Mini-InternVL: A Flexible-Transfer Pocket Multimodal Model with 5% Parameters and 90% Performance

Zhangwei Gao[1,7*] **Zhe Chen**[1,3*] **Erfei Cui**[1,7*] **Yiming Ren**[1,2*] **Weiyun Wang**[1,4*]
Jinguo Zhu[1] **Hao Tian**[6] **Shenglong Ye**[1] **Junjun He**[1] **Xizhou Zhu**[2,1] **Lewei Lu**[6]
Tong Lu[3] **Yu Qiao**[1] **Jifeng Dai**[2,1] **Wenhai Wang**[5,1†]

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

https://github.com/OpenGVLab/InternVL

Abstract

Multimodal large language models (MLLMs) have demonstrated impressive performance in vision-language tasks across a broad spectrum of domains. However, the large model scale and associated high computational costs pose significant challenges for training and deploying MLLMs on consumer-grade GPUs or edge devices, thereby hindering their widespread application. In this work, we introduce Mini-InternVL, a series of MLLMs with parameters ranging from 1B to 4B, which achieves 90% of the performance with only 5% of the parameters. This significant improvement in efficiency and effectiveness makes our models more accessible and applicable in various real-world scenarios. To further promote the adoption of our models, we develop a unified adaptation framework for Mini-InternVL, which enables our models to transfer and outperform specialized models in downstream tasks, including autonomous driving, medical images, and remote sensing. We believe that our study can provide valuable insights and resources to advance the development of efficient and effective MLLMs.

1 Introduction

In recent years, there have been significant advancements in multimodal large language models (MLLMs) [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11], which leverages the powerful capabilities of pre-trained large language models (LLMs) [12, 13, 14, 15, 16, 17] alongside vision foundation models (VFMs) [1, 18, 19]. These models undergo multi-stage training on extensive image-text data, which effectively aligns visual representations from VFMs with the latent space of LLMs, leading to promising performance in general vision-language understanding, reasoning, and interaction tasks. However, *the large computational burden and the poor performance on long-tail domain-specific tasks hinder the widespread application of MLLMs in practical scenarios.*

The emergence of lightweight MLLMs [20, 21, 22, 23, 24, 25] has provided a good balance between parameter size and performance, alleviating the reliance on expensive computing devices and fostering the development of various downstream applications. However, there are still several challenges: (1) Most existing MLLMs use vision encoders like CLIP [18], which are trained on Internet-domain image-text data and are aligned with BERT [26, 27]. *As a result, these vision encoders are not capable*

*Equal contributions
†Corresponding Author

of covering the extensive range of visual domains and are misaligned with LLMs' representations. (2) To adapt MLLMs to specialized domains, existing methods mainly focus on modifying the model architectures, gathering extensive related training data, or customizing the training process for the target domain. *There is still no consensus framework for LLMs' downstream adaptation. Different domains have different model designs, data formats, and training schedules.*

To address these issues, there is a need for a strong vision encoder with comprehensive visual knowledge as well as a general transfer learning paradigm that allows for efficient application across downstream tasks in various domains at a low marginal cost.

In this work, we introduce Mini-InternVL, a series of powerful pocket-sized MLLMs that can be easily transferred to various specialized domains. To this end, we first enhance the representational capabilities of a lightweight vision encoder. We initialize a 300M vision encoder using the weights from CLIP and apply knowledge distillation using InternViT-6B [1] as the teacher model. Subsequently, we develop Mini-InternVL series with 1 billion, 2 billion, and 4 billion parameters, by integrating the vision encoder with the pre-trained LLMs such as Qwen2-0.5B [16], InternLM2-1.8B [14], and Phi-3-Mini [28], respectively. Benefiting from the robust vision encoder, Mini-InternVL exhibits excellent multimodal performance on general multimodal benchmarks like MMBench [29], ChartQA [30], and MathVista [31]. Remarkably, compared with InternVL2-76B, the proposed Mini-InternVL-4B achieves 90% of the performance of larger counterparts while using only 5% of the parameters, significantly reducing computational overhead.

To further adapt our models to specific-domain downstream tasks, we introduce a straightforward yet effective transfer learning paradigm. Within this paradigm, we develop a unified transfer approach applicable to various downstream tasks, including autonomous driving, medical images, and remote sensing. This approach standardizes the model architecture, data format, and training schedule. The results demonstrate the effectiveness of this method in enhancing the models' visual understanding and reasoning capabilities in domain-specific scenarios, enabling them to match the performance of proprietary commercial models within the target domains.

In summary, our contribution has three folds:

(1) We propose Mini-InternVL, a powerful pocket multimodal model, that not only achieves robust multimodal performance with only 4 billion parameters but also easily transfers to downstream tasks across various domains at low marginal cost.

(2) We develop several design features for Mini-InternVL, including a lightweight vision encoder—InternViT-300M, that is robust for various visual domains. Additionally, we introduce a simple but effective paradigm that standardizes model architecture, data format, and training schedule for effective downstream task transfer.

(3) We comprehensively evaluate our models through extensive experiments on general and domain-specific benchmarks. These results show that our multimodal models achieve 90% of the performance using significantly fewer parameters on general multimodal benchmarks. For specific domain tasks, with minimal computational cost for fine-tuning, they can rival closed-source commercial models. We conduct a series of ablation studies to explore the impact of data sample size on domain adaptation, hoping to provide insights into the application of MLLMs in specialized domains.

2 Related Works

Multimodal Large Language Models. Benefiting from the advancement of LLMs, the MLLMs have also achieved great progress. Early works [32, 33, 34] consider multi-modal understanding as one of the tool usage tasks and prompt the LLMs to ask other models to write a caption about the corresponding input modality so that LLMs could understand the multi-modal input. To effectively utilize the ability of pre-trained LLMs and VFMs, a series of works [1, 35, 36, 37, 38, 39, 40] propose to use a connector to align the embedding space between them, which achieve promising performance under a controllable cost. Another series of work [12, 41, 42, 43] extend pre-trained LLMs with extra layers to fuse the vision features, which reduce the number of required visual tokens inputted into LLMs while introducing extra training cost.

Recently, some works, such as Fuyu [44], MoMa [45] and Chameleon [46], propose a vision encoder-free architecture. This type of architecture consists of a single transformer model, which is used to

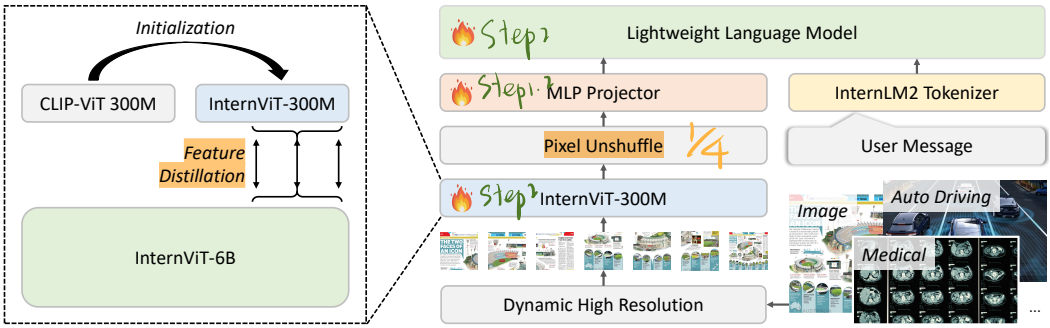


Figure 1: **Training method and architecture of Mini-InternVL.** Left: We employ InternViT-6B [1] as the teacher model to perform knowledge distillation on the student model. Right: Mini-InternVL adopts the ViT-MLP-LLM architecture similar to popular MLLMs [20, 37, 38, 51, 61, 62], combining InternViT-300M with a series of lightweight LLMs through an MLP projector. Here, we employ a simple pixel unshuffle to reduce the number of visual tokens to one-quarter

process both visual and textual information simultaneously without requiring an additional encoder, making it more deployment-friendly. Despite these advancements, the heavy inference cost of these MLLMs hinders their application in downstream tasks. To address such issue, a series of lightweight MLLMs, such as MiniCPM-V [20], are proposed. However, since most of them use CLIP-L [18] as the vision encoder, which is only trained on the natural image domain, these models are limited to the general domain and fail to generalize to other domains. In this work, we propose InternViT-300M, which is distilled from InternViT-6B and trained on a diverse set of image domains.

