VCR: VISUAL CAPTION RESTORATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce Visual Caption Restoration (VCR), a novel vision-language task that challenges models to accurately restore partially obscured texts using pixel-level hints within images. This task stems from the observation that text embedded in images intrinsically differs from common visual elements and natural language due to the need to align the modalities of vision, text, and text embedded in images. While numerous works have integrated text embedded in images into visual question-answering tasks, approaches to these tasks generally rely on optical character recognition or masked language modeling, thus reducing the task to mainly text-based processing. However, text-based processing becomes ineffective in VCR as accurate text restoration depends on the combined information from provided images, context, and subtle cues from the tiny exposed areas of masked texts. We develop a pipeline to generate synthetic images for the VCR task using image-caption pairs, with adjustable caption visibility to control the task difficulty. With this pipeline, we construct a dataset for VCR called VCR-WIKI using images with captions from Wikipedia, comprising $2.11M$ English and $346K$ Chinese entities in both *easy* and *hard* configurations. Our results reveal that current vision language models significantly lag behind human performance in the VCR task, and merely fine-tuning the models on our dataset does not lead to notable improvements. We release VCR-WIKI and the data construction code to facilitate future research.

1 INTRODUCTION

Recent advances in large language models, such as ChatGPT (Ouyang et al., 2022; OpenAI et al., 2023) and Llama (Touvron et al., 2023), have spurred significant interest and progress in the field of vision-language models. With models like GPT-4V (OpenAI et al., 2023) and LLaVA (Liu et al., 2023a; 2024a; 2023b) blending textual and visual information, the intersection of computer vision and natural language processing has become a vibrant research frontier. These integrated models aim to leverage the potential of vision and language modalities to understand and interpret multimedia content more effectively.

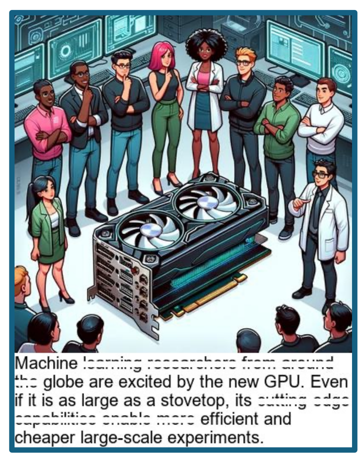

Figure 1: An example of the VCR task.

Amidst this evolving landscape, we introduce VCR, a novel vision-language task designed to challenge existing models uniquely. VCR challenges these models to restore obscured texts within images, which demands an intricate synthesis of text, vision, and text embedded in the image. The VCR task is grounded in two key insights: (1) text embedded within images, with its characteristics different from common visual elements, represents a distinct modality that requires careful alignment of vision, textual data, and the structure of written texts, and (2) neuroscience findings that suggest that humans are proficient in recognizing partially occluded objects through sophisticated visual and cognitive processes (Thinés et al., 2013; Pessoa et al., 1998; van Lier & Gerbino, 2015; Fyall et al., 2017; Li et al., 2023a). By leveraging these insights, VCR seeks to explore how well vision-language models can handle texts embedded within images, aligning visual elements and natural language to mimic human-like multimodal understanding and recognition.

The Visual Question Answering (VQA) task (Antol et al., 2015; Wang et al., 2018; Mishra et al., 2019b; Singh et al., 2019) has been a popular benchmark in assessing how well models align and interpret visual and linguistic information. Traditional VQA approaches, however, predominantly focus on direct queries about visible elements in images and do not address the nuanced relationship between textual content embedded within the image and the overall image context. This gap underscores the limited capabilities of current models in processing integrated visual-textual data, particularly when the textual component, which plays a critical role, is partially obscured or altered.

To address these limitations, our VCR task builds on the premise that effective text restoration from images requires an integrated understanding beyond the capabilities of current VQA benchmarks. For example, in extreme cases, models rely on existing Optical Character Recognition (OCR) system to extract text from documents (Singh et al., 2019; Borisyuk et al., 2018). The extracted text is then used as context for generating answers without a true semantic alignment between the text and the visual elements of the document. This approach, while effective in simple scenarios, falls short in more complex settings where text is intricately woven into the visual narrative of the image.

To develop the VCR task, in this work, we introduce a pipeline for generating synthetic images that allows for manipulation of the visibility of the textual components of the image. This not only enhances the challenge posed by the task but also provides a scalable way to adjust task difficulty. The resulting dataset, VCR-WIKI, comprises 2.11M English data and 346K Chinese data sourced from Wikipedia, featuring captions in both languages across 'easy' and 'hard' difficulty levels. Our evaluations indicate that existing vision-language models significantly underperform compared to human benchmarks, underscoring the need for novel model architectures and training paradigms specifically geared towards this complex intermodal alignment.

By releasing VCR-WIKI and the accompanying dataset construction code, we aim to stimulate further research in this area, encouraging the development of models that can more adeptly navigate the nuanced landscape of the restoration of text embedded in images. This effort aligns with the broader goal of advancing vision-language models to achieve a deeper, more intuitive understanding of multimedia content, bridging the gap between human and machine perception. The code in fully anonymous is available at https://anonymous.4open.science/r/VCR_anonymous/.

Contributions The main contributions of this paper are:

- **C1** Introduce the VCR task to challenge vision-language models to restore occluded texts in images.
- **C2** Develop a pipeline for generating synthetic images with embedded text that allows for adjusting the visibility of such text, thus providing a rich testing environment for VCR.
- **C3** Create and release VCR-WIKI, a dataset with multilingual captions from Wikipedia images, designed to benchmark vision-language models (VLMs) on text restoration tasks.
- **C4** Conduct empirical evaluations that show significant gaps between current models and human performance on the VCR task. This highlights the effectiveness of VCR for assessing advancements in VLMs and underscores the necessity for innovative model architectures and training techniques.

2 VCR TASK DESCRIPTION

In this section, we compare the VCR task with other existing tasks and aim to answer the following questions:

- **Q1** What is the difference between VCR and other visual reconstruction tasks?
- **Q2** Why should we care about VCR?

For better clarity, we define *text embedded in image (TEI)* as text incorporated within the image. The term *visual image (VI)* pertains to the portion of the image that excludes the text embedded in the image. The *string text (ST)* is not part of the image itself but is associated with it as a distinct textual element. It is usually the question prompt in the form of natural language, for example, 'What are the covered texts in the image? Please only guess the covered texts without outputting an explanation.'. Consequently, an element of a VCR task can be expressed as $(ST, (VI, TEI))$, where ST is represented as a string and both VI and TEI are presented in image form. This notation does not imply that VI and TEI can be physically separated into two distinct image components. Instead, this definition is adopted merely to facilitate a clearer explanation of the concepts involved. Please refer to Figure 3 for an illustration of VI, TEI, and ST.

A1 Existing tasks that are similar to VCR are the tasks of VQA and OCR. VQA takes as input images and a natural language question and generates a free-form response. As the ground-truth response is not unique, evaluating VQA poses a major challenge. In contrast to VQA, OCR is a task where the ground-truth responses are unique: OCR takes as input complete characters in image form and outputs a string representing the characters in the image, without considering the image context. Models pretrained with OCR are able to retrieve texts embedded in the input image, even if they are incomplete or vague. However, as the vagueness or occlusion of the textual components of the image increases, retrieving the original text without considering the remaining nontextual image context becomes harder, and OCR is no longer a good approach. VCR bridges the gap between OCR and VQA: it reconstructs the unique text found in the image while also considering the visual context of the rest of the image.

