sujet-ai/Sujet-Finance-QA-Vision-100k`, 2024. 13

[218] Hai-Long Sun, Da-Wei Zhou, Yang Li, Shiyin Lu, Chao Yi, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, De-Chuan Zhan, et al. Parrot: Multilingual visual instruction tuning. *arXiv preprint arXiv:2406.02539*, 2024. 22, 23

[219] Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. Investigating prior knowledge for challenging chinese machine reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:141–155, 2020. 26

[220] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative pretraining in multimodality. In *The International Conference on Learning Representations*, 2024. 1

[221] Tianxiang Sun, Xiaotian Zhang, Zhengfu He, Peng Li, Qinyuan Cheng, Hang Yan, Xiangyang Liu, Yunfan Shao, Qiong Tang, Xingjian Zhao, et al. Moss: Training conversational language models from synthetic data. *arXiv preprint arXiv:2307.15020*, 7, 2023. 1, 13

[222] 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 *International Conference on Document Analysis and Recognition*, pages 1557–1562, 2019. 12, 13

[223] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023. 20, 21

[224] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022. 26

[225] Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. Visualmrc: Machine reading comprehension on document images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13878–13888, 2021. 12

[226] Benny J Tang, Angie Boggust, and Arvind Satyanarayan. Vistext: A benchmark for semantically rich chart captioning. *arXiv preprint arXiv:2307.05356*, 2023. 12, 13

[227] Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, Chunhui Lin, Wanqing Li, Mohamad Fitri Faiz Bin Mahmood, Hao Feng, Zhen Zhao, et al. Mtvqa: Benchmarking multilingual text-centric visual question answering. *arXiv preprint arXiv:2405.11985*, 2024. 13, 23

[228] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1, 23

[229] Qwen Team. Qwen2.5: A party of foundation models. `https://qwenlm.github.io/blog/qwen2.5/`, September 2024. 1, 3, 5, 11

[230] Qwen Team. Qwq: Reflect deeply on the boundaries of the unknown. `https://qwenlm.github.io/blog/qwq-32b-preview/`, November 2024. 1

[231] Ryan Teknium, Jeffrey Quesnelle, and Chen Guang. Hermes 3 technical report. *arXiv preprint arXiv:2408.11857*, 2024. 1

[232] Changyao Tian, Xizhou Zhu, Yuwen Xiong, Weiyun Wang, Zhe Chen, Wenhai Wang, Yuntao Chen, Lewei Lu, Tong Lu, Jie Zhou, et al. Mm-interleaved: Interleaved image-text generative modeling via multi-modal feature synchronizer. *arXiv preprint arXiv:2401.10208*, 2024. 1

[233] Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Hierarchical multimodal transformers for multipage docvqa. *Pattern Recognition*, 144:109834, 2023. 13

[234] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024. 1, 12, 13, 14, 16, 18, 21

[235] 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. *arXiv preprint arXiv:2307.09288*, 2023. 1

[236] Dmitry Ustalov, Nikita Pavlichenko, Sergey Koshelev, Daniil Likhobaba, and Alisa Smirnova. Toloka visual question answering benchmark. *arXiv preprint arXiv:2309.16511*, 2023. 12, 13

[237] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8769–8778, 2018. 12, 13

[238] Andreas Veit, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge Belongie. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140*, 2016. 12, 13

[239] Konstantin Verner. Cyrillic handwriting dataset. `https://www.kaggle.com/datasets/constantinwerner/cyrillic-handwriting-dataset`, 2020. 13

[240] Bryan Wang, Gang Li, Xin Zhou, Zhourong Chen, Tovi Grossman, and Yang Li. Screen2words: Automatic mobile ui summarization with multimodal learning. In *The 34th Annual ACM Symposium on User Interface Software and Technology*, pages 498–510, 2021. 12, 13

[241] Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. Muirbench: A comprehensive benchmark for robust multi-image understanding. *arXiv preprint arXiv:2406.09411*, 2024. 18

[242] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019. 27, 28

[243] Jiaqi Wang, Pan Zhang, Tao Chu, Yuhang Cao, Yujie Zhou, Tong Wu, Bin Wang, Conghui He, and Dahua Lin. V3det: Vast vocabulary visual detection dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19844–19854, 2023. 12, 13

[244] 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. 12

[245] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *arXiv preprint arXiv:2402.14804*, 2024. 14, 15

[246] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1, 2, 3, 7, 8, 14, 15, 16, 17, 18, 20, 21, 22, 23, 24

[247] Peng Wang, Shijie Wang, Junyang Lin, Shuai Bai, Xiaohuan Zhou, Jingren Zhou, Xinggang Wang, and Chang Zhou. One-peace: Exploring one general representation model toward unlimited modalities. *arXiv:2305.11172*, 2023. 22

[248] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023. 1, 22

[249] Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, and Jifeng Dai. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*, 2024. 1, 2