Vision Foundation Models for MLLMs. From a vision-centric perspective, most MLLMs utilize vision models such as CLIP [18, 47] and SigLIP [19], which are trained on large-scale web image-text data. However, such vision encoders face significant limitations in terms of parameter scale and representational ability. Several studies have explored this issue. For instance, Tong *et al.* [48] identified significant differences in the visual patterns of CLIP and DINOv2 [49], leading to the development of a mixture-of-features module that integrates these two VFMs. LLaVA-HR [50] introduces a dual-branch vision encoder that employs CLIP-ViT for low-resolution pathways and CLIP-ConvNext for high-resolution pathways. Similarly, DeepSeek-VL [51] utilizes a dual vision encoder design, incorporating SigLIP-L for low-resolution images and SAM-B [52] for high-resolution images.

However, these methods involve excessively complex pathways, which complicates the practical application of the models. Moreover, such approaches do not resolve the issue of vision encoders lacking comprehensive visual knowledge across various domains. In contrast, InternViT [1] implements progressive image-text alignment, and acquires representational capabilities across multiple domains by performing generative training on datasets spanning various fields. We propose injecting visual knowledge from the capable vision encoder into lightweight visual models, thus avoiding the computational expense associated with iterative generative pre-training.

Domain-Specialized Adaptation of MLLMs. Several methods have been explored to apply MLLMs to specific domains, such as GeoChat [53] and EarthGPT [54] for remote sensing, LLaVA-Med [55] and Qilin-Med-VL [56] for the medical images, ChemVLM [57] for chemistry, DriveVLM [58], DriveMLM [59] and DriveGPT4 [60] for autonomous driving. Although these methods have achieved promising results, they involve modifications to model architectures, the collection of extensive domain-specific training data, or customization of the training process for the target domain. Nonetheless, there is still no universally accepted framework for the downstream adaptation of MLLMs. we propose a straightforward yet effective transfer learning paradigm, aiming to prevent significant disparities among MLLMs in different fields that hinder interoperability.

3 Method

In this section, we introduce Mini-InternVL, a series of lightweight multimodal large language models (MLLMs). Section 3.1 provides a comprehensive overview of Mini-InternVL. Then, Section 3.2

Table 1: Datasets used in knowledge distillation of vision encoder.

Type	Dataset
Natural images	Laion [63], COYO [64], GRIT [39], COCO [65], LVIS [66], Objects365 [67], Flickr30K [68], VG [69], All-Seeing [61, 62], MMInstruct [70], LRV-Instruction [71]
OCR	TextCaps [72], Wukong-OCR [73], CTW [74], MMC-Inst [75], LSVT [76], ST-VQA [77], RCTW-17 [78], ReCTs [79], ArT [80], SynthDoG [81], LaionCOCO-OCR [82], COCO-Text [83], DocVQA [84], TextOCR [85], LLaVAR [86], TQA [87], SynthText [88] DocReason25K [89], Common Crawl PDF
Chart	AI2D [90], PlotQA [91], InfoVQA [92], ChartQA [30], MapQA [93], FigureQA [94], IconQA [95], MMC-Instruction [96]
Multidisciplinary	CLEVR-Math/Super [97, 98], GeoQA+ [99], UniChart [100], ScienceQA [101], Inter-GPS [102], UniGeo [103], PMC-VQA [104], TabMWP [105], MetaMathQA [106]
Other	Stanford40 [107], GQA [108], MovieNet [109], KonIQ-10K [110], ART500K [111], ViQuAE [112]

details InternViT-300M, a lightweight vision model developed through knowledge distillation, which inherits the strengths of a powerful vision encoder. Finally, Section 3.3 describes a transfer learning framework designed to enhance the model's adaptation to downstream tasks.

3.1 Mini-InternVL

As shown in Figure 1, Mini-InternVL consists of three main components: visual encoder, MLP projector, and LLM. We employ InternViT-300M as our visual encoder, a lightweight vision model that inherits the capabilities of a powerful vision encoder. Based on InternViT-300M, we develop three versions of Mini-InternVL: Mini-InternVL-1B, Mini-InternVL-2B, and Mini-InternVL-4B, which are respectively connected to the pre-trained Qwen2-0.5B [16], InternLM2-1.8B [14], and Phi-3-mini [28]. Similar to other open-source MLLMs [2, 5, 37, 62], Mini-InternVL employs an MLP projector to connect the vision encoder and the LLMs.

We adopt a dynamic resolution input strategy similar to that of InternVL 1.5 [2], which improves the model's ability to capture fine-grained details. We also apply a pixel unshuffle operation to reduce the number of visual tokens to one-quarter of the original. Consequently, in our model, a 448×448 image is represented by 256 visual tokens, enabling it to process up to 40 image tiles (*i.e.*, 4K resolution).

The training of Mini-InternVL consists of two stages: (1) Language-image alignment: We keep only the MLP component unfrozen during this stage. Following InternVL 1.5 [2], we use a diverse range of training datasets that encompass various tasks, including captioning, detection, grounding, and OCR. The diversity of these datasets ensures robust pre-training of Mini-InternVL, enabling the model to handle a variety of linguistic and visual elements across different tasks. (2) Visual instruction tuning: We carefully select datasets to enhance the model's performance across a broad spectrum of multimodal tasks, similar to InternVL 1.5. These tasks include image captioning, chart interpretation, OCR, and cross-disciplinary reasoning. We conduct full-parameter fine-tuning with these datasets, further injecting world knowledge and teaching models to follow user instructions.

3.2 InternViT-300M

Most existing MLLMs [20, 37, 38, 51, 61, 62] employ vision encoders that are trained on web-scale image-text paired data, such as CLIP, to obtain their representations. These encoders lack comprehensive knowledge of the visual world, which needs to be acquired through iterative generative pre-training in conjunction with LLMs. Unlike other approaches that enhance the visual foundation models by using auxiliary pathways [21, 48, 51], our method directly leverages a powerful vision model that has undergone generative training on diverse datasets to transfer knowledge to a lightweight vision model. Specifically, we use InternViT-6B as the teacher model and initialize the student model's

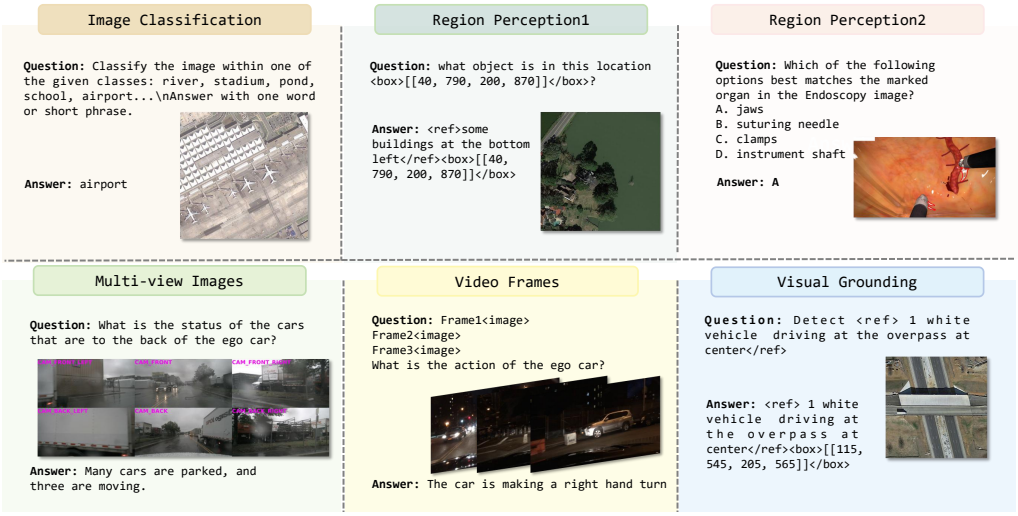


Figure 2: **The data format of our adaptation framework.** We formulate other visual tasks (*i.e.*, image classification, region perception, multi-view image tasks, video-related tasks, and visual grounding) into VQA format in our framework.

weights using CLIP-ViT-L-336px. We align the representations of the student model with those of the teacher model by computing the negative cosine similarity loss between the hidden states of the last K transformer layers. The resulting model is named InternViT-300M.