Figure 2 is an example VCR task in hard mode, and Figure 1 shows an example VCR task in the easy mode. Although humans can still fill the blanks easily in the hard mode, it is nearly impossible for models with only OCR capabilities to recover the covered texts without using the context. This is because the pixel-level hints of single characters no longer correspond to a unique solution.

A2 The proposed VCR task is significant in two aspects.

The first aspect of importance stems from discoveries in neuroscience about human cognitive abilities to recognize partially occluded objects (Fyall et al., 2017; Li et al., 2023a). Although existing models can recognize objects and texts in images, they often struggle with the complexity of occluded objects due to significant information loss in the images. In contrast, humans excel at filling in missing information using a combination of low-level visual processing and high-level cognitive functions, such as those managed by the prefrontal cortex. This cortical area is known to handle complex cognitive processes such as decision-making and memory retention, which are essential for integrating fragmented visual input into coherent objects. We believe that the occlusion restoration task serves as a probe that can effectively distinguish low-level recognition and high-level cognition involving reasoning. In addition, understanding these neural mechanisms can inspire new algorithms capable of mimicking human-like perception and interpretation in dynamic, real-world conditions where occlusion is common.

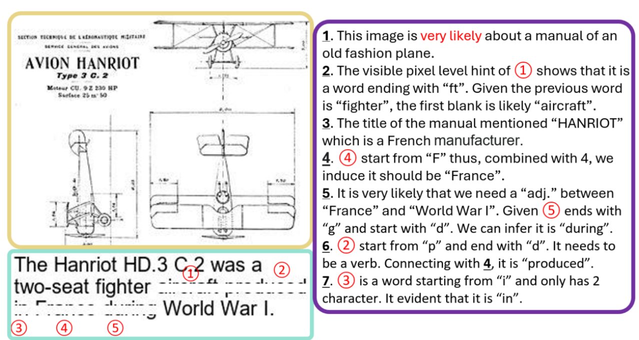

Figure 2: How humans would possibly solve a VCR task.

The second aspect underscores the unique challenge presented by the VCR task, distinguishing it significantly from existing benchmarks, such as traditional VQA or the occluded object restoration task. By occluding texts instead of common visual objects, VCR targets the models' text-image alignment capability, which is one of the major challenges for vision-language models. VCR mandates that models align textual and visual information in a manner that replicates human-like understanding involving the utilization of both textual and visual clues. This task requires a deep integration of visual (VI), embedded textual (TEI), and contextual interpretation across modalities, pushing beyond simple text extraction as performed in OCR tasks. In OCR, the focus is primarily on recognizing visible characters, often without the need to understand their contextual relevance within the image narrative. In contrast, VCR introduces complexity by requiring the model to use available partial texts and the visual context collaboratively to reconstruct the obscured content accurately. This not only tests the model's ability to process TEI-VI modalities effectively but also challenges it to maintain internal consistency, akin to human cognitive processes where context and visual clues guide understanding and response. Besides, the difficulty of the task can be adjusted by altering the extent of text occlusion, offering a scalable and flexible framework for systematically enhancing model capabilities in text-visual alignment and semantic comprehension. This rigorous testing ground will help evolve vision-language models to better grasp the nuanced interplay between text and imagery. We show an example of how humans would solve the VCR task in the "hard" difficulty in Figure 2.

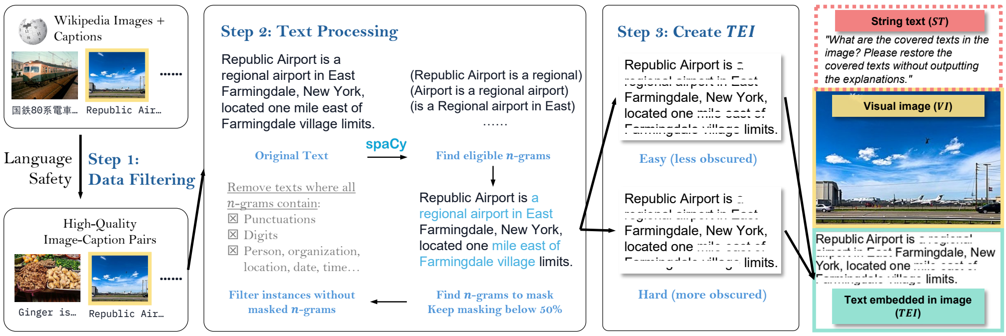

Figure 3: Illustration of the dataset creation pipeline for VCR-WIKI. visual image (VI), text embedded in image (TEI) and string text (ST) in an example of the English Hard configuration of VCR-WIKI. The solid line-enclosed contents (VI and TEI) are part of the image, whereas the dotted line-enclosed content (ST) is given separately from the image.

3 Dataset Creation

The VCR task requires aligning visual images (VI) with text embedded in images (TEI). Therefore, the dataset creation process relies on a set of highly correlated image-text pairs. We utilize the primary images and their corresponding captions from Wikipedia as the data source[1] to create VCR-WIKI, a Wikipedia-based VCR dataset. The pipeline for creating VCR-WIKI is shown in Figure 3. Before constructing the dataset, we first filter out instances with sensitive content, including NSFW and crime-related terms, to mitigate AI risk and biases.

The VCR-WIKI dataset is formatted as a VQA task, where each instance includes an image, a question, and a ground-truth answer. The images are synthesized from text-image pairs by stacking the image (VI) with its corresponding text description (TEI) vertically, mimicking the format of a captioned image. This stacked image is referred to as a stacked VI +TEI image. Each VI +TEI image is resized to a width of 300 pixels. To avoid excessive image height, we truncate TEI to a maximum of five lines. We filter the dataset to exclude instances with VI +TEI images exceeding 900 pixels in height to avoid drastic resolution changes during data pre-processing.

We use spaCy to randomly select 5-grams in the caption for masking. To ensure the restoration process is doable by a human without too much domain knowledge, the 5-grams do not contain numbers, person names, religious or political groups, facilities, organizations, locations, dates, and times labeled by spaCy. The total masked token does not exceed 50% of the tokens in the caption. We pick 5-grams for masking as it balances linguistic complexity and task feasibility, capturing meaningful grammatical structures while avoiding dataset reduction or overly simplified tasks observed with longer or shorter spans. We exclude instances that do not have any maskable 5-grams. The selected 5-grams are partially obscured by a white rectangle that reveals only the upper and lower parts of the text, with the proportion of coverage varying according to task difficulty. Furthermore, to assess the impact of VI on model performance, we create an ablation for each image, maintaining the resolution of the VI +TEI image, but retaining only the TEI part in the center of the image.

The VCR task involves a predefined question that prompts the model to produce the obscured text in the image. The ground-truth answer corresponds to the caption displayed in the uncovered portion of the stacked image. Due to the extensive availability of VLMs and a significant user base in both English and Chinese, we have chosen to develop the dataset in these two languages. For each language, we meticulously select the height of the masking rectangle to create two task variants: (1) an *easy* version, where the task is easy for native speakers but open-source OCR models almost always fail, and (2) a *hard* version, where the revealed part consists of only one to two pixels for the majority of letters or characters, yet the restoration task remains feasible for native speakers.

