

OCEAN-OCR: TOWARDS GENERAL OCR APPLICATION VIA A VISION-LANGUAGE MODEL

Song Chen^{1*} Xinyu Guo^{1*} Yadong Li^{1*} Tao Zhang¹ Mingan Lin¹ Dongdong Kuang^{1,2}
 Youwei Zhang^{1,3} Lingfeng Ming¹ Fengyu Zhang¹ Yuran Wang^{1,4} Jianhua Xu^{1†} Zenan Zhou^{1†} Weipeng Chen¹
¹ Baichuan Inc. ² Beihang University ³ Beijing University of Posts and Telecommunications ⁴ Wuhan University
 {xujianhua, zhouzenan}@baichuan-inc.com

Q <https://github.com/guox25/Ocean-OCR>

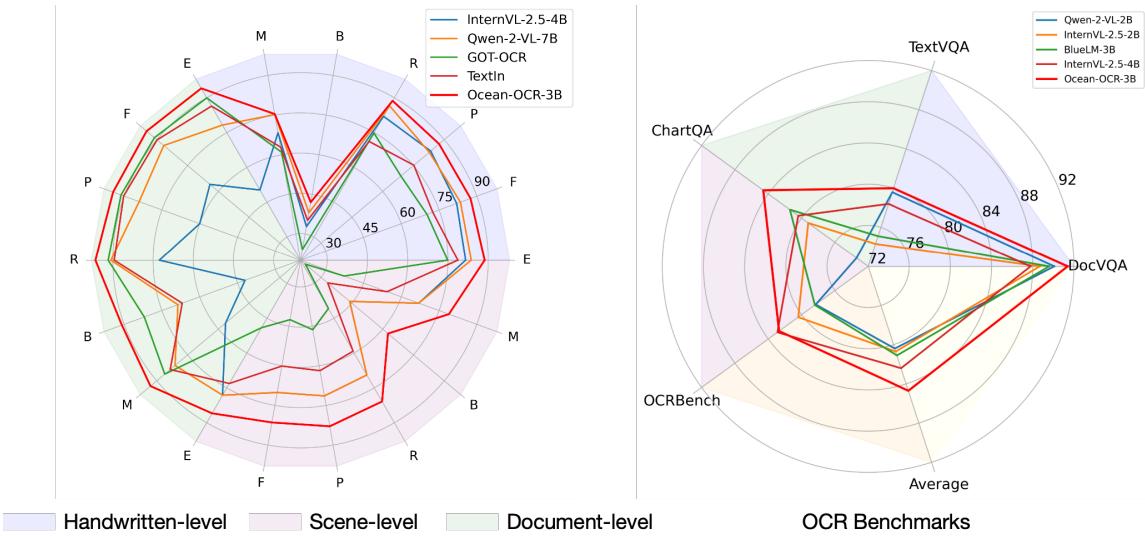


Figure 1: **Comparison with models across various OCR scenarios and benchmarks.** (**Left**) Current mainstream MLLMs and specific OCR models across multiple noteworthy OCR abilities, that is, scene-level, document-level, and handwritten-level text recognition. *E*, *F*, *P*, *R*, *B*, and *M* are the abbreviations for *Edit Distance*, *F1-Score*, *Precision*, *Recall*, *BLEU*, and *METEOR* respectively. For *Edit Distance*, the plotted score is computed with $x_{\text{after}} = 100 - x_{\text{before}}$ for better visualization. (**Right**) Comparison of mainstream MLLMs performance on OCR benchmarks.

ABSTRACT

Multimodal large language models (MLLMs) have shown impressive capabilities across various domains, excelling in processing and understanding information from multiple modalities. Despite the rapid progress made previously, insufficient OCR ability hinders MLLMs from excelling in text-related tasks. In this paper, we present **Ocean-OCR**, a 3B MLLM with state-of-the-art performance on various OCR scenarios and comparable understanding ability on general tasks. We employ Native Resolution ViT to enable variable resolution input and utilize a substantial collection of high-quality OCR datasets to enhance the model performance. We demonstrate the superiority of Ocean-OCR through comprehensive experiments on open-source OCR benchmarks and across various OCR scenarios. These scenarios encompass document understanding, scene text recognition, and handwritten recognition, highlighting the robust OCR capabilities of Ocean-OCR. Note that Ocean-OCR is the first MLLM to outperform professional OCR models such as TextIn and PaddleOCR.

*Equal core contributors.

†Corresponding author.

1 Introduction

Recently, multimodal large language models (MLLMs) [5, 14, 32, 34, 55, 63] have risen to prominence as a crucial advancement in artificial intelligence, demonstrating the ability to process and understand information across various modalities, including text, images, and videos. Nonetheless, developing effective MLLMs remains challenging, which demands sophisticated architectures, comprehensive and high-quality data, and extensive computational resources. Besides, as a vital source of information, recognition of textual content within images asks for a more fine-grained perception, posing a significant obstacle for improving the OCR ability of MLLMs.

Various attempts have been made to empower the OCR ability of MLLMs, including the sliding window strategy [33, 42], the layout-aware compression [20, 22], etc. MLLMs mainly focus on visual reasoning performance, and consequently, their capabilities in perception are not as strong. Given this limitation, some studies argue that MLLMs are not well-suited for OCR tasks [58]. However, our Ocean-OCR model delivers exceptional OCR performance while retaining powerful reasoning abilities.

In this work, we introduce Ocean-OCR, a 3B MLLM that excels in OCR tasks while achieving comparable performance on general-purpose tasks. Ocean-OCR adopts the Native Resolution ViT (NaViT) [12] to address the challenge of varying resolutions present in OCR tasks and employs an MLP to map the visual tokens into the language feature space. We assess Ocean-OCR across a diverse set of comprehensive benchmarks to highlight its broad applicability and robust general-purpose performance. On various OCR-related benchmarks, our model consistently exhibits superior performance, showcasing a clear advantage over other models. To evaluate the OCR ability in real-world applications, we construct extensive evaluation datasets including bilingual dense document understanding, practical scene text recognition, and bilingual handwritten text recognition. In these real-world OCR scenarios, our Ocean-OCR exhibits significantly leading performance. The key advances in Ocean-OCR include the following:

- We introduce Ocean-OCR, a versatile MLLM with 3B parameters that accommodates visual inputs of any resolution. Ocean-OCR is the first MLLM to outperform professional OCR models such as TextIn and PaddleOCR in various OCR scenarios.
- Remarkable performance on various benchmarks. Ocean-OCR demonstrates state-of-the-art performance on a multitude of OCR-related benchmarks, such as DocVQA, ChartQA, TextVQA, and OCRCBench. Besides, Ocean-OCR also achieves comparable results among mainstream MLLMs with similar parameter sizes on general benchmarks, such as SEEDBench.
- Excellence capabilities in real-world OCR applications. We construct comprehensive evaluation datasets covering a wide range of OCR application scenarios. Our model outperforms both previous MLLMs and traditional OCR models in all scenarios.

2 Related Work

2.1 General MLLMs

Large Language Models (LLMs) have exhibited remarkable performance across a wide range of downstream tasks. Building on this progress, Multimodal Large Language Models (MLLMs) integrate vision and language information, equipping LLMs with the ability to process multimodal input. To achieve this, researchers have developed methods such as linear projection [38, 56], Q-Former [30], and Perceiver Resampler [4], each designed to effectively combine visual and textual data. LLaVA series [36–38] scale the resolution by splitting the image into grids using multiple grid configuration. Intern-VL [5, 6, 8] series employ dynamic high resolution to capture detailed information and pixel unshuffle strategy to reduce the visual tokens. Qwen2-VL [56] introduces the Naive Dynamic Resolution mechanism to dynamically process images of varying resolutions. Despite these efforts, current MLLMs still face challenges in capturing fine-grained information within dense images, particularly in OCR tasks that involve the recognition of complex and densely packed text.