The primary goal of this knowledge transfer is to inherit the pre-training knowledge embedded in InternViT-6B. To achieve this, we curate a dataset sourced from a diverse range of publicly accessible resources, as detailed in Table 1. This dataset comprises four main types of data: natural images, OCR images, charts, and multi-disciplinary images. All images are resized to a resolution of 448 $\times$448, and dynamic resolution [2] is disabled for training efficiency. Ultimately, we develop a vision encoder, termed InternViT-300M, which is infused with diverse knowledge and is adaptable to various language models.

3.3 Domain Adaptation

Although many studies [53, 54, 55, 60] have successfully applied MLLMs to downstream tasks, a universally accepted framework for adapting MLLMs to these applications has yet to be established. Differences in model design, data formats, and training strategies across various domains result in significant heterogeneity among MLLMs, making standardization challenging. To address this issue, we propose a straightforward yet effective transfer learning framework.

Data Format. Instruction tuning is a crucial training stage to teach models to follow user instructions, where the training data is formulated as visual question answering (VQA) and conversation format. VQA datasets of the downstream tasks, such as RSVQA [113] and PMC-VQA [104], are directly utilized as instruction-following data. For other conventional tasks, as shown in Figure 2, we formulate them into VQA format according to the following approaches separately:

(1) Image Classification Tasks. In most traditional classification tasks within specialized domains, a wide range of technical terms are involved. In the majority of cases, we can easily format the classification task as a multiple-choice question. Given an image `<image>`, the set of candidate labels O, and the ground truth $G \in O$, the template can be expressed as:

> **USER**: `[Image]` `[Prompt_Prefix]` `[Candidate Labels]` `[Prompt_Suffix]`
> **ASSISTANT**: `[Ground Truth]`

A direct example can be seen in our approach to remote sensing image classification, where we utilize prompts such as "`Classify the image within one of the given classes: dense residential area, ..., school. Answer with one word or short phrase.`", as shown in Figure 2. This method transforms image classification tasks into multiple-choice questions.

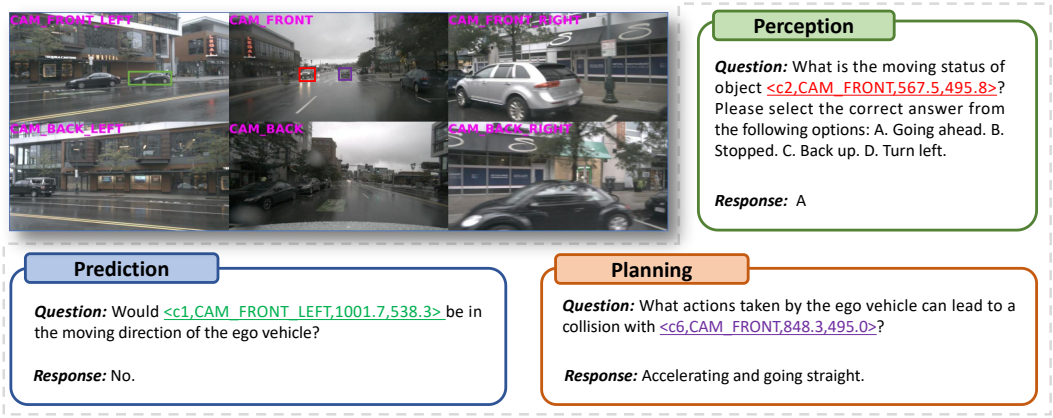


Figure 3: **Qualitative Results of Mini-InternVL-DA.** In the upper left corner is a multi-view image from DriveLM-nuScenes version-1.1 [114] after data processing. The color of the bounding boxes corresponds to the font color of the **c tags** for the objects in question (note that the input images do not contain these manually drawn bounding boxes). We show the predicted answers of our model in perception, prediction, and planning tasks. The model's outputs align with human driving behavior.

For behavior prediction of the ego vehicle in autonomous driving data, we draw inspiration from DriveLM [114] by employing templates like "`Predict the behavior of the ego vehicle. Please select the correct answer from the following options: A. The ego vehicle is going straight. The ego vehicle is not moving. B. ...`".

(2) Visual Grounding Tasks. The native support for the visual grounding task in Mini-InternVL allows the use of a special token, `<ref></ref>`, to enclose the name of the object to be detected. With this token, the model can be directed to provide the object's location enclosed within `<box></box>` in the format `[[x1, y1, x2, y2]]`, where the coordinates range from 0 to 1000. This approach enables us to convert object grounding and referring expression detection into a conversational format. We extensively apply this format to remote sensing instruction data. For example, for the referring expression "`1 overpass near some trees at the center`", we use "`Detect <ref>1 overpass near some trees at the center</ref>`" as the instruction and "`<ref>1 overpass near some trees at the center</ref><box>[[x1, y1, x2, y2]]</box>`" as the response.

(3) Region Perception Tasks. Region-level conversation tasks are prevalent in specialized domains. These tasks involve supplying the model with spatial location information, in addition to the question input. The model is required to focus on objects within the specified attention region to generate a response. Specifically, there are two implementation methods. The first method involves directly annotating the location on the image using bounding boxes, masks, or contours, as illustrated in the "Region Perception2" of Figure 2. The second method denotes the object within the question by `<box>[[x1,y1,x2,y2]]</box>`, where the coordinates are normalized between 0 and 1000. This notation guides the model's attention to specific regions within the image, enabling it to perform tasks such as region-level captioning and region-specific VQA.

For instance, in remote sensing applications, the goal is to train the model to identify objects within specific coordinates `[x1, y1, x2, y2]`. To achieve this, we use a prompt such as "`What object is in this location<box>[[x1, y1, x2, y2]]</box>`" as the input instruction, with "`<ref>object name</ref><box>[[x1, y1, x2, y2]]</box>`" serving as the label.

(4) Multi-View Images. In autonomous driving, the images are captured from six different viewpoints. As shown in Figure 3, we effectively utilize dynamic resolution to accommodate this type of data. Specifically, Mini-InternVL supports splitting images into 448×448-sized tiles based on their aspect ratio. Consequently, we resize each image to 896×448 pixels and then, as illustrated, combine these images in a fixed sequence, resulting in a final resolution of 2688×896. This means that the images are automatically processed into 12 tiles, and an additional thumbnail is added to provide the model with global context. Furthermore, we label each viewpoint image with text indicating its camera position, such as "CAM_FRON".

Table 2: **Comparison with other models on multimodal benchmarks.** We evaluate models using the InternVL and VLMEvalKit [115] repositories. AI2D [90], ChartQA [30], DocVQA [84], InfoVQA [92], and MMBench [29] are tested with InternVL, while MathVista [31] and OCRBench [116] use VLMEvalKit. For MMMU [117], we report scores from the OpenCompass leaderboard. The Avg. Score is the average of the scores from all tested benchmarks, with the OCRBench score divided by 10. The values in parentheses represent the relative parameters and performance of Mini-InternVL compared to InternVL2-Llama3-76B [2], which is considered as 100%.

model	open-source	#param	MMMU (val)	MathVista (testmini)	AI2D	ChartQA	DocVQA	InfoVQA	OCR-Bench	MMB-EN	MMB-CN	Avg. Score
GPT-4V-0409 [118]	✗	-	61.7	58.1	89.4	78.1	87.2	-	678	81.0	80.2	75.4
Gemini-Pro-1.5 [119]	✗	-	60.6	57.7	80.3	81.3	86.5	72.7	754	73.9	73.8	73.6
Claude3.5-Sonnet [120]	✗	-	65.9	67.7	94.7	90.8	95.2	-	788	79.7	80.7	81.7
GPT-4o [4]	✗	-	69.2	63.8	94.2	85.7	92.8	-	736	83.4	82.1	80.6
Cambrian-1 [121]	✓	-	50.4	53.2	79.7	75.6	75.5	-	600	81.4	-	68.0
DeepSeek-VL-1.3B [51]	✓	2B	33.8	29.8	51.5	-	-	-	413	64.6	62.9	47.3
MiniCPM-V 2.0 [20]	✓	3B	38.2	38.7	62.9	-	71.9	-	605	69.1	66.5	58.3
Qwen2-VL-2B [122]	✓	2B	42.2	43.0	74.7	73.5	90.1	65.5	794	74.9	73.5	68.5
InternVL2-Llama3-76B [2]	✓	76B	58.2	65.5	87.6	88.4	94.1	82.0	839	86.5	86.3	81.4
Mini-InternVL-1B	✓	1B (1%)	36.7	37.7	64.1	72.9	81.7	50.9	754	65.4	60.7	60.6 (74%)
Mini-InternVL-2B	✓	2B (3%)	36.3	46.3	74.1	76.2	86.9	58.9	784	73.2	70.9	66.8 (82%)
Mini-InternVL-4B	✓	4B (5%)	48.3	58.6	78.9	81.5	89.2	67.0	788	78.6	73.9	72.8 (90%)

(5) Video Frames. Mini-InternVL supports video frames in an interleaved image format. We represent the frame sequence using a template such as "`Frame1: <img><IMG_CONTEXT></img> Frame2: <img><IMG_CONTEXT></img>.`", where `<IMG_CONTEXT>` denotes the image tokens. For each image at a resolution of 448×448, the model can accommodate sequences of up to 40 frames.