Because of the anonymous, policy, our dataset will be available in the camera-ready version. The VCR codebase provides options to mask complete sentences, specific PoS tags, and letter heights based on optional whitelists or blacklists of words, offering full flexibility for customizing the task.

[1]Datasource: `https://huggingface.co/datasets/wikimedia/wit_base`.

3.1 DATASET FORMAT AND STATISTICS

Table 1: Basic statistics of the dataset. Note that each language's Easy and Hard configurations share the same statistics. We report the mean, standard deviation, and the 5th and 95th percentile ($\eta_{.5}$ and $\eta_{.95}$) for the stacked image height and the number of obscured text spans. Unit is in pixels.

	# Train	# Val	# Test	VI +TEI Image Height				# Obscured Text Spans			
				Mean	SD	$\eta_{.5}$	$\eta_{.95}$	Mean	SD	$\eta_{.5}$	$\eta_{.95}$
English	2095733	5000	5000	375.52	106.01	253	564	1.62	0.63	1	3
Chinese	336448	5000	5000	360.44	102.76	239	562	2.06	0.94	1	4

The VCR dataset comprises four configurations: English Easy, English Hard, Chinese Easy and Chinese Hard. Each configuration can be further divided into training, validation, and test splits. The validation and test splits contain 5,000 entities each. The training set for English configurations and Chinese configurations contains $2,095,733$ and $336,448$ instances, respectively, which can be used for model continuous pretraining. We include more detailed statistics of the dataset in Table 1.

4 EXPERIMENTS

In this section, we report the experiment results of existing state-of-the-art vision-language models on our proposed VCR tasks. The fine-tuning and evaluation of open-source models are conducted on a mix of NVIDIA A100 80G and L40S 48G GPUs in an internal cluster.

4.1 MODELS

Closed-source Models. We evaluate several most advanced proprietary models with their provided APIs. The evaluated models include GPT-4o (gpt-4o-2024-0513), GPT-4 Turbo (gpt-4-turbo-2024-04-09), GPT-4V (gpt-4-1106-vision-preview) (Ouyang et al., 2022; OpenAI et al., 2023), Claude 3 Opus (claude-3-opus-20240229), Claude 3.5 Sonnet (claude-3-5-sonnet-20240620) (Anthropic, 2024), Gemini 1.5 pro (gemini-1.5-pro-001) (Team et al., 2024a), Reka Core (reka-core-20240501) (Team et al., 2024b), and Qwen-VL-Max (tested in May 2024) (Bai et al., 2023).

Open-source Models. We evaluate open-source models with the best performance on the OpenVLM Leaderboard[2] and state-of-the-art Chinese VLM models. The evaluated models include Cambrian-1 (Tong et al., 2024), CogVLM2-19B (Hong et al., 2024), DeepSeek-VL-7B-Chat (Lu et al., 2024), DocOwl-1.5-Omni (Hu et al., 2024a), Idefics3-8B (Laurençon et al., 2024), InternLM-XComposer2-VL-7B (Dong et al., 2024a), InternLM-XComposer2.5-VL (Zhang et al., 2024), InternLM-XComposer2-VL-4K (Dong et al., 2024b), Llama-3.2 (Dubey et al., 2024), MiniCPM-V2.5 (Hu et al., 2024b), Monkey (Liu et al., 2024b; Li et al., 2023b), Pixtral (Mistral, 2024), Qwen-VL-Chat (Bai et al., 2023), Qwen2-VL (Wang et al., 2024), and Yi-VL (01.AI et al., 2024). Out of these models, Cambrian-1, Idefics3 and Llama-3.2 are English-only models, and CogVLM2-Llama3-19B-Chat has its Chinese variant, CogVLM2-Llama3-19B-Chinese-Chat. Please refer to Table 6 for the model specifications.

Finetuned Models. To test whether VLMs can learn to conduct VCR via fine-tuning, we select three models from the open-sourced models: CogVLM2-Llama3-19B-Chat, MiniCPM-Llama3-V2.5, and Qwen2-VL-7B-Instruct, and fine-tune them on a subset of VCR's training set.

More specifically, we fine-tune CogVLM2-Llama3-19B-Chat, MiniCPM-Llama3-V2.5, and Qwen2-VL-7B-Instruct in the English Hard configuration, and CogVLM2-Llama3-19B-Chinese-Chat, MiniCPM-Llama3-V2.5, and Qwen2-VL-7B-Instruct on the Chinese Hard configuration. The models are finetuned using LoRA (Hu et al., 2022) with $r = 8$ and $\alpha = 32$. We adopt the schedule-free

[2]We selected the highest-performing open-source models with fewer than 40 billion parameters from the OpenVLM Leaderboard as of May 23, 2024 and their later versions. Details are available at `https://huggingface.co/spaces/opencompass/open_vlm_leaderboard`.

AdamW optimizer (Defazio et al., 2024) with a learning rate $2e-4$. The effective batch size is 64. Each model is trained on the first 16,000 examples of the training set for 1 epoch. All fine-tuning experiments are performed on a single node with 4 NVIDIA L40S 48G GPUs.

4.2 METRICS

We measure the quality of the model's restoration of each masked n-gram (where $n = 5$ in our setting, as specified in Section 3). Due to the variability of different models' outputs, for each masked n-gram $m \in \mathbb{V}_e^n$, where $\mathbb{V}_e$ is the vocabulary of the evaluation tokenizer[3], we extract the most similar n-gram $\hat{m} \in \mathbb{V}_e^n$ with the least edit distance in the model's generation.

We report the two metrics below in our experiment section to measure the restoration quality: **Exact Match** $(EM) \equiv EM(m, \hat{m}) = \mathbb{I}(m = \hat{m})$, which measures whether the restored n-gram $\hat{m}$ totally matches the ground-truth m; and **Jaccard Index** $(J) \equiv \frac{|S(m) \cap S(\hat{m})|}{|S(m) \cup S(\hat{m})|}$, which measures the similarity of $\hat{m}$ and m as bag-of-words.

4.3 RELATIONSHIP TO OTHER BENCHMARKS.

We evaluated 38 Vision-Language Models (VLMs) across 23 different benchmarks, using the VLM performance scores as features of each benchmark to compute a correlation matrix. Based on this matrix, in Figure 4, we applied K-Means clustering and visualized the results in 2D by plotting the first two principal components derived from the correlation matrix rows for each benchmark. Additionally, we provide a heatmap of the correlation matrix in Appendix B.

$\text{VCR}_{\text{ZH, EASY}}$ and $\text{VCR}_{\text{ZH, HARD}}$ were excluded from these processes due to the limited availability of VLMs that support Chinese. According to Figure 4, $\text{VCR}_{\text{EN, EASY}}$ shows a tentative similarity to ChartQA and TextVQA, as all three benchmarks evaluate the ability to extract and reason about text from natural images and documents. However, $\text{VCR}_{\text{EN, EASY}}$ does not exhibit significant similarity to the other benchmarks. Meanwhile, $\text{VCR}_{\text{EN, HARD}}$ stands apart from all other benchmarks. We attribute this to the fact that $\text{VCR}_{\text{EN, HARD}}$ emphasizes caption recovery with minimal pixel-level information, a skill not tested by any of the other benchmarks. Therefore, we assert that the VCR series benchmarks assess unique aspects of VLMs that are not covered by any of the other benchmarks in our evaluation.