2.2 MLLM-driven OCR

Despite the robust perception and reasoning capabilities of current MLLMs, the increasing demand for text-driven visual understanding necessitates more accurate OCR results. UReader introduces a Shape-Adaptive Cropping Module that divides the original image into multiple low-resolution sub-images and employs a shared low-resolution encoder. Monkey [33] and TextMonkey [42] handle high-resolution OCR images by dividing them into patches. mPLUG-DocOwl [20, 22, 62] series explore the image cropping and visual token compression approach for document understanding. Vary [57] introduces an additional SAM-style [24] visual vocabulary tailored for document and chart data,

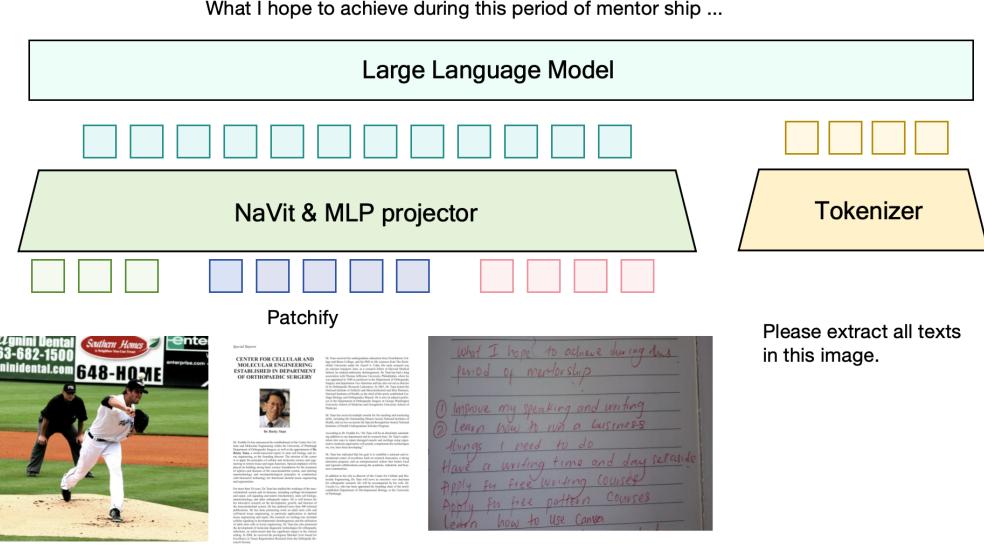


Figure 2: Overview of Ocean-OCR-3B. Following most of current MLLMs [37, 40, 55], Ocean-OCR-3B uses the conventional LLaVA-style structure that consists of a vision encoder, a MLP projector, and a LLM. To better support native dynamic high resolution in various OCR scenarios, we use NaViT-style [12] vision encoder.

which operates alongside the CLIP [49] branch. TextHawk2 [64] designs a resampler that can compress visual tokens by a factor of 16 tokens per image. GOT-OCR [58] proposes an innovative paradigm for OCR, which yielded impressive results. Although these methods have offered valuable inspiration to improve the OCR capabilities of MLLMs, their OCR performance still falls short of practical application requirements. In addition, the applicability of these models remains limited to OCR-specific scenarios and does not extend to general-purpose use.

3 Method

In this section, we provide a comprehensive overview of Ocean-OCR, delving into the details of its model architecture and foundational components.

3.1 Basic Framework

The overall architecture of Ocean-OCR is shown in Fig. 2, which is composed of the following components.

Dynamic Resolution and Image Encoder. Ocean-OCR employs Native Resolution ViT (NaViT) [12] as visual encoder. The visual encoder supports dynamic resolution, enabling Ocean-OCR to process images of any resolution and dynamically convert images into a variable number of visual tokens. This design ensures that Ocean-OCR can effectively handle a variety of image sizes while maintaining the integrity and detail of the visual information.

MLP Projector. We utilize a simple MLP layer as the projector to map the visual tokens to the input space of LLM. To address the issue of an excessive number of visual tokens in high-resolution images, we implement a strategy that compresses adjacent 2×2 tokens into a single token. This approach reduces the total number of visual tokens, thereby alleviating computational load while preserving essential visual information.

LLM. For the LLM of Ocean-OCR, we employ the Qwen-2.5-3B [52]. Considering the ease of use and practical deployment, we opted for this 3B model as it strikes a balance between model capability and size. This choice ensures efficient performance while maintaining robust language understanding and generation abilities, making it well-suited for a wide range of applications.

3.2 High-Quality Multimodal Data

We construct a comprehensive high-quality multimodal dataset from multiple sources to power Ocean-OCR-3B. As shown in Table 1, the training data cover a wide range of types, such as pure text data, caption data, interleaved data, and OCR data. The training process can be divided into three distinct stages: (1) vision-language alignment, (2)

Phase	Type	Public Datasets	Public	In-House
Alignment&Pretrain	Pure-Text	-	-	150.7M
	Caption	[31] [23] [67] [9]	33.2M	49.1M
	Interleaved	[26]	19.1M	28.7M
	OCR	[21]	12.4M	7.8M
Supervised fine-tuning	General QA	[27]	3.6M	-
	OCR QA	[53]	3M	1.9M
Total	-	-	71.3M	238.2M

Table 1: Detailed statistics of the training data of Ocean-OCR-3B.

vision-language pretraining, and (3) supervised fine-tuning. In the following section, we outline the datasets utilized in each of these stages.

3.2.1 Vision-language alignment and pretraining data

Pure text data. To reserve the strong comprehension abilities of the language model, it is necessary to contain pure text data in the training stage. For the development of a high-caliber text corpus, we collect data from an extensive variety of sources such as web pages, books, scholarly articles, programming code, and additional resources. Following the data processing protocols outlined in prior research [13, 43], we design a meticulous selection process to enhance both the diversity and the quality of our text corpus. This emphasis on diversity ensures that our training dataset covers a wide array of subjects and linguistic patterns, making it applicable to a multitude of uses. Additionally, our advanced data processing methods are tailored to remove redundancies and eliminate noise, thus amplifying the dataset’s informational richness and overall effectiveness. For the vision-language pretraining stage of Ocean-OCR, we maintain a ratio of 50% pure text data and 50% vision-language data.

Interleaved image-text data. To strengthen the model’s capability in handling interleaved image-text data, we utilize the open-source OBELICS [26] as base data. Furthermore, a comprehensive in-house dataset is developed to enrich the model’s scope of real-world knowledge. We utilize in-house collected books and papers and parse them to generate interleaved image-text data. These data are highly complete, specialized, and knowledge intensive. The ratio of OBELICS to our in-house data is approximately 4: 6.

Image caption data. As an essential part in the training of MLLMs, image caption data directly connect the visual content with textual descriptions. We adopt various open-source caption datasets, including DenseFusion-1M [31], Synthdog [23], DreamLIP [67], InternVL-SA-1B-Caption [7, 9]. In addition, considering the drop-out of OCR information in these caption data, we synthesized extensive image caption data with OCR hints using PaddleOCR [15] and GPT-4o [48]. The images of the synthetic data come form open-source datasets like Wukong [18] and Laion-2B [50].

OCR data. To enhance the model’s OCR performance, we utilize both open-source and synthetic OCR datasets. The open-source datasets is composed of DocStruct4M [21], RenderedText [1], AnyWord-3M [53], TinyChartData [21]. Our synthetic OCR data contains scene data, PDF document data, and bilingual handwritten text recognition in Chinese and English. For the natural scene data, the Chinese and English images are sampled from the Wukong [18] and Laion-2B [50] datasets, respectively. Specifically, we use the PaddleOCR [15] tools to generate pseudo-ground truth and then utilize GPT-4o to integrate them into the caption. We crawl PDF document data from in-house E-book data and use pure-text corpus for rendering handwritten text recognition data.