[250] Weiyun Wang, Yiming Ren, Haowen Luo, Tiantong Li, Chenxiang Yan, Zhe Chen, Wenhai Wang, Qingyun Li, Lewei Lu, Xizhou Zhu, et al. The all-seeing project v2: Towards general relation comprehension of the open world. *arXiv preprint arXiv:2402.19474*, 2024. 12, 13, 21

[251] Weiyun Wang, Min Shi, Qingyun Li, Wenhai Wang, Zhenhang Huang, Linjie Xing, Zhe Chen, Hao Li, Xizhou Zhu, Zhiguo Cao, et al. The all-seeing project: Towards panoptic visual recognition and understanding of the open world. In *The International Conference on Learning Representations*, 2024. 12

[252] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14408–14419, 2023. 1

[253] Xinyu Wang, Yuliang Liu, Chunhua Shen, Chun Chet Ng, Canjie Luo, Lianwen Jin, Chee Seng Chan, Anton van den Hengel, and Liangwei Wang. On the general value of evidence, and bilingual scene-text visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10126–10135, 2020. 12, 13

[254] Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Gedas Bertasius, Mohit Bansal, et al. Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. *arXiv preprint arXiv:2401.10529*, 2024. 12, 13

[255] Yejie Wang, Keqing He, Dayuan Fu, Zhuoma Gongque, Heyang Xu, Yanxu Chen, Zhexu Wang, Yujia Fu, Guanting Dong, Muxi Diao, et al. How do your code llms perform? empowering code instruction tuning with high-quality data. *arXiv preprint arXiv:2409.03810*, 2024. 13

[256] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, et al. Internvideo2: Scaling video foundation models for multimodal video understanding. *arXiv preprint arXiv:2403.15377*, 2024. 3

[257] Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, et al. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *arXiv preprint arXiv:2406.18521*, 2024. 16, 17

[258] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021. 13

[259] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. *arXiv preprint arXiv:2405.09711*, 2024. 12, 13

[260] Chaoyi Wu. Pmc-casereport. `https://huggingface.co/datasets/chaoyi-wu/PMC-CaseReport`, 2023. 12, 13

[261] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology. *arXiv preprint arXiv:2308.02463*, 2023. 13

[262] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *arXiv preprint arXiv:2407.15754*, 2024. 24

[263] Renqiu Xia, Bo Zhang, Haoyang Peng, Hancheng Ye, Xiangchao Yan, Peng Ye, Botian Shi, Junchi Yan, and Yu Qiao. Structchart: Perception, structuring, reasoning for visual chart understanding. *arXiv preprint arXiv:2309.11268*, 2023. 12, 13

[264] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *European Conference on Computer Vision*, pages 418–434, 2018. 28

[265] Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. Wizardlm: Empowering large pre-trained language models to follow complex instructions. In *The International Conference on Learning Representations*, 2024. 13

[266] Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. *arXiv preprint arXiv:2406.08464*, 2024. 13

[267] B. Yan, Yi Jiang, Jiannan Wu, D. Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 22

[268] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. 4

[269] Dongjie Yang, Suyuan Huang, Chengqiang Lu, Xiaodong Han, Haoxin Zhang, Yan Gao, Yao Hu, and Hai Zhao. Vript: A video is worth thousands of words. *arXiv preprint arXiv:2406.06040*, 2024. 13

[270] Jianxin Yang. Firefly: A chinese conversational large language model. `https://github.com/yangjianxin1/Firefly`, 2023. 13

[271] Jianxin Yang. Longqlora: Efficient and effective method to extend context length of large language models. *arXiv preprint arXiv:2311.04879*, 2023. 13

[272] Wenjuan Yang, Xuhui Zhang, Bing Ma, Yanqun Wang, Yujia Wu, Jianxing Yan, Yongwei Liu, Chao Zhang, Jicheng Wan, Yue Wang, et al. An open dataset for intelligent recognition and classification of abnormal condition in longwall mining. *Scientific Data*, 10(1):416, 2023. 13

[273] Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas Guibas, and Li Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *2011 International Conference on Computer Vision*, pages 1331–1338, 2011. 12

[274] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 14, 16, 18, 21, 24

[275] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv preprint arXiv:2311.04257*, 2023. 1, 23

[276] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*, 2019. 12, 13

[277] Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, Jiayi Lei, Quanfeng Lu, Runjian Chen, Peng Xu, Renrui Zhang, Haozhe Zhang, Peng Gao, Yali Wang, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint arXiv:2404.16006*, 2024. 18

[278] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023. 12

[279] Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*, 2024. 4

[280] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85, 2016. 12, 13

[281] 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. 12, 13

[282] Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*, 2024. 13

[283] 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. 20, 21

[284] Weihao Yu, Zhengyuan Yang, Linfeng Ren, Linjie Li, Jianfeng Wang, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Lijuan Wang, and Xinchao Wang. Mm-vet v2: A challenging benchmark to evaluate large multimodal models for integrated capabilities. *arXiv preprint arXiv:2408.00765*, 2024. 20, 21