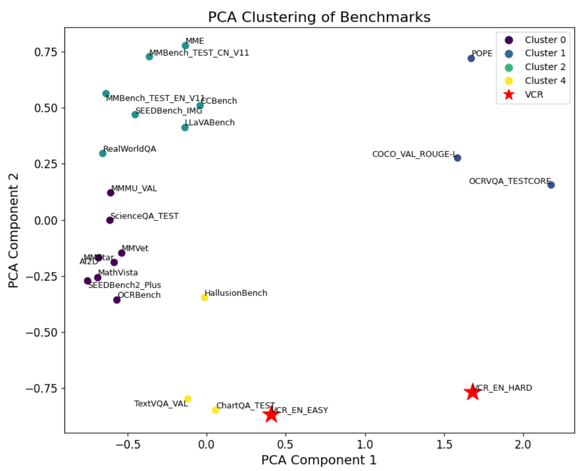

Figure 4: Projection of benchmarks onto the first two principal components derived from the correlation matrix of VLM performance scores. Each point represents a benchmark, and proximity indicates higher similarity based on model performance correlation.

4.4 EXPERIMENTAL RESULTS

Table 2 presents the exact match scores and Jaccard indices for our evaluations. Additionally, Figure 5 illustrates a comparative analysis of our fine-tuned models against their base counterparts across all four VCR settings. We also provide experiment results on smaller test sets of VCR-WIKI for faster benchmarking. Please refer to Table 4 and 5 in the Appendix.

The VCR task and the VCR-WIKI dataset are specifically designed to advance the development of future VLMs. In this section, we conduct a detailed analysis of the performance of current state-of-the-art (SOTA) models on the VCR task, highlighting key insights through comparative evaluations.

[3] We utilize spaCy's `en_core_web_sm`'s and `zh_core_web_sm`'s tokenizer for English and Chinese evaluation, respectively.

VCR Remains a Challenging Task for SOTA VLMs. Despite high-performing models like Qwen2-VL excelling in $\text{VCR}_{\text{EN, EASY}}$ and $\text{VCR}_{\text{EN, HARD}}$ settings, most recent models struggle significantly, especially under harder settings where metrics approach zero. This highlights not only the inherent difficulty of the VCR task but also that subpar performance on VCR-WIKI stems from *a lack of reasoning capabilities or sufficient text-image alignment rather than unfamiliarity with the underlying text*, as many VLMs are pretrained on similar data. These results emphasize the need for advancements in VLM designs to achieve robust performance across all settings.

Enhanced OCR Capabilities Do Not Necessarily Translate to Improved VCR Performance. Our analysis reveals that models proficient in OCR, such as InternLM-XComposer2-VL, and those excelling in image document understanding, like DocOwl 1.5 and Monkey, demonstrate subpar performance across most VCR settings. This discrepancy suggests that while these models can accurately recognize text within images, they lack the advanced reasoning capabilities required to effectively interpret and utilize this information within the context of the VCR task. This finding also highlights a fundamental distinction between OCR tasks and the more complex VCR task.

Language-Specific Performance: Need for Enhanced Multilingual Capabilities. A significant performance degradation is observed when models are evaluated on Chinese configurations, despite assertions of basic English-Chinese bilingual capabilities. This decline is particularly surprising given the logographic nature of Chinese characters, which theoretically offer higher recognizability compared to alphabetic scripts (Wu et al., 2024; Zhao et al., 2022). These results indicate a critical need for targeted improvements in multilingual support to ensure consistent performance across different languages.

Model Size Does Not Guarantee Superior Performance. Comparative analysis between Llama-3.2-11B and Llama-3.2-90B models reveals that both exhibit similar performance levels on the $\text{VCR}_{\text{EN, EASY}}$ and $\text{VCR}_{\text{EN, HARD}}$ settings. This observation suggests that merely increasing model size does not inherently enhance VCR performance. Instead, advancements in the cognitive abilities of models, achieved through improved training strategies, reasoning frameworks, or architectural innovations, are essential for meaningful performance gains in VCR tasks.

Model Resolution Is Not Directly Correlated with Performance Enhancement. InternLM-XComposer2-VL-4K, despite its higher resolution, demonstrates significantly lower performance on the $\text{VCR}_{\text{EN, EASY}}$ setting compared to its lower-resolution counterparts. This decline may result from more aggressive image partitioning strategies that disrupt the spatial continuity of text or from more intensive pixel or token compression techniques that lead to the loss of crucial local details. Both factors are critical for the accurate interpretation required in VCR tasks.

Inclusion of VI Input Images Negatively Impacts Performance. The addition of VI generally results in negative performance changes ($\Delta < 0$), indicating that the image information is not being effectively leveraged by the models. This negative impact may stem from the importance of key information locations, which could be compromised by image partitioning strategies that fail to preserve spatial relationships essential for accurate reasoning.

Model Design Influences Performance Gains from VCR-WIKI Finetuning. Finetuning on the VCR-WIKI dataset yields varying performance improvements across different model designs. Specifically: 1) **CogVLM2** demonstrates substantial performance enhancements across all four settings after finetuning, indicating that its overall design may be well-aligned with the image-text reasoning demands of VCR, though further empirical validation is needed to substantiate this hypothesis. 2) **MiniCPM-V2.5** shows only marginal performance increases from an already low baseline, particularly in the $\text{VCR}_{\text{ZH, EASY}}$ and $\text{VCR}_{\text{ZH, HARD}}$ settings. This limited improvement indicates potential design limitations that hinder its ability to effectively address the complexities of the VCR task. 3) **Qwen2-VL-7B** maintains relatively high performance both before and after finetuning, implying that the model is sufficiently pre-trained on relevant tasks to perform well on the VCR task without extensive additional training.

We hope that through controlled variables, the VCR-WIKI dataset delivers *fully comparable* results across languages, difficulty levels, image inclusion, and fine-tuning stages. Each comparison is intended to guide specific and targeted improvements in VLM development.

Table 2: Performance of vision language models on the VCR task in English (EN) and Chinese (ZH), for easy and hard modes. We label the best result of each setting with **bold fonts**, and the best open-source model with underline. A superscript of * marks that the model was released after the initial public release of the VCR-WIKI dataset (June 10, 2024). Subscripts show bootstrapped standard deviation.