3.2.2 Supervised fine-tuning

The SFT data of Ocean-OCR is composed of open-source data and in-house synthetic data. The following illustrates the details of the SFT data.

General Visual-Question Answering. To enhance the general visual-question answering ability of Ocean-OCR, we utilize the open-source dataset Cauldron [27]. We perform some data filtering strategy to address certain limitations in the open-source data: (1) Open-source data originates from a wide variety of sources, leading to inconsistent response lengths. (2) The OCR quality is often low. (3) There are instances of hallucinations in the data. We concatenate the image and text data into a dialogue template and then use Qwen2-VL-72B [55] to evaluate the accuracy of the response. This process helps filter out any question-answer pairs that are not sufficiently accurate.

OCR data. Given the limited volume of open-source data, particularly for specific OCR tasks, we have expanded our OCR dataset by synthesizing several types of data: (1) Scene OCR Data: We synthesize scene OCR data using

sources such as COCO-Text [54], ICDAR2019 ArT [10], and Incidental Scene Text [60]. For this synthesis, we employ GPT-4o [48] to generate realistic text visual-question-answers within images. (2) Handwritten OCR Data: To create synthetic handwritten OCR data, we utilize a variety of handwriting styles from different fonts. This data is generated based on corpus content to mimic authentic handwritten text. (3) In-House Document PDF Data: We also include data derived from in-house document PDFs to further enrich the dataset with diverse document layouts and content. This approach ensures that our OCR dataset is more comprehensive and better suited to handle a wide range of OCR challenges, including scene text recognition, handwritten text recognition, and document understanding.

3.3 Training Pipelines

The training process of Ocean-OCR is a three-phase pipeline designed to progressively enhance its multimodal capabilities. (1) First, we focus on training the vision-language projector MLP while keeping the vision encoder and the language model parameters fixed. (2) Second, we engage in comprehensive vision-language pre-training using the datasets described in Section 3.2.1. During this stage, all model parameters are unfrozen and trained concurrently to enhance the model’s multimodal understanding. (3) Third, we perform supervised fine-tuning with the datasets outlined in Section 3.2.2. Similar to the pretraining stage, all components of the model are updated. For all of the three stages, we utilize the next token prediction loss on the text tokens to optimize the model.

Vision-language alignment. Based on the pre-trained Qwen-2.5-3B language model, our main objective is to project the visual tokens to the feature space of text tokens, thereby enhancing the model’s capability to process visual inputs effectively. We introduce NaViT as a dynamic vision encoder to accommodate high-resolution images more flexibly. In this stage, the vision encoder and language model remain frozen, we focus on optimizing the vision-language projector.

Vision-language pretraining. Since we have established the vision-language alignment, the following stage is vision-language pre-training dedicated to building extensive joint vision-language knowledge across a variety of tasks. During this phase, all modules are trainable, including the NaViT vision encoder, the MLP projector, and the LLM. After the vision-language pretraining, we can boost the model’s multimodal understanding ability while preserving its robust language capabilities in LLM.

Supervised fine-tuning. In the supervised fine-tuning phase, the primary goal is to enhance the instruction following ability of Ocean-OCR and maintain the general ability.

4 Experiment

In this section, we carry out a series of experiments to validate the effectiveness of our proposed approaches and demonstrate the strengths in benchmark accuracy and real-world OCR scenarios.

4.1 General Benchmarks

To evaluate the general performance of Ocean-OCR, we use comprehensive benchmarks, including MMMU [65], MMBench-EN [39], MMBench-CN [39], MathVista [44], MME [17], SEEDBench [29], RealWorldQA [11], and HallusionBench [19]. To ensure consistent and reproducible evaluation results, we uniformly employ VLMEvalKit [16] for all evaluations. Every evaluation is performed in a zero-shot setting, following the models’ original configurations to maintain fairness and uniformity across various models and benchmarks. In Table 2, Ocean-OCR shows promising performance compared to other models with a similar number of parameters ($\leq 4\text{B}$). In particular, we find Ocean-OCR has extraordinary performance on SEEDBench and Hallusion Bench.

4.2 OCR Benchmarks

To demonstrate the superior OCR ability of our Ocean-OCR, we evaluate the performance on representative open-source benchmarks related to OCR, including DocVQA [47], TextVQA [51], ChartQA [46] and OCRCBench [41]. These open-source OCR benchmarks assess the OCR capabilities across various dimensions. The experimental results in Table 3 demonstrate that our model significantly outperforms other models with comparable parameter sizes in OCR tasks.

4.3 OCR practical scenarios

In this section, we verify the performance of Ocean-OCR on 4 different OCR practical scenarios, including (1) document understanding; (2) scene text recognition; (3) handwritten recognition. Note that for each benchmark, the test



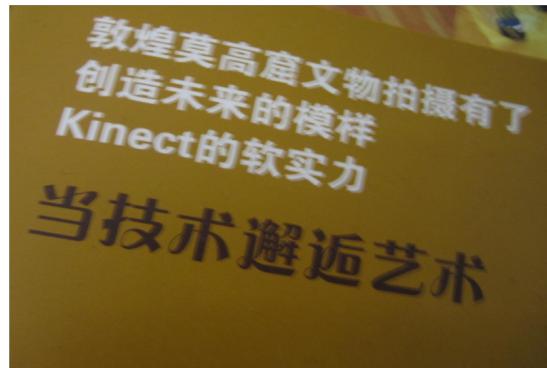
中国中钢集团公司
SINO STEEL CORPORATION
资源开发 贸易物流 工程科技
设备制造 投资发展 专业服务
www.sinosteel.com



家乐福中关村广场店 免费送货 大宗购物热
线: 51721515
CARREFOUR ZGC PLAZA STORE FREE
DELIVERY BIG PURCHASE SERVICE
LINE: 51721515



津乐汇百货 La Vita
步行街 Pedestrian Street
大食代 Food Republic
家乐福 Carrefour
自动扶梯 Elevator



敦煌莫高窟文物拍摄有了
创造未来的模样 Kinect的软实力
当技术邂逅艺术



Watsons
your personal store



Serving Soul since Twenty Twelve WELCOME
MARLOWE'S WAY PRUDOLY INDEPENDENT
FAMILY OWNED CAFE & BAR FOR MARLOWE &
HOPETON WE PLAY THIS MUSIC BECAUSE IT'S
GOOD SOUL JAZZ REGGAE HIPHOP FUNK LATIN
AFRIC & THE BLUES TRADING HOURS SEVEN
AM TO THREE PM MONDAY TO FRIDAY
CASUALLY ON INSTAGRAM TOO

Figure 3: Strong OCR ability of Ocean-OCR-3B. Our model shows strong text recognition ability across various real-world scenarios. We simply use *What is written in this image?* as prompt.