Training Strategy. During the domain adaptation phase, we perform full-parameter fine-tuning on Mini-InternVL. For a domain-specific application scenario, we convert corresponding data into the required format and incorporate it into our training dataset. Adding a certain proportion of general multimodal data during the domain adaptation phase will not affect the performance in the specific domain, while retaining the model's general multimodal capability. In our experiments, we find that adding general data can improve the generalization ability of the model on other tasks. Therefore, when performing domain adaptation, we can choose the appropriate general data ratio on the premise of balancing computational overhead and performance.

4 Experiments

In this section, we begin by conducting a comprehensive comparison of our Mini-InternVL with leading multi-modal large language models (MLLMs) on representative vision-language benchmarks (Section 4.1). Following this, in Section 4.2, we apply the domain adaptation framework introduced in Section 3.3 to transfer our models to three specialized domains: autonomous driving (Section 4.2.1 and Section 4.2.2), medical images (Section 4.2.3), and remote sensing (Section 4.2.4). Additionally, we perform an extensive ablation study to explore the impact of data sample size and model size on domain adaptation (Section 4.3).

4.1 Results on General Multimodal Benchmark.

Settings. In this section, we present a comprehensive evaluation of our model's multimodal understanding and reasoning capabilities across a variety of benchmarks. The benchmarks used in our study are categorized into four distinct types: **OCR-related tasks**, including DocVQA [84], OCRBench [116] and InfographicVQA [92]; **chart and diagram understanding**, including AI2D [90] and ChartQA [30]; **general multimodal tasks**, such as MMBench [29]; and **multimodal reasoning**, including MMMU [117] and MathVista [31]. Additionally, we calculate the average score across these benchmarks, with the OCRBench score divided by 10.

Results. As shown in Table 2, Mini-InternVL demonstrates strong performance across the majority of benchmarks. Our smallest model contains only 1 billion parameters, yet it demonstrates

Table 3: **The sources of our general datasets**. In each task, we sample a specific amount of data from the data source at a predetermined ratio to balance the training dataset. In each domain adaptation case, we sample a certain amount of data from each data source.

Data Source	Size	Data Type
ShareGPT4V [123]	767K	Captioning
AllSeeingV2 [62]	127K	Grounding VQA
LLaVA-Instruct-ZH[3]	158K	VQA
DVQA [124]	200K	Diagram VQA
ChartQA [125]	18K	Diagram VQA
AI2D [126]	12K	Diagram VQA
DocVQA [84]	10K	Document VQA
GeoQA+ [127]	72K	Geometric VQA
SynthDoG-EN [128]	30K	OCR

Table 4: **The results on driving with language official leaderboard [114].** "DA" means model after domain adaptation on DriveLM. The other results in the table are taken from the CVPR 2024 Autonomous Driving Challenge leaderboard. MTMM†, MMFM_AD, and Team NVIDIA are team names on the challenge leaderboard, which we use to represent their methods.

Method	#Param	Accuracy	ChatGPT	Bleu 1	Bleu 2	Bleu 3	Bleu 4	ROUGE L	CIDEr	Match	Final Score
InternVL4Drive-v2 [129]	26B	0.7339	65.25	0.7787	0.7176	0.6608	0.6059	0.7449	0.2061	47.65	0.6002
MTMM†	-	0.7473	65.59	0.76	0.70	0.64	0.59	0.74	0.18	0.45	0.5974
Team NVIDIA	-	0.7746	59.89	-	-	-	-	-	-	-	0.5884
MMFM_AD	-	0.6658	63.92	-	-	-	-	-	-	-	0.5732
Mini-InternVL-4B	4B	0.0	54.45	0.2405	0.0801	0.0252	0.0084	0.1927	0.0018	34.30	0.3051
InternVL2-Llama3-76B	76B	0.0	52.50	0.2100	0.0884	0.0249	0.0078	0.1848	0.0001	34.22	0.2963
Mini-InternVL-DA-1B	1B	0.7007	63.84	0.7362	0.6767	0.6214	0.5678	0.7365	0.1669	39.76	0.5686
Mini-InternVL-DA-2B	2B	0.7628	65.23	0.7616	0.7012	0.6452	0.5908	0.7447	0.1914	43.24	0.5958
Mini-InternVL-DA-4B	4B	0.7296	63.97	0.7642	0.7032	0.6463	0.5914	0.7427	0.1976	42.16	0.5821

performance comparable to 2 billion parameter models, such as DeepSeek-VL-1.3B and MiniCPM-V 2.0. Compared to other lightweight models, our Mini-InternVL-4B excels across most benchmarks, particularly in MMbench, ChartQA, DocVQA, and MathVista, where its performance is on par with commercial models like Gemini-Pro-1.5. Notably, compared to InternVL2-Llama3-76B, which utilizes the larger InternViT-6B, Mini-InternVL achieves approximately 90% of its performance while using 5% parameters. This highlights the effectiveness of our knowledge distillation strategy.

4.2 Transfer to Various Specialized Domains

4.2.1 Multi-View Image-Based Autonomous Driving

Settings. We select DriveLM-nuScenes version 1.1 [114] as our training dataset, which contains 317K training samples and encompasses various aspects of the driving process. This dataset includes data for perception, prediction, and planning, offering a comprehensive understanding of autonomous driving scenarios.

In DriveLM-nuScenes, the images are captured from six different viewpoints. We effectively utilize dynamic resolution features to accommodate this type of data. Specifically, Mini-InternVL supports splitting images into 448×448-sized tiles based on their aspect ratio. As illustrated in Figure 3, we resize the image of each view to 896×448 pixels and then combine these images in a fixed sequence, resulting in a final resolution of 2688×896. This means that the images are automatically processed into 12 tiles, and an additional thumbnail is added to provide the model with global context. Furthermore, we mark the image of each view with text indicating its camera position, such as "CAM_FRON".

As shown in Figure 3, DriveLM-nuScenes contains QA pairs with coordinates, thus we need to normalize them to a range of 0 to 1000 to align with the output of Mini-InternVL. In the dataset,

[3] https://huggingface.co/datasets/openbmb/llava_zh

objects are represented by **c tags**. We use a tailored prompt: "`Objects are encoded using <c, CAM, [cx,cy]>, where c is the identifier, CAM indicates the camera where the object's center point is situated, and x, y represent the horizontal and vertical coordinates of the center point of the 2D bounding box.`" to guide the model on the composition of **c tags**. The ground truth responses typically include bounding box annotations for questions requiring a list of all objects. Therefore, We annotate the bounding box as `<box>[[x1,y1,y2,y3]]</box>`, where `<box>` and `</box>` are special tokens in Mini-InternVL. Additionally, we incorporate general datasets into the training set, to prevent the model from losing its general domain perception capabilities, maintaining a 1:4 ratio of general to domain-specific data. The sources of the general datasets are shown in Table 3.

Finally, we conduct full-parameter fine-tuning of Mini-InternVL using 8 A100 GPUs, training the model for 1 epoch with a learning rate of 1e-5. We report the performance of our model after transfer learning on the CVPR 2024 Autonomous Driving Challenge [114]. Furthermore, we evaluate our model on autonomous driving scenarios from MME-Realworld [130], where we separately assess its performance on Perception and Reasoning tasks.

Results. We test our model using DriveLM-nuScenes-version-1.1-val [114], and the results are presented in Table 4. Our final score of Mini-InternVL-2B is 0.5958, which is comparable to the best result on the CVPR 2024 Autonomous Driving Challenge Leaderboard[4], InternVL4Drive-v2 [129]. Notably, our model uses only one-tenth of the parameters of InternVL4Drive-v2.