Language	Mode	Open/closed source	Model name	Model size	Exact match (%) ↑			Jaccard index (%) ↑		
					$VI+TEI$	TEI	Δ	$VI+TEI$	TEI	Δ
English	Easy	Closed	Claude 3 Opus	-	$62.0_{0.13}$	$77.0_{0.5}$	-15	$77.67_{0.32}$	$88.41_{0.39}$	-10.74
			Claude 3.5 Sonnet	-	$63.85_{1.71}$	$72.8_{1.56}$	-8.94	$74.65_{1.33}$	$83.48_{1.14}$	-8.83
			Gemini 1.5 Pro	-	$62.73_{1.66}$	$82.98_{1.3}$	-20.25	$77.71_{1.21}$	$91.56_{0.76}$	-13.85
			GPT-4 Turbo	-	$78.74_{0.13}$	$81.94_{0.25}$	-3.2	$88.54_{0.24}$	$92.18_{0.3}$	-3.65
			GPT-4o	-	$\mathbf{91.55_{0.29}}$	$\mathbf{94.56_{0.13}}$	-3.01	$\mathbf{96.44_{0.11}}$	$\mathbf{97.76_{0.06}}$	-1.32
			GPT-4V	-	$52.04_{0.24}$	$37.86_{0.22}$	14.17	$65.36_{0.39}$	$54.13_{0.41}$	11.23
			Qwen-VL-Max	-	$76.8_{0.5}$	$85.53_{0.19}$	-8.74	$85.71_{0.28}$	$91.45_{0.29}$	-5.74
			Reka Core	-	$66.46_{1.64}$	$78.51_{1.42}$	-12.05	$84.23_{0.86}$	$90.45_{0.7}$	-6.22
		Open	Cambrian-1*	34B	$79.69_{0.43}$	$81.28_{0.43}$	-1.59	$89.27_{0.28}$	$92.54_{0.19}$	-3.27
			CogVLM2	19B	$83.25_{0.07}$	$78.29_{0.04}$	4.96	$89.75_{0.1}$	$88.07_{0.08}$	1.68
			DeepSeek-VL	1.3B	$23.04_{0.05}$	$31.09_{0.12}$	-8.04	$46.84_{0.07}$	$52.36_{0.06}$	-5.52
			DeepSeek-VL	7B	$38.01_{0.12}$	$45.94_{0.1}$	-7.93	$60.02_{0.15}$	$64.72_{0.04}$	-4.7
			DocOwl-1.5-Omni	8B	$0.84_{0.01}$	$1.55_{0.02}$	-0.71	$13.34_{0.03}$	$14.62_{0.04}$	-1.28
			Idefics3*	8B	$25.99_{0.48}$	$31.43_{0.51}$	-5.44	$47.22_{0.42}$	$54.00_{0.39}$	-6.78
			InternLM-XComposer2-VL	7B	$46.64_{0.1}$	$46.4_{0.11}$	0.24	$70.99_{0.1}$	$72.14_{0.07}$	-1.14
			InternLM-XComposer2-VL-4K	7B	$5.32_{0.24}$	$3.71_{0.21}$	1.60	$22.14_{0.28}$	$18.78_{0.25}$	3.37
			InternLM-XComposer2.5-VL*	7B	$41.35_{0.55}$	$25.37_{0.51}$	15.97	$63.04_{0.42}$	$49.95_{0.41}$	13.09
			InternVL-V2*	40B	$84.67_{0.40}$	$87.71_{0.37}$	-3.04	$92.64_{0.22}$	$95.10_{0.16}$	-2.47
			Llama-3.2*	11B	$79.85_{0.45}$	$67.53_{0.53}$	12.32	$90.58_{0.22}$	$81.11_{0.33}$	9.47
			Llama-3.2*	90B	$80.54_{0.43}$	$71.05_{0.51}$	9.48	$89.81_{0.26}$	$84.22_{0.30}$	5.59
			MiniCPM-V2.5	8B	$31.81_{0.08}$	$40.05_{0.09}$	-8.25	$53.24_{0.1}$	$63.2_{0.1}$	-9.96
			Monkey	7B	$50.66_{0.1}$	$56.2_{0.08}$	-5.54	$67.6_{0.09}$	$72.82_{0.08}$	-5.22
			Pixtral*	12B	$18.41_{0.42}$	$11.60_{0.36}$	6.81	$41.25_{0.37}$	$31.60_{0.33}$	9.65
			Qwen-VL	7B	$49.71_{0.17}$	$52.15_{0.15}$	-2.44	$69.94_{0.07}$	$72.28_{0.08}$	-2.34
			Qwen2-VL*	7B	$\underline{89.70_{0.34}}$	$\underline{93.44_{0.26}}$	-3.74	$\underline{93.84_{0.24}}$	$\underline{97.47_{0.12}}$	-3.62
			Yi-VL	34B	$0.82_{0.03}$	$1.61_{0.04}$	-0.79	$5.59_{0.04}$	$7.72_{0.03}$	-2.13
			Yi-VL	6B	$0.75_{0.01}$	$1.65_{0.01}$	-0.9	$5.54_{0.02}$	$7.76_{0.03}$	-2.22
	Hard	Closed	Claude 3 Opus	-	$37.8_{0.28}$	$50.0_{0.33}$	-12.2	$57.68_{0.8}$	$70.16_{0.64}$	-12.48
			Claude 3.5 Sonnet	-	$41.74_{1.69}$	$44.72_{1.78}$	-2.98	$56.15_{1.46}$	$58.54_{1.6}$	-2.4
			Gemini 1.5 Pro	-	$28.07_{1.58}$	$38.76_{1.68}$	-10.68	$51.9_{1.22}$	$59.62_{1.27}$	-7.72
			GPT-4 Turbo	-	$45.15_{0.28}$	$48.64_{0.57}$	-3.5	$65.72_{0.25}$	$67.86_{0.2}$	-2.14
			GPT-4o	-	$73.2_{0.16}$	$\mathbf{82.43_{0.17}}$	-9.22	$\mathbf{86.17_{0.21}}$	$\mathbf{92.01_{0.2}}$	-5.84
			GPT-4V	-	$25.83_{0.44}$	$14.95_{0.3}$	10.87	$44.63_{0.48}$	$30.08_{0.67}$	14.56
			Qwen-VL-Max	-	$41.65_{0.32}$	$52.72_{0.2}$	-11.07	$61.18_{0.35}$	$70.19_{0.37}$	-9.01
			Reka Core	-	$6.71_{0.89}$	$11.18_{1.15}$	-4.47	$25.84_{0.95}$	$35.83_{1.05}$	-9.99
		Open	Cambrian-1*	34B	$27.20_{0.48}$	$29.68_{0.50}$	-2.48	$50.04_{0.40}$	$55.66_{0.39}$	-5.62
			CogVLM2	19B	$37.98_{0.18}$	$17.68_{0.06}$	20.3	$59.99_{0.05}$	$39.69_{0.03}$	20.3
			DeepSeek-VL	1.3B	$0.16_{0.01}$	$0.39_{0.02}$	-0.23	$11.89_{0.02}$	$11.47_{0.03}$	0.42
			DeepSeek-VL	7B	$1.0_{0.02}$	$1.75_{0.03}$	-0.75	$15.9_{0.08}$	$17.2_{0.04}$	-1.3
			DocOwl-1.5-Omni	8B	$0.04_{0.0}$	$0.02_{0.0}$	0.01	$7.76_{0.01}$	$7.74_{0.02}$	0.03
			Idefics3*	8B	$0.60_{0.08}$	$0.37_{0.07}$	0.23	$10.37_{0.15}$	$9.59_{0.13}$	0.79
			InternLM-XComposer2-VL	7B	$0.7_{0.01}$	$0.92_{0.01}$	-0.22	$12.51_{0.02}$	$13.23_{0.02}$	-0.72
			InternLM-XComposer2-VL-4K	7B	$0.21_{0.05}$	$0.18_{0.05}$	0.02	$9.52_{0.12}$	$9.52_{0.12}$	-0.00
			InternLM-XComposer2.5-VL*	7B	$0.93_{0.11}$	$1.11_{0.11}$	-0.18	$13.82_{0.16}$	$14.72_{0.18}$	-0.89
			InternVL-V2*	40B	$13.10_{0.37}$	$19.16_{0.44}$	-6.06	$33.64_{0.36}$	$41.35_{0.39}$	-7.71
			Llama-3.2*	11B	$14.09_{0.40}$	$6.92_{0.27}$	7.17	$35.26_{0.36}$	$26.35_{0.29}$	8.90