2007年6月17日，一场大雨过后，黄河兰州段又遭污染，而这次污染的是大量

2017年6月17日，一场大雨过后，黄河兰州段又遭污染，而这次污染的是大量

由约翰、塔杜诺领导的这个研究小组通过磁化石对早期磁

gunpowder exclaim

由约翰、塔杜诺领导的这个研究小组通过磁化石对早期磁

gunpowder exclaim

6月19日，北京铁路暑运方案出台。暑运期间，北京铁路局预计发送旅客310万人次，同比增长10%。在运能安排上，北京铁路局将加开北京至哈尔滨、大连、烟台等方向临客25对。

2007年暑运自7月1日起至8月31日止，共计62天。北京铁路局根据客流调查对暑运客流情况进行了预测。预计暑运期间将发送旅客310万人次，比去年同期增加283万人次，暑运高峰日将达到58万人次，比去年同期高峰日增加5万人次。暑期客流以学生、旅游观光、休闲度假为主，客流方向集中在哈尔滨、大连、青岛、烟台、南京、上海等方向。

6月19日，北京铁路暑运方案出台。暑运期间，北京铁路局预计发送旅客310万人次，同比增长10%。在运能安排上，北京铁路局将加开北京至哈尔滨、大连、烟台等方向临客25对。

2007年暑运自7月1日起至8月31日止，共计62天。北京铁路局根据客流调查对暑运客流情况进行了预测。预计暑运期间将发送旅客310万人次，比去年同期增加283万人次，暑运高峰日将达到58万人次，比去年同期高峰日增加5万人次。暑期客流以学生、旅游观光、休闲度假为主，客流方向集中在哈尔滨、大连、青岛、烟台、南京、上海等方向。

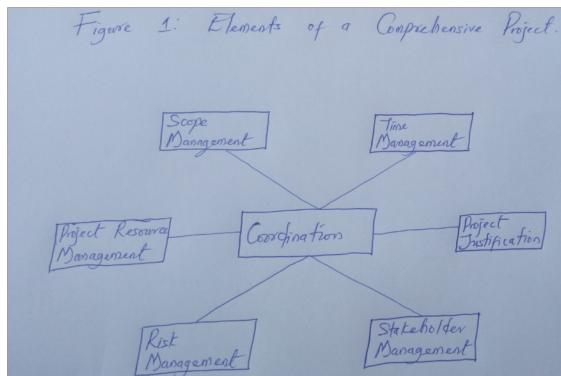
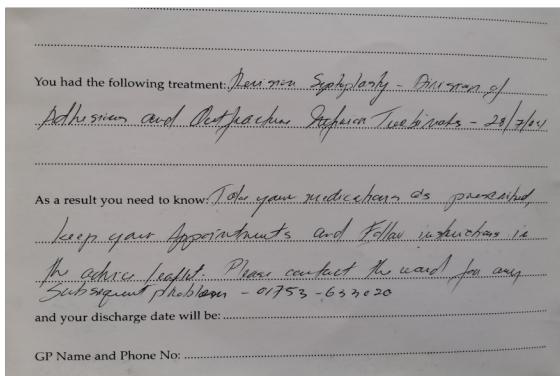


Figure 1: Elements of a Comprehensive Project. Scope Management Time Management Project Resource Management Coordination Project Justification Risk Management Stakeholder Management

Figure 4: Strong OCR ability of Ocean-OCR-3B. Our model shows strong ability for handwritten text recognition in Chinese and English. We simply use *Please extract all texts in this image.* as prompt.

年第期先锋农业先锋顾问江树人裴维蕃吴常信李沙民李德发名誉主编唐运新编委主任李明阳主编岳高峰副主任田飞先锋农业编委张玉枝孙小平臧勇军责任主编李东阳先锋农业农先锋裴维蕃吴常信李德发编委主任李明阳副主任田飞孙小平臧勇军中国农业大学进军西部年月日在开发中西部情况通报暨动员会上中国农业大学江树人校长向全校各部门各地系发出了进军西部的动员令。这次会议由副校长傅泽田

年第三期先锋农业先锋顾问江树人裴维蕃吴常信李沙民李德发名誉主编唐运新编委主任李明阳主编岳高峰副主任田飞先锋农业编委张玉枝孙小平臧勇军责任编辑主编李东阳先锋农业农先锋裴维蕃吴常信李德发编委主任李明阳副主任田飞孙小平臧勇军中国农业大学进军西部年月日在开发中西部情况通报暨动员会上中国农业大学江树人校长向全校各部门各地系发出了进军西部的动员令。这次会议由副校长傅泽田



You had the following treatment: Revision Septoplasty - Dni-ram of Adhesions and Septorhinoplasty - 28/7/21 As a result you need to know: Take your medication as prescribed, Keep your appointments and follow instructions in the advice leaflet. Please contact the ward for any subsequent problems - 01753-633020 and your discharge date will be: GP Name and Phone No:

整合进入世界的家庭系统。

在《消失的地域》一书中，约书亚·梅洛维茨观察说，电子媒体可以把信息和体验从一个地方带到另一个地方”因为当人们看到和听到经电子传播而散发的内容时，常常是在同样的时间内获得同样的意象，他们会感到自己被输送到一个同样的空间内。也正是为此，与“敝视监狱”相比，传播和信息技术形成了权力和控制的同样的撒播，但不再受边沁的砖石原型的限制。

在这个意义上，社会学家托马斯·麦谢森认为福柯忽略了大众媒体这种前现代过渡到现代的新的权力技术，他因而提出了“单视监狱”的概念（*Synopticon*），即同福柯设想的少数观看多数（*the few watch the many*）的模式不同，大众媒体，特别是电视，构成了多数观看少数（*the many watch the few*）的模式。虽然观看者彼此距离遥远，但观看的动作本身把全世界的观众带进同一个电子空间，只有少数人才能成为被观看者，大多数人都是观众，而被观看的少数人成了多数人景仰与效仿的榜样。

杰弗里·罗森进一步发挥了两个人的概念，提出“全视监狱”（*Omni pticon*），即多数观看多数（*the many watch the many*），毫无疑问这构成了互联网时代的权力技术。生活在全视监狱之中，我们从来不知道在任意时间内我们看到谁，谁在观看我们，个人不得不担心自己在公开和私下场合表现的一致性。

网络时代的个人，无论是公开，还是私下，无论是网上，还是线下，都不能再对自己的言行粗心大意或是轻佻妄动。香港巴士阿叔吵架，希拉里唱国歌跑调，新加坡南洋理工大学女生 Tammy 手机自拍做爱录像被传到网上，北京海淀艺校辱师视频引起千夫所指，都证明了今天的技术使得任何人——或许是自私的、不负责任的和怀有恶意的人——拥有不花任何代价在全球范围内侵犯隐私的能力。他们所需的只是一部电脑和一根网线，而很多博客服务是完全免费的。网络时代的个人，就如同 The Police 所唱的那首歌一样：

整合进入世界的家庭系统。在《消失的地域》一书中，约书亚·梅洛维茨观察说，电子媒体可以把信息和体验从一个地方带到另一个地方”因为当人们看到和听到经电子传播而散发的内容时，常常是在同样的时间内获得同样的意象，他们会感到自己被输送到一个同样的空间内。也正是为此，与“敝视监狱”相比，传播和信息技术形成了权力和控制的同样的撒播，但不再受边沁的砖石原型的限制。在这个意义上，社会学家托马斯·麦谢森认为福柯忽略了大众媒体这种前现代过渡到现代的新的权力技术，他因而提出了“单视监狱”的概念（*Synopticon*），即同福柯设想的少数观看多数（*the few watch the many*）的模式不同，大众媒体，特别是电视，构成了多数观看少数（*the many watch the few*）的模式。虽然观看者彼此距离遥远，但观看的动作本身把全世界的观众带进同一个电子空间，只有少数人才能成为被观看者，大多数人都是观众。而被观看的少数人成了多数人景仰与效仿的榜样。杰弗里·罗森进一步发挥了两个人的概念，提出“全视监狱”（*Omni pticon*），即多数观看多数（*the many watch the many*），毫无疑问这构成了互联网时代的权力技术。生活在全视监狱之中，我们从来不知道在任意时间内我们看到谁，谁在观看我们，个人不得不担心自己在公开和私下场合表现的一致性。网络时代的个人，无论是公开，还是私下，无论是网上，还是线下，都不能再对自己的言行粗心大意或是轻佻妄动。香港巴士阿叔吵架，希拉里唱国歌跑调，新加坡南洋理工大学女生 Tammy 手机自拍做爱录像被传到网上，北京海淀艺校辱师视频引起千夫所指，都证明了今天的技术使得任何人——或许是自私的、不负责任的和怀有恶意的人——拥有不花任何代价在全球范围内侵犯隐私的能力。他们所需的只是一部电脑和一根网线，而很多博客服务是完全免费的。网络时代的个人，就如同 The Police 所唱的那首歌一样：

ND axons were distributed widely throughout both the proximal and distal parts of the penis (Fig. 2a-e). Most of these axons were varicose and very fine and they were generally more prevalent in the tissues of the proximal region. Endothelial cells lining many blood vessels were more faintly stained, whereas within the cavernous spaces some groups of endothelial cells exhibited darker staining (Fig. 2f). ND epithelial cells lined the urethra (Fig. 2d).