Our model scores slightly lower in the Match metric, which might be due to Mini-InternVL's lack of proficiency in predicting object center points. InternVL4Drive-v2 [129] offers a viable solution by using Segment Anything [132] to convert object center points into object bounding boxes. In Figure 3, we show the predicted answers of our model in perception, prediction, and planning tasks, demonstrating alignment with human driving behavior. Furthermore, our 4B-parameter model performs similarly to our 2B-parameter model. We attribute this potentially to the limitations of the existing training data and evaluation criteria, which might constrain larger models from achieving significant performance gains.

Table 5: **Results on Autonomous Driving domain of MME-RealWorld.** "DA" means model after domain adaptation on DriveLM.

Method	Perception	Reasoning	Avg.
GPT-4o [4]	21.14	26.41	24.60
Claude 3.5 Sonnet [131]	32.43	31.92	32.10
LLaVA-OneVision-7B [5]	45.77	34.08	41.75
Qwen2-VL-7B [122]	34.62	31.47	33.54
InternVL2-Llama3-76B [2]	47.46	35.71	44.30
Mini-InternVL-1B	31.34	22.47	28.96
Mini-InternVL-2B	39.86	30.13	37.25
Mini-InternVL-4B	38.96	33.11	37.39
Mini-InternVL-DA-1B	42.95	27.75	38.87
Mini-InternVL-DA-2B	50.74	40.48	47.98
Mini-InternVL-DA-4B	53.14	39.14	49.38

As shown in Table 5, in the autonomous driving scenarios of MME-Realworld, we observe that even when using only DriveLM as domain-specific training data, our model achieves an improvement of over 10 points. The transferred model surpasses the best-performing model on this task, LLaVA-OneVision-7B [5], as well as several commercial closed-source models such as GPT-4o [4] and Claude 3.5 Sonnet [131], demonstrating the strong generalization capability of our model.

4.2.2 Autonomous Driving with Temporal Information

Settings. Using single-frame images alone is insufficient for accurate perception and prediction of vehicle behavior. Therefore, we explore temporal expansion. Specifically, we utilize instruction-following data constructed by DriveGPT4 [60] from the BDD-X dataset [133] as our training set, which comprises 26K video clips. Multiple video frames are organized as described in Section 3.3. Each training sample contains four aspects of question-answer data: Action Description, Action Justification, Speed Signal Prediction, and Turning Angle Signal Prediction. We set the proportion of general to domain-specific data at 2:1.

Following DriveGPT4 [60], we report several metric scores widely used in the NLP community, including CIDEr, BLEU4, and ROUGE-L, to evaluate the action descriptions and justifications. For

[4] https://opendrivelab.com/challenge2024/#driving_with_language

Table 6: **The results on action tasks of BDD-X dataset.** We provide evaluation results on action description, action justification, and full-text generation (*i.e.*, combining description and justification). "B4" stands for BLEU4. "DA" means model after domain adaptation on BDD-X.

Method	Description			Justification			Full		
	CIDEr	B4	ROUGE	CIDEr	B4	ROUGE	CIDEr	B4	ROUGE
ADAPT [134]	219.35	33.42	61.83	94.62	9.95	32.01	93.66	17.76	44.32
DriveGPT4 [60]	254.62	35.99	63.97	101.55	10.84	31.91	102.71	19.00	45.10
Mini-InternVL-DA-1B	223.85	34.17	62.11	95.52	9.70	32.58	83.72	16.78	44.29
Mini-InternVL-DA-2B	242.14	35.77	63.03	105.06	10.63	32.46	98.47	18.05	44.52
Mini-InternVL-DA-4B	237.41	35.94	63.67	104.62	9.51	32.23	97.42	17.70	44.98

Table 7: **Quantitative results of control signals prediction on BDD-X test dataset.** RMSE denotes the root mean squared error, and A_τ measures the proportion of test samples with prediction errors less than τ. "DA" means model after domain adaptation on BDD-X.

Method	Speed(m/s)					Turningangle(degree)				
	RMSE↓	$A_{0.1}$↑	$A_{0.5}$↑	$A_{1.0}$↑	$A_{5.0}$↑	RMSE↓	$A_{0.1}$↑	$A_{0.5}$↑	$A_{1.0}$↑	$A_{5.0}$↑
ADAPT [134]	3.02	9.56	24.77	37.07	90.39	11.98	27.93	66.83	75.13	89.45
DriveGPT4 [60]	1.30	30.09	60.88	79.92	98.44	8.98	59.23	72.89	79.59	95.32
Mini-InternVL-DA-1B	1.28	29.44	60.38	79.34	98.67	9.45	59.34	73.54	80.28	92.76
Mini-InternVL-DA-2B	1.26	27.96	59.23	80.06	98.78	9.52	57.40	72.54	80.06	92.04
Mini-InternVL-DA-4B	1.31	28.84	60.94	78.78	98.61	9.46	59.12	73.15	80.17	92.65

Table 8: **The sources of our medical data.** The table presents the data types and the sample sizes collected from each source. We sample a total of 500K image-text pairs from multiple publicly available datasets of different data types as our medical training data.

Data	Size	Description
PMC-VOA [135]	238K	The datasets include image-text pairs containing X-rays, pathology images, and images of affected areas, extracted from open-source websites or journal articles.
MedICaT [136]	31K	
PMC-Image [104, 135, 137]	29K	
Open-i [138]	1K	
MedPix [139]	6K	
Quilt-1M [140]	95K	A medical dataset includes image-text pairs of histopathology images.
RP3D [137]	82K	A medical dataset includes image-text pairs of X-ray images.
MIMIC-CXR [141]	14K	
Retina Image Bank [142]	4K	A medical dataset includes image-text pairs of retinal images.

open-loop control signal prediction, we use root mean squared error (RMSE) and threshold accuracies (A_τ) for evaluation.

Results. We report our scores on the BDD-X testing set in Table 6 and Table 7. Although our model has not undergone pre-training on large amounts of proprietary domain data like DriveGPT, it still performs comparably to DriveGPT4 across the four tasks and surpasses ADAPT [134]. Note that the performance of the three models on this dataset is similar, which aligns with the observations discussed in Section 4.2.1.

4.2.3 Medical Image Question Answering

Settings. We utilize several publicly available medical image-text datasets to improve the model's understanding of medical images. These datasets include a wide range of medical images, such as photos, X-rays, and pathology images. The datasets include PMC-OA [135], MedICaT [136], PMC-Image [104, 135, 137], Open-i [138], MedPix [139], Quilt-1M [140], RP3D [137], MIMIC-

Table 9: **The results of our model on GMAI-MMBench.** The results of other models are taken from the GMAI-MMBench leaderboard. "DA" means model after domain adaptation on medical data. After supervised fine-tuning, our model shows significant improvement and outperforms several medical-specialized models (*e.g.*, LLaVA-Med, RadFM) and some commercial closed-source models (*e.g.*, Claude3-Opus) on most metrics.

Model	Size	Seg C	Seg M	2D Cls	2D Det	2D Mcls_acc	2D Mcls_recall
Qwen-VL-Chat [38]	9.6B	34.45	35.20	39.55	22.04	42.88	81.23
LLaVA-NeXT-mistral-7B [143]	7.6B	36.29	35.20	39.34	27.87	44.05	47.70
LLaVA-Med [55]	-	18.45	18.97	21.15	17.14	45.84	41.19
RadFM [137]	14B	20.43	20.27	25.71	18.83	40.98	57.45
Claude3-Opus [131]	-	33.56	33.36	32.17	24.72	45.31	38.98
GPT-4V [3]	-	47.87	46.58	42.24	30.32	45.21	40.59
Mini-InternVL-1B	1B	34.30	34.55	36.02	24.08	21.67	8.57
Mini-InternVL-2B	2B	35.33	35.61	38.08	25.31	43.52	16.13
Mini-InternVL-4B	4B	36.60	36.99	38.74	26.01	43.99	16.25
Mini-InternVL-DA-1B	1B	38.67	39.44	35.87	23.09	22.79	8.99
Mini-InternVL-DA-2B	2B	40.22	39.46	39.34	25.59	44.33	16.20
Mini-InternVL-DA-4B	4B	41.41	40.45	41.34	24.84	44.33	16.59

CXR [141], and Retina Image Bank [142], which together provide a substantial collection of medical images. From these datasets, we sampled 500K image-text pairs for the model's training set. Finally, We add general data in a 1:1 ratio to the domain-specific data and conduct full-parameter training of the model for one epoch.