			Llama-3.2*	90B	$14.91_{0.40}$	$13.06_{0.37}$	1.85	$35.44_{0.35}$	$34.44_{0.35}$	1.00
			MiniCPM-V2.5	8B	$1.41_{0.03}$	$1.96_{0.02}$	-0.55	$11.94_{0.02}$	$13.37_{0.04}$	-1.43
			Monkey	7B	$1.96_{0.04}$	$2.43_{0.03}$	-0.48	$14.02_{0.03}$	$14.11_{0.03}$	-0.09
			Pixtral*	12B	$0.44_{0.08}$	$0.64_{0.09}$	-0.20	$10.99_{0.13}$	$11.45_{0.15}$	-0.46
			Qwen-VL	7B	$2.0_{0.03}$	$2.32_{0.03}$	-0.32	$15.04_{0.05}$	$14.27_{0.05}$	0.77
			Qwen2-VL*	7B	$\underline{\mathbf{74.32_{0.47}}}$	$\underline{75.20_{0.49}}$	-0.88	$\underline{85.47_{0.30}}$	$\underline{87.63_{0.27}}$	-2.15
			Yi-VL	34B	$0.07_{0.0}$	$0.05_{0.0}$	0.02	$4.31_{0.02}$	$5.89_{0.02}$	-1.58
			Yi-VL	6B	$0.06_{0.0}$	$0.04_{0.0}$	0.02	$4.46_{0.02}$	$5.91_{0.01}$	-1.46
Chinese	Easy	Closed	Claude 3 Opus	-	$0.9_{0.3}$	$1.0_{0.31}$	-0.1	$11.5_{0.48}$	$10.0_{0.49}$	1.49
			Claude 3.5 Sonnet	-	$1.0_{0.31}$	$0.8_{0.28}$	0.2	$7.54_{0.54}$	$7.5_{0.51}$	0.03
			Gemini 1.5 Pro	-	$1.1_{0.32}$	$0.5_{0.22}$	0.6	$11.1_{0.56}$	$11.47_{0.48}$	-0.37
			GPT-4o	-	$14.87_{1.14}$	$22.46_{1.35}$	-7.58	$39.05_{0.99}$	$48.24_{1.09}$	-9.19
			GPT-4 Turbo	-	$0.2_{0.14}$	$0.1_{0.1}$	0.1	$8.42_{0.36}$	$6.97_{0.29}$	1.45
			Qwen-VL-Max	-	$6.34_{0.08}$	$9.92_{0.09}$	-3.58	$13.45_{0.41}$	$22.86_{0.46}$	-9.42
			Reka Core	-	$0.0_{0.0}$	$0.0_{0.0}$	0	$3.43_{0.26}$	$3.15_{0.2}$	0.28
		Open	CogVLM2-Chinese	19B	$33.24_{0.04}$	$30.7_{0.07}$	2.54	$57.57_{0.06}$	$53.66_{0.04}$	3.91
			DeepSeek-VL	1.3B	$0.0_{0.0}$	$0.0_{0.0}$	0	$6.56_{0.01}$	$3.17_{0.02}$	3.4
			DeepSeek-VL	7B	$0.0_{0.0}$	$0.0_{0.0}$	0	$4.08_{0.01}$	$6.84_{0.01}$	-2.76
			DocOwl-1.5-Omni	8B	$0.0_{0.0}$	$0.0_{0.0}$	0	$1.14_{0.01}$	$3.38_{0.01}$	-2.23
			InternLM-XComposer2-VL	7B	$0.27_{0.01}$	$0.23_{0.01}$	0.04	$12.32_{0.02}$	$12.28_{0.03}$	0.04
			InternLM-XComposer2-VL-4K	7B	$0.46_{0.07}$	$0.46_{0.07}$	0.00	$12.31_{0.14}$	$13.37_{0.14}$	-1.05
			InternLM-XComposer2.5-VL*	7B	$0.46_{0.07}$	$0.58_{0.08}$	-0.12	$12.97_{0.16}$	$14.99_{0.17}$	-2.01
			InternVL-V2*	40B	$22.09_{0.41}$	$17.26_{0.39}$	4.84	$47.62_{0.34}$	$37.93_{0.35}$	9.69
			MiniCPM-V2.5	8B	$4.1_{0.02}$	$5.05_{0.08}$	-0.95	$18.03_{0.07}$	$22.94_{0.04}$	-4.9
			Monkey	7B	$0.62_{0.01}$	$1.44_{0.01}$	-0.82	$8.34_{0.06}$	$10.95_{0.03}$	-2.61
			Qwen-VL	7B	$0.04_{0.01}$	$0.0_{0.0}$	0.04	$1.5_{0.01}$	$0.34_{0.01}$	1.15
			Qwen2-VL*	7B	$\underline{\mathbf{59.94_{0.49}}}$	$\underline{\mathbf{67.48_{0.47}}}$	-7.54	$\underline{\mathbf{76.95_{0.32}}}$	$\underline{\mathbf{82.63_{0.28}}}$	-5.67
			Yi-VL	34B	$0.0_{0.0}$	$0.0_{0.0}$	0	$4.44_{0.01}$	$1.8_{0.01}$	2.64
			Yi-VL	6B	$0.0_{0.0}$	$0.0_{0.0}$	0	$4.37_{0.01}$	$1.76_{0.0}$	2.6
	Hard	Closed	Claude 3 Opus	-	$0.3_{0.18}$	$0.1_{0.1}$	0.2	$9.22_{0.38}$	$8.09_{0.33}$	1.13
			Claude 3.5 Sonnet	-	$0.2_{0.15}$	$0.0_{0.0}$	0.2	$4.0_{0.33}$	$2.37_{0.23}$	1.63
			Gemini 1.5 Pro	-	$0.7_{0.26}$	$0.5_{0.23}$	0.2	$11.82_{0.51}$	$11.75_{0.44}$	0.07
			GPT-4o	-	$2.2_{0.47}$	$1.8_{0.4}$	0.4	$22.72_{0.67}$	$22.89_{0.65}$	-0.17
			GPT-4 Turbo	-	$0.0_{0.0}$	$0.2_{0.13}$	-0.2	$8.58_{0.3}$	$6.87_{0.28}$	1.72
			Qwen-VL-Max	-	$0.89_{0.06}$	$1.38_{0.1}$	-0.49	$5.4_{0.19}$	$12.29_{0.18}$	-6.89
			Reka Core	-	$0.0_{0.0}$	$0.0_{0.0}$	0	$3.35_{0.23}$	$2.97_{0.2}$	0.38
		Open	CogVLM2-Chinese	19B	$1.34_{0.03}$	$2.67_{0.02}$	-1.32	$17.35_{0.03}$	$19.51_{0.03}$	-2.16
			DeepSeek-VL	1.3B	$0.0_{0.0}$	$0.0_{0.0}$	0	$6.46_{0.01}$	$3.22_{0.02}$	3.24
			DeepSeek-VL	7B	$0.0_{0.0}$	$0.0_{0.0}$	0	$5.11_{0.01}$	$7.21_{0.01}$	-2.1
			DocOwl-1.5-Omni	8B	$0.0_{0.0}$	$0.0_{0.0}$	0	$1.37_{0.01}$	$4.07_{0.02}$	-2.7
			InternLM-XComposer2-VL	7B	$0.07_{0.01}$	$0.09_{0.0}$	-0.02	$8.97_{0.02}$	$8.51_{0.01}$	0.46
			InternLM-XComposer2-VL-4K	7B	$0.05_{0.02}$	$0.05_{0.02}$	0.00	$7.67_{0.10}$	$7.72_{0.10}$	-0.04
			InternLM-XComposer2.5-VL*	7B	$0.11_{0.04}$	$0.12_{0.04}$	-0.01	$10.95_{0.11}$	$11.43_{0.12}$	-0.48
			InternVL-V2*	40B	$0.48_{0.07}$	$0.74_{0.08}$	-0.26	$12.57_{0.14}$	$13.31_{0.15}$	-0.74
			MiniCPM-V2.5	8B	$0.09_{0.0}$	$0.08_{0.0}$	0.01	$7.39_{0.02}$	$7.89_{0.01}$	-0.5
			Monkey	7B	$0.12_{0.01}$	$0.07_{0.0}$	0.05	$6.36_{0.01}$	$6.68_{0.03}$	-0.32
			Qwen-VL	7B	$0.01_{0.0}$	$0.01_{0.0}$	0	$1.17_{0.01}$	$0.12_{0.0}$	1.06
			Qwen2-VL*	7B	$\underline{\mathbf{18.33_{0.37}}}$	$\underline{\mathbf{27.58_{0.44}}}$	-9.26	$\underline{\mathbf{43.55_{0.34}}}$	$\underline{\mathbf{54.24_{0.34}}}$	-10.69
			Yi-VL	34B	$0.0_{0.0}$	$0.0_{0.0}$	0	$4.12_{0.0}$	$1.81_{0.01}$	2.31
			Yi-VL	6B	$0.0_{0.0}$	$0.0_{0.0}$	0	$4.0_{0.01}$	$1.88_{0.01}$	2.12