The majority of blood vessels throughout the penis were supplied by very fine, varicose ND axons (Fig. 2a-e). Particularly delicate ND axons were associated with tissues of the skin of the distal penis, including the blood vessels, smooth muscle and, occasionally, approaching the epithelium (Fig. 2c). In the proximal penis fine ND axons in the outer tissues were sparser, but bundles of non-varicose ND axons were more common, than in the distal penis. Near the urethral epithelium varicose ND axons formed a dense plexus of very delicate fibres (Fig. 2d). ND somata and varicose axons were prevalent in sections of the trabeculae of the cavernous spaces and amongst the connective tissues of the corpora cavernosa (Fig. 2e).

ND somata and varicose axons were prevalent in sections of rectum, the somata restricted to the myenteric plexus and the axons to the circular and longitudinal muscle layers. No ND axons were found within sections of bladder wall.

These studies have shown that many rat pelvic neurons contain NADPH-diaphorase activity and therefore probably contain NOS. The distribution of these neurons within the MPG and penile nerve is identical to that of retrogradely-labelled penile neurons, which are concentrated near and within the penile nerve and constitute the vast majority of pelvic neurons in this area [3, 4, 11]. Further pieces of evidence to suggest that the NADPH-diaphorase-stained cells project to the penis are: (1) many deeply stained axons are found in the penile nerve, and (2) stained varicose axons are widely distributed to many structures within the penis. Although the MPG is the likely origin of many axons within the penis, additional sources may exist (e.g., sensory or sympathetic ganglia). A sensory origin is quite possible for the skin innervation, as ND staining has been observed in rat sacral dorsal root ganglia [1].

The results have not ruled out the possibility that some ND pelvic neurons (such as those located more ventrally or in the accessory ganglia) project to targets other than the penis. The present study eliminated the possibility of a projection to the bladder, but the rectum remains another potential target. Many ND axons were found in this tissue and, although some of these may arise from the ND somata of enteric neurons, an extrinsic source

from the MPG could not be discounted. ND axons in the accessory reproductive organs were not investigated.

In conclusion, relaxation of cavernous smooth muscle, as well as dilation of other penile vascular beds during erection, may involve neurons which contain NOS. Virtually all of the MPG neurons which project to the penis contain VIP [3, 4, 11], and many are cholinergic [7]. The present studies have labelled a similar number of MPG neurons with the NADPH diaphorase technique as labelled previously with retrograde tracing from the penis, and it is likely that the MPG neurons which project to the penis contain the combination of VIP, acetylcholine and NOS. Activation of these neurons could therefore elicit the release of one or more of these vasodilators. The role of each of these substances and their possible interactions with the NOS of endothelial origin, will be important to establish.

I wish to thank Mandy Bauer for excellent technical assistance. This study was supported by the Australian National Health and Medical Research Council.

1 Aimi, Y., Fujimura, M., Vincent, S.R. and Kimura, H., Localization of NADPH-diaphorase-containing neurons in sensory ganglia of the rat, *J. Comp. Neurol.*, 306 (1991) 382-392.

2 Bredt, D.S., Hwang, P.M. and Snyder, S., Localization of nitric oxide synthase indicating a neural role for nitric oxide, *Nature*, 347 (1990) 768-770.

3 Dail, W.G., Moll, M.A. and Weber, K., Localization of vasoactive intestinal polypeptide in penile erectile tissue and in the major pelvic ganglion of the rat, *Acta Physiol. Scand.*, 109 (1973) 137-138.

4 Dail, W.G., Trujillo, D., de la Rosa, D. and Walton, G., Autonomic innervation of the reproductive organs: analysis of the neurons whose axons project in the main penile nerve in the pelvic plexus of the rat, *Anat. Rec.*, 224 (1989) 94-101.

5 Dawson, T.M., Bredt, D.S., Futai, M., Hwang, P.M. and Snyder, S., Nitric oxide synthase and NADPH-diaphorase are identical in brain and peripheral tissues, *Proc. Natl. Acad. Sci. USA*, 88 (1991) 7797-7801.

6 Dail, W.G., Steers, W.D., Manzarekas, K., Moll, M.A. and Minorsky, N., The hypogastric nerve innervates a population of penile neurons in the pelvic plexus, *Neuroscience*, 16 (1985) 1041-1046.

7 Dail, W.G., Trujillo, D., de la Rosa, D. and Walton, G., Autonomic innervation of the reproductive organs: analysis of the neurons whose axons project in the main penile nerve in the pelvic plexus of the rat, *Anat. Rec.*, 224 (1989) 94-101.

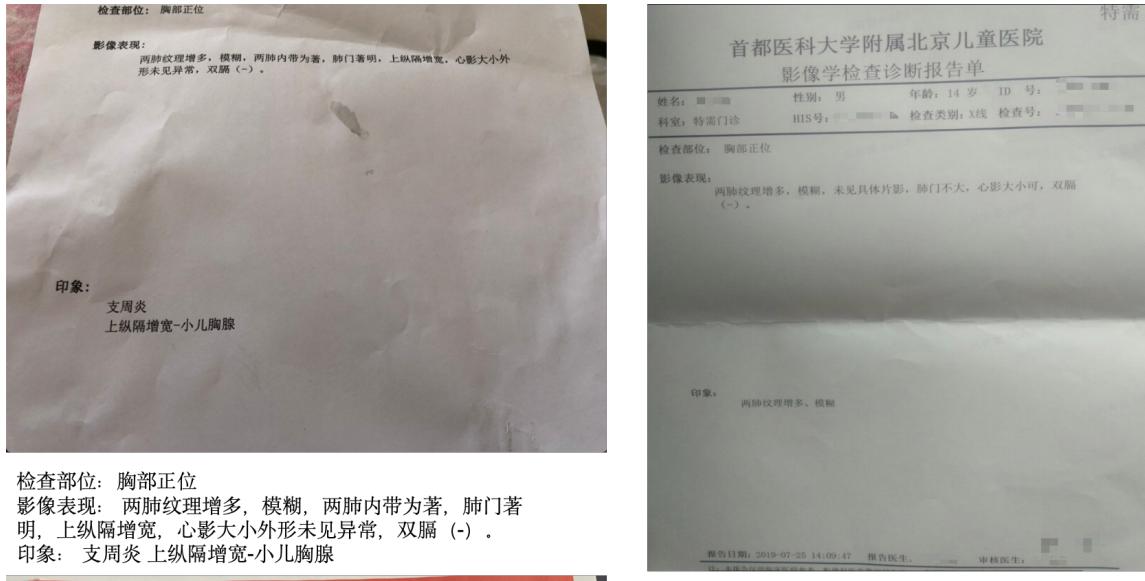
8 Holmqvist, F., Hedlund, H. and Andersson, K.-E., L-N'-Nitroarginine inhibits non-adrenergic non-cholinergic relaxation of human isolated corpus cavernosum, *Acta Physiol. Scand.*, 141 (1991) 441-442.

9 Hope, B.T., Michael, G.J., Knigge, K.M. and Vincent, S.R., Nitric oxide synthase is a nitric oxide synthase, *Proc. Natl. Acad. Sci. USA*, 88 (1991) 2811-2814.

10 Ignarro, L.J., Bush, P.A., Baga, G.M. and Wood, K.S., Nitric oxide and cyclic GMP formation upon electrical field stimulation cause relaxation of corpus cavernosum smooth muscle, *Biochem. Biophys. Res. Commun.*, 170 (1990) 843-850.