Results. In this section, we present the performance of Mini-InternVL and its fine-tuned variant, **Mini-InternVL-DA**, on a comprehensive medical AI benchmark, GMAI-MMBench [144]. Our evaluation is conducted using the VLMEvalKit[5] framework. Table 9 illustrates the performance of various models on the medical VQA tasks.

After supervised fine-tuning, our model shows significant improvement across most evaluation metrics. Specifically, it excels in 2D classification (2D Cls), 2D detection (2D Det), and 2D multi-class accuracy (2D Mcls_acc). These results highlight its strong multimodal understanding capabilities in complex medical visual question-answering tasks.

Furthermore, our model of 4B size outperforms several medical-specialized models (*e.g.*, LLaVA-Med [55], RadFM [137]) and some commercial closed-source models (*e.g.*, Claude3-Opus [131]) on most metrics. However, there is no improvement in multiple-choice questions after SFT, which we attribute to the lack of multiple-choice question data in the training dataset.

4.2.4 Remote Sensing

Settings. The training data is summarized in Table 10. The GeoChat instruction set [53] serves as the primary component of our training dataset. To enrich the dataset with high-resolution imagery, we also incorporate the RSVQA-HR dataset [113]. Additionally, we include 100K VQA instances sampled from FIT-RS [145] to further expand our training set. For the visual grounding task, we integrate the DIOR-RSVG dataset [146] into our training process. All data are reformatted according to the methods outlined in Section 3.3.

A single epoch of training on the visual grounding data is found to be insufficient, so we repeat the DIOR-RSVG training samples multiple times. Finally, we incorporate 20% of the total general domain training samples into the training data and conduct training following the settings described in Section 4.2.1.

We assess the performance of our transferred model using the RSVQA dataset for the VQA task and the DIOR-RSVG dataset for the visual grounding task. Following the methodology outlined in [53],

[5] https://github.com/open-compass/VLMEvalKit

Table 11: **The results on remote sensing VQA and visual grounding tasks.** For the VQA datasets, we omit area and count questions during evaluation. "DA" means model after domain adaptation on remote sensing.

Method	Size	RSVQA-LR				RSVQA-HR-Test1			RSVQA-HR-Test2			DIOR-RSVG
		Rural/Urban	Presence	Compare	Avg.	Presence	Compare	Avg.	Presence	Compare	Avg.	(acc@0.5)
RSVQA [147]	-	90.00	87.46	81.50	86.32	90.43	88.19	83.12	86.26	85.94	77.50	-
Bi-Modal [148]	-	92.66	91.06	92.66	91.63	92.03	91.83	84.98	89.37	89.62	80.54	-
EasyToHard [149]	-	91.67	90.66	87.49	89.94	91.39	89.75	93.97	87.97	87.68	79.06	-
GeoChat [53]	7B	94.00	91.09	90.33	90.70	-	-	-	-	58.45	83.19	72.30
SkyEyeGPT [150]	-	75.00	88.93	88.63	84.19	84.95	85.63	85.29	83.50	80.28	81.89	88.59
SkySenseGPT [145]	-	95.00	91.07	92.00	92.69	-	-	-	69.14	84.14	76.64	-
Mini-InternVL-4B	4B	66.00	64.64	73.26	69.55	62.42	79.20	71.72	65.73	79.70	73.55	16.89
InternVL2-Llama3-76B	76B	61.00	66.29	79.61	73.77	60.79	77.47	70.04	63.30	78.32	71.71	29.65
Mini-InternVL-DA-1B	1B	95.00	81.39	91.6	87.37	92.24	91.76	91.98	89.38	90.91	90.24	89.73
Mini-InternVL-DA-2B	2B	93.00	87.07	91.85	89.87	92.33	92.21	92.27	89.60	90.86	90.30	89.24
Mini-InternVL-DA-4B	4B	92.00	85.69	92.18	89.46	92.42	92.12	92.25	89.25	90.92	90.18	92.04

Table 12: **Comparison between Mini-InternVL-2B with Mini-InternVL-CLIP-2B across various tasks.** We test the model on DriveLM val set after fine-tuning them using domain-specific data.

Method	MMB-EN	MMB-CN	ChartQA (test)	DocVQA (val)	InfoVQA (val)	MMMU	MME-RW (AD)	DriveLM
Mini-InternVL-CLIP-2B	70.3	68.1	70.9	77.5	49.6	32.9	43.7	0.580
Mini-InternVL-2B	73.2	70.9	76.2	85.9	57.7	34.3	48.0	0.578

we chose the Presence, Comparison, and Rural/Urban subsets of the RSVQA-LR and RSVQA-HR datasets for assessment.

Results. Table 11 presents the performance of our models on remote sensing VQA and visual grounding tasks. On the RSVQA task, our model demonstrates strong performance under both high-resolution and low-resolution conditions. Unlike existing remote sensing MLLMs such as GeoChat [53] and SkySenseGPT [150], which support only single-resolution images, our model leverages dynamic resolution to effectively benefit from high-resolution training data. Compared to traditional models in the remote sensing domain–such as RSVQA [147], Bi-Modal [148], and EasyToHard [149]–our model achieves superior scores on both RSVQA-HR-Test1 and RSVQA-HR-Test2, showcasing its generalization ability. Furthermore, our models of three different sizes outperform SkyEyeGPT on DIOR-RSVG, indicating that our framework can effectively model visual grounding tasks.

Table 10: **Details on the training samples used to adapt remote sensing.**

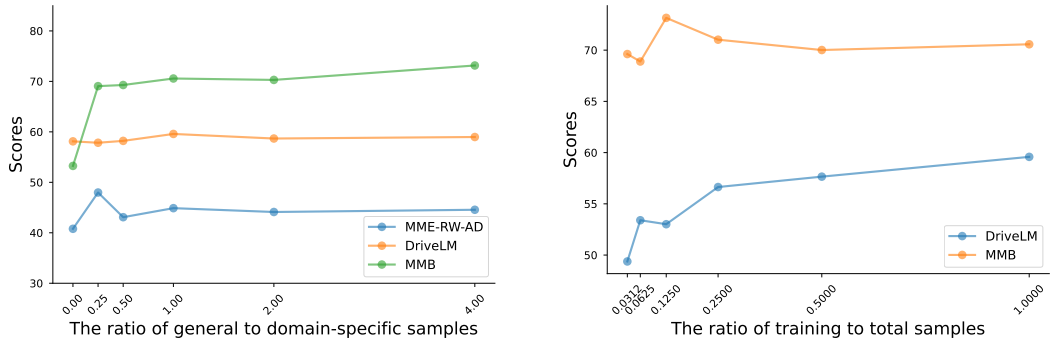
Dataset	Data type	Size
GeoChat [53]	Detailed Description	75K
	Multi-Round Conversation	65K
	Complex Questions	10K
	VQA	91.5K
	Region Captioning	40K
	Visual Grounding	25K
RSVQA-HR [113]	VQA	50K
FIT-RS-VQA [145]	VQA	100K
DIOR-RSVG [146]	Visual Grounding	26K

4.3 Ablation Study

The Importance of Knowledge Distillation. We demonstrate the effectiveness of knowledge distillation in this study. We train a new 2B-sized MLLM, which we refer to as Mini-InternVL-CLIP-2B. Specifically, we maintain the structure of the MLLM while replacing the vision en-

Table 13: **Comparing the per-GPU memory usage and training speed of 3 methods.** Our experiment use ZeRO1 [151] and is conducted on 8 A100 GPUs.

Method	GPU Memory (GB)	Training Speed (iterations per hour)
LoRA	23.40	263.02
Freezing ViT	29.87	260.76
Full-parameter	33.11	185.13



(a) The impact of the proportion of general data.

(b) The impact of the number of training samples.

Figure 4: **The impact of the proportion of general data and the number of training samples.** MME-RW-AD, MMB, and DriveLM refer to the Autonomous Driving domain of MME-RealWorld, the general multimodal benchmark MMBench, and DriveLM Challenge, respectively.