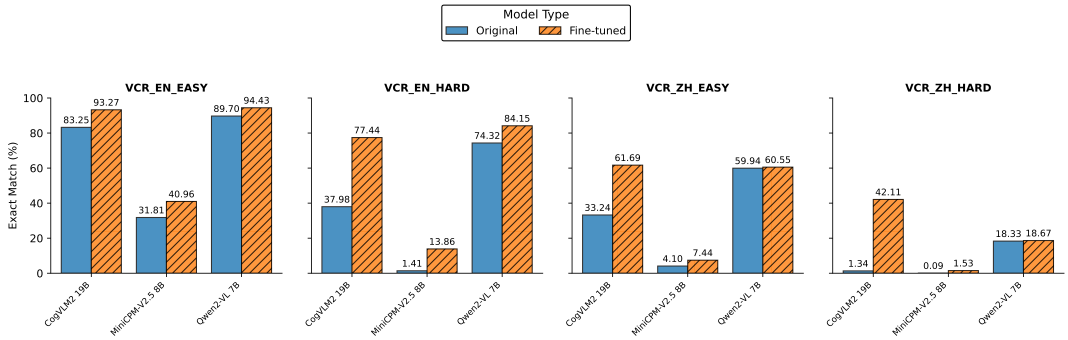

Figure 5: Exact Match performance on VCR before and after LoRA fine-tuning for selected models. The results demonstrate varying degrees of improvement across different tasks and models, highlighting the heterogeneous responses to fine-tuning.

4.5 Human Evaluation

We recruited 7 volunteers to perform human evaluation on a subset of the samples from our datasets. Two out of the seven evaluators are native English speakers, while five are native Chinese speakers who are also fluent in English[4]. All volunteers have earned postgraduate degrees, majoring in one of the following fields: biology, statistics, computer science, and economics. The evaluations were conducted on a voluntary basis, and participants received no rewards.

We gave the volunteers the following instructions: (1) We asked the volunteers to focus on the puzzles. Each example in the hard collection may require 30 seconds to 2 minutes of focused attention, and (2) we asked the volunteers to utilize the context rather than directly brute-force the puzzle.

Every sample is solved by at least 3 volunteers. In English, we release the exact match score in 2 splits: all errors counted (All), and only counting errors not related to dates and person names (Filtered).

The human evaluation results are shown in Table 3. Although current SOTA models suffer from the challenge, fluent speakers can easily achieve more than 90 percent accuracy across difficulties. Please refer to Table 4 to compare all models with human evaluation results using the same test cases.

Table 3: Human evaluation results on the VCR task in terms of exact matches. N is the number of puzzles in each language.

	VCR$_{\text{EN, EASY}}$ (N = 169)		**VCR$_{\text{EN, HARD}}$ (N = 169)**		**VCR$_{\text{ZH, EASY}}$ (N = 188)**		**VCR$_{\text{ZH, HARD}}$ (N = 188)**	
	Mean (%)	**SD (%)**	**Mean (%)**	**SD (%)**	**Mean (%)**	**SD (%)**	**Mean (%)**	**SD (%)**
All	96.65	0.34	91.12	1.18	98.58	0.31	91.84	0.81
Filtered	98.62	0.34	97.63	2.13	99.47	0.00	96.63	1.11

5 Related Work

Visual Question Answering (VQA). Several datasets have been proposed for visual question answering VQA Antol et al. (2015); Zhang et al. (2016); Goyal et al. (2017); Mishra et al. (2019b). FVQA Wang et al. (2018) and OK-VQAMarino et al. (2019) are datasets about knowledge-based visual question answering and contains questions that necessitate the usage of external knowledge resources. CLEVR Johnson et al. (2017) is a synthetic VQA dataset that mainly focuses on visual reasoning abilities. Recognizing the need to develop VQA models that can understand text, Text-VQA Singh et al. (2019); Biten et al. (2019); Mishra et al. (2019a); Wang et al. (2020) aims to read and reason about texts embedded within images in the context of image-question answering. Several datasets Singh et al. (2019); Biten et al. (2019); Mishra et al. (2019a) have been developed for the Text-VQA task, such as the TextVQA dataset Singh et al. (2019) and the ST-VQA dataset Biten et al.

[4]The TOEFL scores of the non-native English-speaking participants range from 102/120 to 112/120.

(2019) on natural images, the OCR-VQA dataset Mishra et al. (2019b) on book or movie covers, the InfographicVQA Mathew et al. (2022) dataset on infographics, and the DocVQA dataset Mathew et al. (2021) on document images.

Vision Language Model. Vision-language models are designed for tasks that involve understanding and generating content from images and text Sun et al. (2023); Liu et al. (2023b); Laurençon et al. (2023; 2024). For example, models have been developed to combine Llama3 with advanced vision-language processing capabilities to handle complex multimodal tasks Yu et al. (2024); Xu et al. (2024); Hu et al. (2023); Yu et al. (2023); Wang et al. (2023b); Dong et al. (2024a). Qwen-VL Bai et al. (2023) enhances visual-linguistic representations for more accurate contextual interpretations, while OpenGVLab-InternVL-Chat Chen et al. (2023; 2024b) merges the InternVL framework with interactive chat capabilities. These studies typically employ a multimodal encoder (Radford et al., 2021; Zhai et al., 2023; Wu et al., 2022) to process multimodal data, which is then mapped to the same input space of the language model. General-purpose models such as the GPT-4 series models (Ouyang et al., 2022; OpenAI et al., 2023), the Claude series models (Anthropic, 2024), the Gemini series models (Team et al., 2024a) and the Reka series models (Team et al., 2024b) have also been adapted for vision-language tasks, demonstrating strong performance in multimodal tasks. Finally, DocLLM Wang et al. (2023a) specializes in document understanding by integrating visual and textual data to enhance the interpretation and generation of document-related content. These models collectively represent significant advancements in vision-language integration, contributing unique capabilities and enhancements to the understanding and generation of multimodal information.