ND axons were distributed widely throughout both the proximal and distal parts of the penis (Fig. 2a-e). Most of these axons were varicose and very fine and they were generally more prevalent in the tissues of the proximal region. Endothelial cells lining many blood vessels were more faintly stained, whereas within the cavernous spaces some groups of endothelial cells exhibited darker staining (Fig. 2f). ND epithelial cells lined the urethra (Fig. 2d). The majority of blood vessels throughout the penis were supplied by very fine, varicose ND axons (Fig. 2a-e). Particularly delicate ND axons were associated with tissues of the skin of the distal penis, including the blood vessels, smooth muscle and, occasionally, approaching the epithelium (Fig. 2c). In the proximal penis fine ND axons in the outer tissues were sparser, but bundles of non-varicose ND axons were more common, than in the distal penis. Near the urethral epithelium varicose ND axons formed a dense plexus of very delicate fibres (Fig. 2d). ND somata and varicose axons were prevalent in sections of rectum, the somata restricted to the myenteric plexus and the axons to the circular and longitudinal muscle layers. No ND axons were found within sections of bladder wall. These studies have shown that many rat pelvic neurons contain NADPH diaphorase activity and therefore probably contain NOS. The distribution of these neurons within the MPG and penile nerve is identical to that of retrogradely-labelled penile neurons, which are concentrated near and within the penile nerve and constitute the vast majority of pelvic neurons in this area [3, 4, 11]. Further pieces of evidence to suggest that the NADPH-diaphorase-stained cells project to the penis are: (1) many deeply stained axons are found in the penile nerve, and (2) stained varicose axons are widely distributed to many structures within the penis. Although the MPG is the likely origin of many axons within the penis, additional sources may exist (e.g., sensory or sympathetic ganglia). A sensory origin is quite possible for the skin innervation, as ND staining has been observed in rat sacral dorsal root ganglia [1]. The results have not ruled out the possibility that some ND pelvic neurons (such as those located more ventrally or in the accessory ganglia) project to targets other than the penis. The present study eliminated the possibility of a projection to the bladder, but the rectum remains another potential target. Many ND axons were found in this tissue and, although some of these may arise from the ND somata of enteric neurons, an extrinsic source

Figure 5: Strong OCR ability of Ocean-OCR-3B. Our model shows strong ability for PDF document text recognition in Chinese and English. We simply use *Please extract all texts in this image.* as prompt.



检查部位：胸部正位
影像表现：两肺纹理增多，模糊，两肺内带为著，肺门著明，上纵隔增宽，心影大小外形未见异常，双膈（-）。
印象：支周炎 上纵隔增宽-小儿胸膜

报告日期：2019-07-25 14:09:47 报告医生： 审核医生：
首都医科大学附属北京儿童医院
影像学检查诊断报告单
姓名： 性别：男 年龄：14岁 ID号：
科室：特需门诊 HIS号： 检查类别：X线 检查号：
检查部位：胸部正位
影像表现：两肺纹理增多，模糊，未见具体片影，肺门不大，心影大小可，双膈（-）。
报告标题： 血细胞分析+C反应蛋白
报告时间： 2023-12-12 11:18:37
项目	结果	参考值	单位
白细胞 | 6.53 | 3.5--9.5 | 10^9/l
中性细胞绝对值 | 3.04 | 1.8--6.3 | 10^9/l
淋巴细胞绝对值 | 2.57 | 1.1--3.2 | 10^9/l
单核细胞绝对值 | 0.69 | 0.1--0.6 | 10^9/l
嗜酸细胞绝对值 | 0.22 | 0.02--0.52 | 10^9/l
嗜碱细胞绝对值 | 0.01 | 0--0.06 | 10^9/l
中性细胞百分比 | 46.5 | 40--75 | %
淋巴细胞百分比 | 39.3 | 20--50 | %
单核细胞百分比 | 10.7 | 3--10 | %
嗜酸细胞百分比 | 3.40 | 0.4--8 | %

夕阳云医院化子及儿免疫检验报告单
样本号：_____
姓名：_____性别：男 门诊号：_____科室：儿科门诊 申请医生：_____
采集时间：2023.12.17 10:21 采样时间：2023.12.17 10:33
床号：_____条形码号：_____临床诊断：支气管炎 接收时间：
2023.12.17 10:33 备注：
编号 项目名称 英文名称 结果 单位 参考区间 方法
1 肺炎支原体IgM抗体 MP-M 204.44 AU/ml 0-20 化学发光法
使用免疫分析仪器及其配套试剂 报告时间：2023.12.17 11:48 检验者：审核者：
此结果仅对本次标本负责，如有疑问请于3天内咨询本科室。

报告标题：血细胞分析+C反应蛋白
报告时间：2023-12-12 11:18:37

项目	结果	参考值	单位
白细胞	6.53	3.5--9.5	10^9/l
中性细胞绝对值	3.04	1.8--6.3	10^9/l
淋巴细胞绝对值	2.57	1.1--3.2	10^9/l
单核细胞绝对值	0.69	0.1--0.6	10^9/l
嗜酸细胞绝对值	0.22	0.02--0.52	10^9/l
嗜碱细胞绝对值	0.01	0--0.06	10^9/l
中性细胞百分比	46.5	40--75	%
淋巴细胞百分比	39.3	20--50	%
单核细胞百分比	10.7	3--10	%
嗜酸细胞百分比	3.40	0.4--8	%

返回

Figure 6: Strong OCR ability of Ocean-OCR-3B. Our model shows strong ability for medical report texts recognition. We simply use *Please extract all texts in this image.* as prompt.

Model	MMMU -val	MMBench -EN	MMBench -CN	MathVista -mini	MME	SEED Bench	RealWorld QA	Hallusion Bench
MM1.5-3B [66]	37.1	-	-	44.4	1798.0	72.4	56.9	-
Qwen-2-VL-2B [55]	40.0	72.2	70.1	43.2	1890.0	72.8	63.1	38.8
InternVL-2.5-2B [5]	43.6	74.7	71.9	51.3	2138.2	-	60.1	42.6
TextHawk2-7B [64]	45.0	77.5	77.6	54.5	2125.9	74.3	66.8	49.5
BlueLM-3B [45]	45.1	-	-	60.8	-	-	66.7	48
Megrez-3B-Omni [3]	51.9	80.8	82.3	62	2315	-	71.9	50.1
Phi-3.5-Vision-4B* [2]	44.0	74.1	59.9	44.7	1531.6	70.9	58.7	39.8
InternVL-2.5-4B [5]	52.3	81.1	79.3	60.5	2337.5	-	64.3	46.3
Ocean-OCR(Ours)	42.0	75.3	73.0	55.6	2094	72.5	61.2	46.0

Table 2: **Comparison of performance on general benchmarks.** * denotes the results are reproduced by ourselves and the others denote officially reported results.

Model	DocVQA	TextVQA	ChartQA	OCRBench	Average
MM1.5-3B [66]	87.7	76.5	74.2	65.7	76.0
Phi-3.5-Vision-4B* [2]	84.4	73.3	81.2	63.9	75.7
Qwen-2-VL-2B [55]	90.1	79.6	73.4	78.3	80.4
InternVL-2.5-2B [5]	88.7	74.3	79.2	80.4	80.7
TextHawk2-7B [64]	89.6	75.1	81.4	78.4	81.1
BlueLM-3B [45]	87.8	78.4	80.4	82.9	82.4
InternVL-2.5-4B [5]	91.6	76.8	84.0	82.8	83.8
Megrez-3B-Omni [3]	91.6	80.3	-	82.8	-
Ocean-OCR(Ours)	91.4	80.0	84.6	82.7	84.7

Table 3: **Comparison of performance on OCR benchmarks.** * denotes the results are reproduced by ourselves and the others denote officially reported results.

data is carefully filtered for text similarity to ensure it does not appear in the training data. GOT [58] is a MLLM designed specifically for OCR tasks. It cannot follow complex instructions and is therefore not compatible with general-purpose tasks. TextIn³ and PaddleOCR⁴ are well-known specialized models in the field of OCR.