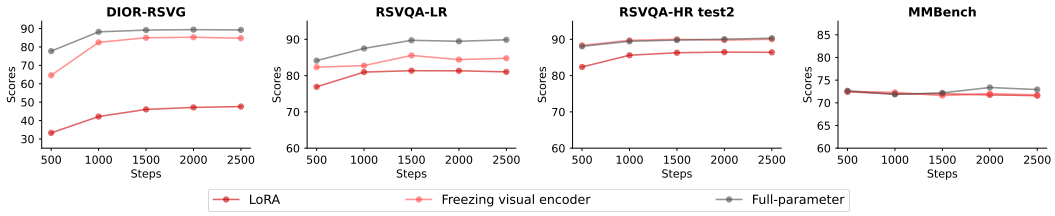


Figure 5: **Performance of 3 methods at varying training steps on 4 benchmarks.** After 1500 steps, the model's performance converges. The scores for the three adaptation methods remain stable across different training steps on the general multimodal benchmark MMBench [29].

coder from InternViT with CLIP-ViT-L-336px [18], employing the same training recipe as described in Section 3.1. As shown in Table 12, the results across multiple benchmarks indicate that Mini-InternVL-2B significantly outperforms Mini-InternVL-CLIP-2B on general benchmarks, particularly in document-related tasks. This clearly illustrates that our knowledge distillation in our method effectively enables InternViT to acquire visual world knowledge. Furthermore, we adapt both models to autonomous driving tasks, and the results demonstrate the advantages of our model in proprietary domain transfer.

The Impact of Data Ratio Balancing. In this study, we investigate the effect of the proportion between general data and domain-specific data on model transferability. We conduct experiments in the autonomous driving domain, utilizing all samples from DriveLM and supplementing them with general training data at multiples of r times the number of DriveLM samples. The results are shown in Figure 4a. Our findings indicate that relying exclusively on domain-specific training data does not yield optimal performance on downstream tasks. Introducing a specific ratio of general data not only enhances performance on domain-specific tasks but also reduces performance degradation on general multimodal benchmarks. This demonstrates that incorporating an appropriate proportion of general data is crucial for improving the model's generalization ability and maintaining its general capabilities. For our autonomous driving scenario, we observe that performance peaks at $r = 0.25$; beyond this point, performance slightly declines as r increases. This suggests that we can achieve benefits without significantly increasing computational load.

Influence of Training Sample Size. We investigate how varying the quantity of training samples affects performance on downstream tasks. In this experiment, we use different amounts of training data while maintaining a 1:1 ratio between general data and domain-specific data, as shown in Figure 4b. Training the model with only one-quarter of the full dataset significantly reduces the computational load during training while resulting in only a minor loss in performance. Notably, the

model's score on general benchmarks remains largely unchanged when the proportion of different training data is kept constant.

Effect of Different Adaptation Methods. In this study, we examine the effects of different adaptation methods–LoRA, freezing the vision encoder, and full-parameter fine-tuning–on model performance. Using the dataset described in Section 4.2.4, we apply these three methods to train the model across varying numbers of steps and evaluate its performance on three tasks: general multimodal VQA, remote sensing VQA, and visual grounding. We record memory consumption and training speed during the process, as shown in Table 13. As illustrated in Figure 5, model performance converges after 1500 steps, with each method exhibiting distinct performance ceilings. Notably, full-parameter fine-tuning achieves the highest scores on domain-specific tasks. Additionally, we find that LoRA underperforms on the visual grounding task, and freezing the vision encoder strikes a balance between performance and computational efficiency. The scores for all three adaptation methods remain stable across different training steps on the general multimodal benchmark, maintaining strong performance in the general domain even with extended training.

5 Conclusion

In this work, we introduce Mini-InternVL, a series of lightweight, open-source MLLMs designed to tackle the challenges of deploying MLLMs in resource-constrained environments. Mini-InternVL utilizes InternViT-300M as a compact vision encoder, integrating world knowledge across multiple domains through knowledge distillation from a more capable teacher model, thereby addressing the limitations of encoders like CLIP-ViT. Mini-InternVL achieves approximately 90% of the performance of larger models using significantly fewer parameters, excelling particularly in tasks such as OCR and domain-specific image understanding. To facilitate the application of small-scale multimodal models in specialized fields, we employ a unified transfer format, enabling our models to be effectively adapted to multiple specific domains, where they achieve comparable performance to other domain-specific approaches. We hope that this work provides valuable insights into the application of MLLM.

References

[1] 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 *CVPR*, pages 24185–24198, 2024.

[2] 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.

[3] OpenAI. Gpt-4v(ision) system card. https://cdn.openai.com/papers/GPTV_System_Card.pdf, 2023.

[4] OpenAI. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/, 2024.

[5] 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.

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

[7] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.

[8] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *CVPR*, pages 26689–26699, 2024.

[9] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*, 2024.

[10] Chunjiang Ge, Sijie Cheng, Ziming Wang, Jiale Yuan, Yuan Gao, Jun Song, Shiji Song, Gao Huang, and Bo Zheng. Convllava: Hierarchical backbones as visual encoder for large multimodal models. *arXiv preprint arXiv:2405.15738*, 2024.

[11] 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.

[12] 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.

[13] 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.

[14] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*, 2024.

[15] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[16] 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.

[17] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, and Jack Clark. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[19] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, pages 11975–11986, 2023.

[20] 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.

[21] 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:2403.18814*, 2023.

[22] Xiangxiang Chu, Limeng Qiao, Xinyang Lin, Shuang Xu, Yang Yang, Yiming Hu, Fei Wei, Xinyu Zhang, Bo Zhang, Xiaolin Wei, et al. Mobilevlm: A fast, reproducible and strong vision language assistant for mobile devices. *arXiv preprint arXiv:2312.16886*, 2023.

[23] Xiangxiang Chu, Limeng Qiao, Xinyu Zhang, Shuang Xu, Fei Wei, Yang Yang, Xiaofei Sun, Yiming Hu, Xinyang Lin, Bo Zhang, et al. Mobilevlm v2: Faster and stronger baseline for vision language model. *arXiv preprint arXiv:2402.03766*, 2024.

[24] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

[25] Muyang He, Yexin Liu, Boya Wu, Jianhao Yuan, Yueze Wang, Tiejun Huang, and Bo Zhao. Efficient multimodal learning from data-centric perspective. *arXiv preprint arXiv:2402.11530*, 2024.

[26] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[27] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916. PMLR, 2021.

[28] 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.

[29] 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.

[30] 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 *ACL*, pages 2263–2279, 2022.

[31] 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.

[32] Zhaoyang Liu, Yinan He, Wenhai Wang, Weiyun Wang, Yi Wang, Shoufa Chen, Qinglong Zhang, Zeqiang Lai, Yang Yang, Qingyun Li, et al. Interngpt: Solving vision-centric tasks by interacting with chatgpt beyond language. *arXiv preprint arXiv:2305.05662*, 2023.

[33] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *NeurIPS*, 36, 2024.

[34] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023.

[35] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900. PMLR, 2022.

[36] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, pages 19730–19742. PMLR, 2023.

[37] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 36, 2024.

[38] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

[39] 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.

[40] Weiyun Wang, Shuibo Zhang, Yiming Ren, Yuchen Duan, Tiantong Li, Shuo Liu, Mengkang Hu, Zhe Chen, Kaipeng Zhang, Lewei Lu, et al. Needle in a multimodal haystack. *arXiv preprint arXiv:2406.07230*, 2024.

[41] 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. *NeurIPS*, 35:23716–23736, 2022.

[42] 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.

[43] 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.

[44] Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. Introducing our multimodal models, 2023. URL https://www.adept.ai/blog/fuyu-8b.

[45] Xi Victoria Lin, Akshat Shrivastava, Liang Luo, Srinivasan Iyer, Mike Lewis, Gargi Gosh, Luke Zettlemoyer, and Armen Aghajanyan. Moma: Efficient early-fusion pre-training with mixture of modality-aware experts. *arXiv preprint arXiv:2407.21770*, 2024.

[46] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

[47] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

[48] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *CVPR*, pages 9568–9578, 2024.

[49] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[50] Gen Luo, Yiyi Zhou, Yuxin Zhang, Xiawu Zheng, Xiaoshuai Sun, and Rongrong Ji. Feast your eyes: Mixture-of-resolution adaptation for multimodal large language models. *arXiv preprint arXiv:2403.03003*, 2024.

[51] 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.

[52] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

[53] Kartik Kuckreja, Muhammad S. Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad S. Khan. Geochat: Grounded large vision-language model for remote sensing. *CVPR*, 2024.

[54] Wei Zhang, Miaoxin Cai, Tong Zhang, Yin Zhuang, and Xuerui Mao. Earthgpt: A universal multi-modal large language model for multi-sensor image comprehension in remote sensing domain. *TGRS*, 2024.