[285] Ya-Qi Yu, Minghui Liao, Jiwen Zhang, and Jihao Wu. Texthawk2: A large vision-language model excels in bilingual ocr and grounding with 16x fewer tokens. *arXiv preprint arXiv:2410.05261*, 2024. 22

[286] Yijiong Yu. "paraphrasing the original text" makes high accuracy long-context qa. *arXiv preprint arXiv:2312.11193*, 2023. 13

[287] 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*, 34:509–521, 2019. 12, 13

[288] Ye Yuan, Xiao Liu, Wondimu Dikubab, Hui Liu, Zhilong Ji, Zhongqin Wu, and Xiang Bai. Syntax-aware network for handwritten mathematical expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4553–4562, 2022. 12, 13

[289] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*, 2023. 2, 3, 14, 15

[290] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Ming Yin, Botao Yu, Ge Zhang, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024. 14, 15

[291] Abhay Zala, Jaemin Cho, Satwik Kottur, Xilun Chen, Barlas Oguz, Yashar Mehdad, and Mohit Bansal. Hierarchical video-moment retrieval and step-captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23056–23065, 2023. 13

[292] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, 2019. 26

[293] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023. 1

[294] Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[295] Bo-Wen Zhang, Yan Yan, Lin Li, and Guang Liu. Infinitymath: A scalable instruction tuning dataset in programmatic mathematical reasoning. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 5405–5409, 2024. 13

[296] Haotian Zhang, Mingfei Gao, Zhe Gan, Philipp Dufter, Nina Wenzel, Forrest Huang, Dhruti Shah, Xianzhi Du, Bowen Zhang, Yanghao Li, et al. Mm1.5: Methods, analysis & insights from multimodal llm fine-tuning. *arXiv preprint arXiv:2409.20566*, 2024. 22

[297] Haotian Zhang, Haoxuan You, Philipp Dufter, Bowen Zhang, Chen Chen, Hong-You Chen, Tsu-Jui Fu, William Yang Wang, Shih-Fu Chang, Zhe Gan, et al. Ferret-v2: An improved baseline for referring and grounding with large language models. *arXiv preprint arXiv:2404.07973*, 2024. 22

[298] Jiajie Zhang, Yushi Bai, Xin Lv, Wanjun Gu, Danqing Liu, Minhao Zou, Shulin Cao, Lei Hou, Yuxiao Dong, Ling Feng, and Juanzi Li. Longcite: Enabling llms to generate fine-grained citations in long-context qa. *arXiv preprint arXiv:2409.02897*, 2024. 13

[299] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer, 2025. 14, 15

[300] Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Yichi Zhang, Ziyu Guo, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, Shanghang Zhang, et al. Mavis: Mathematical visual instruction tuning. *arXiv preprint arXiv:2407.08739*, 2024. 12, 13

[301] 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 *International Conference on Document Analysis and Recognition*, pages 1577–1581, 2019. 12, 13

[302] Tianyu Zhang, Suyuchen Wang, Lu Li, Ge Zhang, Perouz Taslakian, Sai Rajeswar, Jie Fu, Bang Liu, and Yoshua Bengio. Vcr: Visual caption restoration. *arXiv preprint arXiv:2406.06462*, 2024. 13, 16, 17

[303] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023. 12, 13

[304] Xiaotian Zhang, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. Evaluating the performance of large language models on gaokao benchmark. *arXiv preprint arXiv:2305.12474*, 2023. 26

[305] 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. 12, 13

[306] Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *arXiv preprint arXiv:2408.13257*, 2024. 18, 19

[307] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. 13

[308] Bingchen Zhao, Yongshuo Zong, Letian Zhang, and Timothy Hospedales. Benchmarking multi-image understanding in vision and language models: Perception, knowledge, reasoning, and multi-hop reasoning. *arXiv preprint arXiv:2406.12742*, 2024. 18

[309] Bo Zhao, Boya Wu, and Tiejun Huang. Svit: Scaling up visual instruction tuning. *arXiv preprint arXiv:2307.04087*, 2023. 12

[310] Mingyu Zheng, Xinwei Feng, Qingyi Si, Qiaoqiao She, Zheng Lin, Wenbin Jiang, and Weiping Wang. Multimodal table understanding. *arXiv preprint arXiv:2406.08100*, 2024. 13

[311] Tianyu Zheng, Ge Zhang, Tianhao Shen, Xueling Liu, Bill Yuchen Lin, Jie Fu, Wenhu Chen, and Xiang Yue. Opencodeinterpreter: Integrating code generation with execution and refinement. *arXiv preprint arXiv:2402.14658*, 2024. 13

[312] Xinyi Zheng, Douglas Burdick, Lucian Popa, Xu Zhong, and Nancy Xin Ru Wang. Global table extractor (gte): A framework for joint table identification and cell structure recognition using visual context. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 697–706, 2021. 12, 13

[313] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 633–641, 2017. 28, 29

[314] Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36, 2024. 13

[315] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024. 24

[316] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *The International Conference on Learning Representations*, 2024. 3

[317] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4995–5004, 2016. 12, 13