4.3.1 Document extraction

To assess the OCR capability in understanding bilingual dense document images, we compiled an evaluation dataset consisting of 100 images from English papers and 100 images from Chinese papers. We evaluated the model using a comprehensive set of metrics, including Normalized Edit Distance, F1 Score, Precision, Recall, BLEU, and METEOR. The evaluation results are shown in Table 4. Ocean-OCR demonstrates excellent performance in PDF text recognition and document understanding, highlighting its strong capabilities in this OCR task. Fig. 5 shows two cases about document information extraction. Besides, we find that our Ocean-OCR also has excellent performance in text extraction in medical report image. We show some cases in Fig. 6. In our future work, we will establish a benchmark related to the extraction of textual information from medical reports. This initiative aims to evaluate the performance of mainstream MLLMs as well as specialized OCR models.

³For document extraction we use https://api.textin.com/ai/service/v1/pdf_to_markdown, for scene text and handwritten recognition we use <https://api.textin.com/ai/service/v2/recognize/multipage>

⁴We download model weights in <https://github.com/PaddlePaddle/PaddleOCR> for corresponding scenarios

Model	Edit Distance ↓		F1-score ↑		Precision↑		Recall↑		BLEU↑		METEOR↑	
	en	zh	en	zh	en	zh	en	zh	en	zh	en	zh
InternVL-2.5-2B [5]	0.328	0.616	0.725	0.599	0.691	0.555	0.799	0.675	0.581	0.269	0.710	0.443
Phi-3.5-Vision-4B [2]	0.295	-	0.807	-	0.785	-	0.863	-	0.704	-	0.816	-
InternVL-2.5-4B [5]	0.304	0.690	0.756	0.526	0.729	0.471	0.809	0.644	0.622	0.218	0.744	0.386
Qwen-2-VL-7B [55]	0.165	0.270	0.849	0.883	0.834	0.847	0.873	0.942	0.795	0.578	0.859	0.763
MiniCPM-V2.6-8B [61]	0.244	0.437	0.804	0.778	0.793	0.721	0.837	0.875	0.695	0.431	0.640	0.642
GOT [58]	0.084	0.117	0.895	0.928	0.891	0.934	0.906	0.929	0.835	0.805	0.874	0.848
TextIn	0.055	0.217	0.861	0.936	0.856	0.948	0.866	0.924	0.773	0.566	0.887	0.782
PaddleOCR	0.323	0.649	0.707	0.864	0.690	0.912	0.730	0.821	0.517	0.537	0.674	0.699
Ocean-OCR(Ours)	0.057	0.062	0.937	0.962	0.932	0.956	0.956	0.974	0.906	0.912	0.945	0.916

Table 4: **Comparison of performance on dense English(en) and Chinese(zh) OCR for document-level pages.**
Phi-3.5-Vision-4B fails in following instruction on Chinese document-level pages.

Model	Edit Distance↓	F1-score ↑	Precision ↑	Recall↑	BLEU↑	METEOR↑
Qwen-2-VL-2B [55]	0.292	0.710	0.705	0.757	0.283	0.641
InternVL-2.5-2B [5]	0.193	0.807	0.807	0.824	0.293	0.683
Phi-3.5-Vision-4B [2]	0.452	0.595	0.498	0.595	0.152	0.398
InternVL-2.5-4B [5]	0.184	0.820	0.834	0.820	0.328	0.683
Qwen-2-VL-7B [55]	0.163	0.835	0.827	0.865	0.380	0.752
MiniCPM-V2.6-8B [61]	0.146	0.857	0.844	0.887	0.372	0.734
GOT [58]	0.251	0.702	0.689	0.748	0.241	0.610
TextIn	0.213	0.725	0.752	0.712	0.352	0.629
PaddleOCR	0.130	0.837	0.828	0.858	0.387	0.720
Ocean-OCR(Ours)	0.113	0.875	0.875	0.887	0.420	0.754

Table 5: **Comparison of performance on OCR for scene texts.**

4.3.2 Scene text recognition

Scene text is ubiquitous in daily life, found on everything from street signs to product packaging, highlighting the crucial need for accurate recognition of scene text. Effective scene text recognition not only enhances the accessibility of information but also plays a vital role in various applications, from assistive technologies to automated systems. We have assembled a scene text OCR benchmark comprising 260 natural images, evenly divided between Chinese and English. These images are sampled from MSRA-TD500-Dataset [59]. Each image is first passed through PaddleOCR tools to get pseudo ground truth, and then each one has been manually verified and corrected to ensure the accuracy of the ground truth annotations. As illustrated in Table 5, Ocean-OCR performs outstandingly in natural scene OCR tasks, accurately identifying text even when it comprises only a small part of the image. For example, Ocean-OCR with 3B parameters even surpasses Qwen2-VL-7B and MiniCPM-V2.6-8B in all the six indicators, such as *Edit Distance* (0.163 in Qwen2-VL-7B, 0.146 in MiniCPM-V2.6-8B, and 0.113 in Ocean-OCR) and *METEOR* (0.752 in Qwen2-VL-7B, 0.734 in MiniCPM-V2.6-8B, and 0.754 in Ocean-OCR). This underscores the model’s capability to handle intricate and varied real-world scenes with high accuracy. We show some cases in Fig. 3.

Model	Edit Distance ↓		F1-score ↑		Precision↑		Recall↑		BLEU↑		METEOR↑	
	en	zh	en	zh	en	zh	en	zh	en	zh	en	zh
InternVL-2.5-2B [5]	0.227	0.255	0.619	0.717	0.633	0.733	0.614	0.709	0.363	0.464	0.611	0.661
InternVL-2.5-4B [5]	0.197	0.240	0.661	0.741	0.674	0.754	0.655	0.734	0.406	0.473	0.652	0.687
Qwen-2-VL-7B [55]	0.127	0.113	0.760	0.881	0.773	0.884	0.754	0.884	0.490	0.666	0.756	0.859
MiniCPM-V2.6-8B [61]	0.147	0.175	0.727	0.810	0.747	0.811	0.714	0.812	0.443	0.583	0.727	0.774
GOT [58]	0.616	0.402	0.283	0.568	0.309	0.618	0.273	0.544	0.151	0.295	0.255	0.492
TextIn	0.358	0.180	0.362	0.840	0.368	0.869	0.362	0.822	0.098	0.567	0.337	0.751
PaddleOCR	0.418	0.325	0.237	0.664	0.232	0.646	0.263	0.700	0.069	0.431	0.236	0.648
Ocean-OCR(Ours)	0.145	0.106	0.774	0.885	0.780	0.912	0.782	0.862	0.532	0.736	0.772	0.885

Table 6: **Comparison of performance on English(en) and Chinese(zh) OCR for handwritten recognition.** We find that Qwen-2-VL-2B and Phi-3.5-Vision-4B have trouble following instructions in this scenario.

4.3.3 Handwritten recognition

Handwritten text recognition is also a crucial component in evaluating a model’s OCR capabilities. To provide a thorough assessment, we have developed a multi-granularity handwritten text recognition evaluation dataset that incorporates both real and synthetic bilingual data. This dataset specifically includes: (1) Paragraph-level real Chinese and English data (from CASIA-HWDB [35] and GNHK [28]); (2) Line-level real Chinese and English data (from CASIA-HWDB and BRUSH [25]); (3) Paragraph-level synthetic Chinese and English data; (4) Line-level synthetic Chinese and English data. Each category contains 100 samples. On the constructed dataset, the evaluation metrics we use include Normalized Edit Distance, F1 Score, Precision, Recall, BLEU, and METEOR. As demonstrated in Table 6, Ocean-OCR also shows impressive performance in challenging tasks such as handwritten text recognition, demonstrating its robust capabilities in handling complex and varied handwriting styles. In Fig. 4, we show several cases of handwritten recognition in Chinese and English.