[55] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*, 2023.

[56] Junling Liu, Ziming Wang, Qichen Ye, Dading Chong, Peilin Zhou, and Yining Hua. Qilin-med-vl: Towards chinese large vision-language model for general healthcare. *arXiv preprint arXiv:2310.17956*, 2023.

[57] Junxian Li, Di Zhang, Xunzhi Wang, Zeying Hao, Jingdi Lei, Qian Tan, Cai Zhou, Wei Liu, Weiyun Wang, Zhe Chen, et al. Seeing and understanding: Bridging vision with chemical knowledge via chemvlm. *arXiv preprint arXiv:2408.07246*, 2024.

[58] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Chenxu Hu, Yang Wang, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*, 2024.

[59] Wenhai Wang, Jiangwei Xie, ChuanYang Hu, Haoming Zou, Jianan Fan, Wenwen Tong, Yang Wen, Silei Wu, Hanming Deng, Zhiqi Li, et al. Drivemlm: Aligning multi-modal large language models with behavioral planning states for autonomous driving. *arXiv preprint arXiv:2312.09245*, 2023.

[60] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K. Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *arXiv preprint arXiv:2310.01412*, 2023.

[61] 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 *ICLR*, 2024.

[62] 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.

[63] 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. *NeurIPS*, 35: 25278–25294, 2022.

[64] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset, 2022.

[65] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[66] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019.

[67] 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 *ICCV*, pages 8430–8439, 2019.

[68] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, pages 2641–2649, 2015.

[69] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123:32–73, 2017.

[70] 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.

[71] 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.

[72] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *ECCV*, pages 742–758, 2020.

[73] 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. *NeurIPS*, 35:26418–26431, 2022.

[74] 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.

[75] 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.

[76] 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 *ICDAR*, pages 1557–1562, 2019.

[77] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *ICCV*, pages 4291–4301, 2019.

[78] 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 *ICDAR*, volume 1, pages 1429–1434, 2017.

[79] 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 *ICDAR*, pages 1577–1581, 2019.

[80] 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 *ICDAR*, pages 1571–1576, 2019.

[81] 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 *ECCV*, 2022.

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

[83] 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.

[84] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *WACV*, pages 2200–2209, 2021.

[85] 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 *CVPR*, pages 8802–8812, 2021.

[86] 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.

[87] 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 *CVPR*, pages 4999–5007, 2017.

[88] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data for text localisation in natural images. In *CVPR*, pages 2315–2324, 2016.

[89] 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.

[90] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, pages 235–251, 2016.

[91] Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots. In *WACV*, pages 1527–1536, 2020.

[92] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *WACV*, pages 1697–1706, 2022.

[93] Shuaichen Chang, David Palzer, Jialin Li, Eric Fosler-Lussier, and Ningchuan Xiao. Mapqa: A dataset for question answering on choropleth maps. In *NeurIPS Workshop*, 2022.

[94] 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.

[95] 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. In *NeurIPS*, 2021.

[96] 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.

[97] 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 *CVPR*, pages 14963–14973, 2023.

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

[99] Jie Cao and Jing Xiao. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. In *COLING*, pages 1511–1520, 2022.

[100] 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.

[101] 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. *NeurIPS*, 35:2507–2521, 2022.

[102] Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *ACL*, 2021.

[103] 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.

[104] 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.

[105] 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. In *ICLR*, 2023.

[106] 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.

[107] 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 *ICCV*, pages 1331–1338, 2011.

[108] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, pages 6700–6709, 2019.

[109] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In *ECCV*, 2020.

[110] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *TIP*, 29:4041–4056, 2020.

[111] Hui Mao, Ming Cheung, and James She. Deepart: Learning joint representations of visual arts. In *ACM MM*, pages 1183–1191. ACM, 2017.

[112] 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 *SIGIR*, pages 3108–3120, 2022.

[113] Sylvain Lobry, Diego Marcos, Jesse Murray, and Devis Tuia. Rsvqa: Visual question answering for remote sensing data. *TGRS*, 58(12):8555–8566, 2020.

[114] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. *arXiv preprint arXiv:2312.14150*, 2023.

[115] 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. *arXiv preprint arXiv:2407.11691*, 2024.

[116] 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.

[117] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, pages 9556–9567, 2024.

[118] 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.

[119] 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.

[120] Anthropic. Claude 3.5 sonnet model card addendum. https://www.anthropic.com, 2024. URL https://www-cdn.anthropic.com/fed9cc193a14b84131812372d8d5857f8f304c52/Model_Card_Claude_3_Addendum.pdf.

[121] 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.

[122] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[123] 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.

[124] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *CVPR*, pages 5648–5656, 2018.

[125] 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 *ACL*, pages 2263–2279, 2022.

[126] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, pages 235–251. Springer, 2016.

[127] Jie Cao and Jing Xiao. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. In *COLING*, pages 1511–1520, 2022.

[128] 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 *ECCV*, 2022.

[129] Jiajhan Li and Tong Lu. Driving with internvl. 2024.

[130] 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.

[131] Anthropic. The claude 3 model family: Opus, sonnet, haiku. https://www.anthropic.com, 2024. URL https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.

[132] 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 *ICCV*, pages 4015–4026, 2023.

[133] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual explanations for self-driving vehicles. *ECCV*, 2018.

[134] Bu Jin, Xinyu Liu, Yupeng Zheng, Pengfei Li, Hao Zhao, Tong Zhang, Yuhang Zheng, Guyue Zhou, and Jingjing Liu. Adapt: Action-aware driving caption transformer. *arXiv preprint arXiv:2302.00673*, 2023.

[135] Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-clip: Contrastive language-image pre-training using biomedical documents. *arXiv preprint arXiv:2303.07240*, 2023.

[136] Sanjay Subramanian, Lucy Lu Wang, Sachin Mehta, Ben Bogin, Madeleine van Zuylen, Sravanthi Parasa, Sameer Singh, Matt Gardner, and Hannaneh Hajishirzi. Medicat: A dataset of medical images, captions, and textual references. In *EMNLP*, 2020.

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

[138] https://openi.nlm.nih.gov/.

[139] https://medpix.nlm.nih.gov/home.

[140] Wisdom Oluchi Ikezogwo, Mehmet Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Stefan Chan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. Quilt-1m: One million image-text pairs for histopathology. *arXiv preprint arXiv:2306.11207*, 2023.

[141] Alistair E W Johnson, David J Stone, Leo A Celi, and Tom J Pollard. The mimic code repository: enabling reproducibility in critical care research. *JAMIA*, 25(1):32–39, 2018.

[142] https://imagebank.asrs.org/.

[143] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, pages 26296–26306, 2024.

[144] Pengcheng Chen, Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, et al. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. *arXiv preprint arXiv:2408.03361*, 2024.

[145] Junwei Luo, Zhen Pang, Yongjun Zhang, Tingzhu Wang, Linlin Wang, Bo Dang, Jiangwei Lao, Jian Wang, Jingdong Chen, Yihua Tan, and Yansheng Li. Skysensegpt: A fine-grained instruction tuning dataset and model for remote sensing vision-language understanding. *arXiv preprint arXiv:2406.10100*, 2024.

[146] Yang Zhan, Zhitong Xiong, and Yuan Yuan. Rsvg: Exploring data and models for visual grounding on remote sensing data. *TGRS*, 61:1–13, 2023.

[147] Zixiao Zhang, Licheng Jiao, Lingling Li, Xu Liu, Puhua Chen, Fang Liu, Yuxuan Li, and Zhicheng Guo. A spatial hierarchical reasoning network for remote sensing visual question answering. *TGRS*, 61:1–15, 2023.

[148] Yakoub Bazi, Mohamad Mahmoud Al Rahhal, Mohamed Lamine Mekhalfi, Mansour Abdulaziz Al Zuair, and Farid Melgani. Bi-modal transformer-based approach for visual question answering in remote sensing imagery. *TGRS*, 60:1–11, 2022.

[149] Zhenghang Yuan, Lichao Mou, Qi Wang, and Xiao Xiang Zhu. From easy to hard: Learning language-guided curriculum for visual question answering on remote sensing data. *TGRS*, 60: 1–11, 2022.

[150] Yang Zhan, Zhitong Xiong, and Yuan Yuan. Skyeyegpt: Unifying remote sensing vision-language tasks via instruction tuning with large language model. *arXiv preprint arXiv:2401.09712*, 2024.

[151] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE, 2020.