5 Conclusion

In this study, we introduced Ocean-OCR, a 3B MLLM that addresses the OCR limitations of existing multimodal models. Using NaViT, Ocean-OCR handles variable resolution inputs effectively, enhancing its adaptability to different image qualities. Trained on high-quality OCR datasets, Ocean-OCR excels in diverse OCR scenarios. Our experiments on open-source OCR benchmarks and real-world applications demonstrate Ocean-OCR’s superior performance and robustness. This work sets a new benchmark for multimodal learning-based OCR tasks, providing a powerful tool for accurate visual-textual information processing.

References

- [1] Renderedtext. <https://huggingface.co/datasets/wendlerc/RenderedText/>, 2024.
- [2] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.
- [3] Infinigence AI. Megrez-3b-omni. <https://huggingface.co/Infinigence/Megrez-3B-Omni/>, 2024.
- [4] 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.
- [5] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

- [6] 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. *Science China Information Sciences*, 67(12):220101, 2024.
- [7] 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.
- [8] 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.
- [9] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023.
- [10] Chee Kheng Chng, Yuliang Liu, Yipeng Sun, Chun Chet Ng, Canjie Luo, Zihan Ni, ChuanMing Fang, Shuaítáo Zhang, Junyu Han, Errui Ding, et al. Icdar2019 robust reading challenge on arbitrary-shaped text-rrc-art. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1571–1576. IEEE, 2019.
- [11] X.AI Corp. Grok-1.5 vision preview: Connecting the digital and physical worlds with our first multimodal model., 2024.
- [12] Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim M Alabdulmohsin, et al. Patch n’pack: Navit, a vision transformer for any aspect ratio and resolution. *Advances in Neural Information Processing Systems*, 36, 2024.
- [13] Guosheng Dong, Da Pan, Yiding Sun, Shusen Zhang, Zheng Liang, Xin Wu, Yanjun Shen, Fan Yang, Haoze Sun, Tianpeng Li, et al. Baichuanseed: Sharing the potential of extensive data collection and deduplication by introducing a competitive large language model baseline. *arXiv preprint arXiv:2408.15079*, 2024.
- [14] 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.
- [15] Yuning Du, Chenxia Li, Ruoyu Guo, Cheng Cui, Weiwei Liu, Jun Zhou, Bin Lu, Yehua Yang, Qiwen Liu, Xiaoguang Hu, et al. Pp-ocrv2: Bag of tricks for ultra lightweight ocr system. *arXiv preprint arXiv:2109.03144*, 2021.
- [16] 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.
- [17] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Li Ke, Sun Xing, Wu Yunsheng, and Ji Rongrong. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.
- [18] 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.
- [19] 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 and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385, 2024.
- [20] 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.
- [21] 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.
- [22] Anwen Hu, Haiyang Xu, Liang Zhang, Jiabo Ye, Ming Yan, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. mplug-docowl2: High-resolution compressing for ocr-free multi-page document understanding. *arXiv preprint arXiv:2409.03420*, 2024.

- [23] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, JinYeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyo Han, and Seunghyun Park. Ocr-free document understanding transformer. In *European Conference on Computer Vision (ECCV)*, 2022.
- [24] 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.
- [25] Atsunobu Kotani, Stefanie Tellex, and James Tompkin. Generating handwriting via decoupled style descriptors. In *European Conference on Computer Vision*, pages 764–780. Springer, 2020.
- [26] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36, 2024.
- [27] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models?, 2024.
- [28] Alex WC Lee, Jonathan Chung, and Marco Lee. Gnkh: a dataset for english handwriting in the wild. In *Document Analysis and Recognition—ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part IV 16*, pages 399–412. Springer, 2021.
- [29] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308, 2024.
- [30] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.
- [31] Xiaotong Li, Fan Zhang, Haiwen Diao, Yueze Wang, Xinlong Wang, and Ling-Yu Duan. Densefusion-1m: Merging vision experts for comprehensive multimodal perception. *arXiv preprint arXiv:2407.08303*, 2024.
- [32] 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.
- [33] 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. *CoRR*, abs/2311.06607, 2023.
- [34] 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.
- [35] Cheng-Lin Liu, Fei Yin, Da-Han Wang, and Qiu-Feng Wang. Casia online and offline chinese handwriting databases. In *2011 international conference on document analysis and recognition*, pages 37–41. IEEE, 2011.
- [36] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.
- [37] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024.
- [38] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.
- [39] 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? In *European conference on computer vision*, pages 216–233. Springer, 2025.
- [40] Yuan Liu, Le Tian, Xiao Zhou, Xinyu Gao, Kavio Yu, Yang Yu, and Jie Zhou. Points1. 5: Building a vision-language model towards real world applications. *arXiv preprint arXiv:2412.08443*, 2024.
- [41] Yuliang Liu, Zhang Li, Biao Yang, Chunyuan Li, Xucheng Yin, Cheng-lin Liu, Lianwen Jin, and Xiang Bai. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023.
- [42] Yuliang Liu, Biao Yang, Qiang Liu, Zhang Li, Zhiyin Ma, Shuo Zhang, and Xiang Bai. Textmonkey: An ocr-free large multimodal model for understanding document. *arXiv preprint arXiv:2403.04473*, 2024.

- [43] Keer Lu, Zheng Liang, Xiaonan Nie, Da Pan, Shusen Zhang, Keshi Zhao, Weipeng Chen, Zenan Zhou, Guosheng Dong, Wentao Zhang, et al. Datasculpt: Crafting data landscapes for llm post-training through multi-objective partitioning. *arXiv preprint arXiv:2409.00997*, 2024.
- [44] 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.
- [45] 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.
- [46] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022.
- [47] 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.
- [48] OpenAI. Hello gpt-4o, 2024.
- [49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [50] 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.
- [51] 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.
- [52] Qwen Team. Qwen2.5: A party of foundation models, September 2024.
- [53] Yuxiang Tuo, Wangmeng Xiang, Jun-Yan He, Yifeng Geng, and Xuansong Xie. Anytext: Multilingual visual text generation and editing. 2023.
- [54] 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.
- [55] 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.
- [56] 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.
- [57] Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, Jinrong Yang, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. Vary: Scaling up the vision vocabulary for large vision-language model. In *European Conference on Computer Vision*, pages 408–424. Springer, 2025.
- [58] Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun, Yuang Peng, et al. General ocr theory: Towards ocr-2.0 via a unified end-to-end model. 2024.
- [59] Cong Yao, Xiang Bai, Wenyu Liu, Yi Ma, and Zhuowen Tu. Detecting texts of arbitrary orientations in natural images. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1083–1090. IEEE, 2012.
- [60] Cong Yao, Jianan Wu, Xinyu Zhou, Chi Zhang, Shuchang Zhou, Zhimin Cao, and Qi Yin. Incidental scene text understanding: Recent progresses on icdar 2015 robust reading competition challenge 4. *arXiv preprint arXiv:1511.09207*, 2015.
- [61] 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.
- [62] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, et al. mplug-docowl: Modularized multimodal large language model for document understanding. *arXiv preprint arXiv:2307.02499*, 2023.
- [63] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023.

- [64] 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.
- [65] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024.
- [66] 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 lilm fine-tuning. *arXiv preprint arXiv:2409.20566*, 2024.
- [67] Kecheng Zheng, Yifei Zhang, Wei Wu, Fan Lu, Shuailei Ma, Xin Jin, Wei Chen, and Yujun Shen. Dreamlip: Language-image pre-training with long captions. In *ECCV*, 2024.