Optical Character Recognition (OCR). OCR Nagy (2000) and its subproblems Howe (2013); Smith (1995); Shafait et al. (2008); Frinken et al. (2011) have been well-studied in the literature in the constrained setting. However, classical OCR methods often cannot perform well on images captured in the wild in an unconstrained setting. Many new methods have been developed for advancing scene-text recognition on camera-captured images Bissacco et al. (2013); Gupta et al. (2016); Huang et al. (2014); Jaderberg et al. (2014); Wang et al. (2012); Shi et al. (2017); Zhou et al. (2017); Lee & Osindero (2016). In addition to the detection and recognition of OCR tasks, visual question answering has emerged as an important downstream task in the OCR literature. With the development of Text-VQA, new methods for improving the reading abilities in VQA utilizing OCR have been proposed. For example, LoRRA Singh et al. (2019) extends a VQA model Pythia Jiang et al. (2018) with an OCR module to better handle Text-VQA tasks. TAP Yang et al. (2021) incorporates scene texts that are generated from OCR engines during pretraining to further improve Text-VQA capabilities.

6 Conclusion

In this work, we introduced the VCR task, a novel vision-language challenge aimed at promoting the integration of visual and textual modalities, including text embedded in both natural language tokens and image formats and highly obscured text embedded in the image. We developed a specialized pipeline to create a dataset tailored to this task, utilizing correlated image-text pairs. This task stands out from existing methods by requiring a more profound integration of visual cues and partially obscured text, highlighting its uniqueness and importance in the field.

We conducted extensive evaluations of state-of-the-art vision-language models (VLMs) in both English and Chinese. The results demonstrated significant room for improvement, suggesting that current models have not yet fully exploited the capabilities necessary for VCR. We selected models representing both the highest and average performance tiers for additional fine-tuning with our dataset. Although fine-tuning exhibited potential for enhancing VCR capabilities, it did not consistently result in significant improvements, indicating the complexity and challenges of adapting models to this task.

By introducing the VCR task and its specialized dataset, we aim to advance research in vision-language interaction. The unique challenges of VCR seek to improve model development and training, extending the limits of multimodal AI. VCR provides a controllable testbed for fine-grained analysis of model behavior across languages, difficulty levels, image inclusion, and fine-tuning stages. We invite the community to utilize our dataset and develop innovative strategies to boost the performance of vision-language models.

REFERENCES

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

Anthropic. The claude 3 model family: Opus, sonnet, haiku. 2024.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015.

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.

Alessandro Bissacco, Mark Cummins, Yuval Netzer, and Hartmut Neven. Photoocr: Reading text in uncontrolled conditions. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2013.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marcal Rusinol, Ernest Valveny, C.V. Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Fedor Borisyuk, Albert Gordo, and Viswanath Sivakumar. Rosetta: Large scale system for text detection and recognition in images. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pp. 71–79, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355520. doi: 10.1145/3219819.3219861. URL https://doi.org/10.1145/3219819.3219861.

Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. M^{3}cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. *arXiv preprint arXiv: 2405.16473*, 2024a.

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.

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*, 2024b.

Aaron Defazio, Xingyu Yang, Harsh Mehta, Konstantin Mishchenko, Ahmed Khaled, and Ashok Cutkosky. The road less scheduled, 2024.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv: 2401.16420*, 2024a.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Songyang Zhang, Haodong Duan, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Zhe Chen, Xinyue Zhang, Wei Li, Jingwen Li, Wenhai Wang, Kai Chen, Conghui He, Xingcheng Zhang, Jifeng Dai, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer2-4khd: A pioneering large vision-language model handling resolutions from 336 pixels to 4k hd. *arXiv preprint arXiv:2404.06512*, 2024b.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, and et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Volkmar Frinken, Andreas Fischer, R Manmatha, and Horst Bunke. A novel word spotting method based on recurrent neural networks. *IEEE transactions on pattern analysis and machine intelligence*, 34(2):211–224, 2011.

Amber M Fyall, Yasmine El-Shamayleh, Hannah Choi, Eric Shea-Brown, and Anitha Pasupathy. Dynamic representation of partially occluded objects in primate prefrontal and visual cortex. *eLife*, 6:e25784, September 2017. ISSN 2050-084X. doi: 10.7554/eLife.25784.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data for text localisation in natural images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*, 2024.

Nicholas R Howe. Document binarization with automatic parameter tuning. *International journal on document analysis and recognition (ijdar)*, 16:247–258, 2013.

Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *arXiv preprint arXiv: 2403.12895*, 2024a.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=nZeVKeeFYf9`.

Jinyi Hu, Yuan Yao, Chongyi Wang, Shan Wang, Yinxu Pan, Qianyu Chen, Tianyu Yu, Hanghao Wu, Yue Zhao, Haoye Zhang, Xu Han, Yankai Lin, Jiao Xue, Dahai Li, Zhiyuan Liu, and Maosong Sun. Large multilingual models pivot zero-shot multimodal learning across languages. *arXiv preprint arXiv:2308.12038*, 2023.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Minicpm: Unveiling the potential of small language models with scalable training strategies, 2024b.

Weilin Huang, Yu Qiao, and Xiaoou Tang. Robust scene text detection with convolution neural network induced mser trees. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 497–511, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10593-2.

Max Jaderberg, Andrea Vedaldi, and Andrew Zisserman. Deep features for text spotting. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 512–528, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10593-2.

Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Pythia v0.1: the winning entry to the vqa challenge 2018, 2018.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. Obelics: An open web-scale filtered dataset of interleaved image-text documents, 2023.

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

Chen-Yu Lee and Simon Osindero. Recursive recurrent nets with attention modeling for ocr in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Bao Li, Chi Zhang, Long Cao, Panpan Chen, Tianyuan Liu, Hui Gao, Linyuan Wang, Bin Yan, and Li Tong. Brain Functional Representation of Highly Occluded Object Recognition. *Brain Sciences*, 13(10):1387, October 2023a. ISSN 2076-3425. doi: 10.3390/brainsci13101387.

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. *arXiv preprint arXiv: 2311.06607*, 2023b.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916. Curran Associates, Inc., 2023b. URL `https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf`.

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 2024a. URL `https://llava-vl.github.io/blog/2024-01-30-llava-next/`.

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*, 2024b.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. Deepseek-vl: Towards real-world vision-language understanding, 2024.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Minesh Mathew, Dimosthenis Karatzas, and C.V. Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2200–2209, January 2021.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C.V. Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1697–1706, January 2022.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 947–952, 2019a. doi: 10.1109/ICDAR.2019.00156.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 947–952, 2019b. doi: 10.1109/ICDAR.2019.00156.

AI Mistral. Announcing pixtral 12B, September 2024.

G. Nagy. Twenty years of document image analysis in pami. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(01):38–62, jan 2000. ISSN 1939-3539. doi: 10.1109/34.824820.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, et al. Gpt-4 technical report. *arXiv preprint arXiv: 2303.08774*, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Luiz Pessoa, Evan Thompson, and Alva Noë. Finding out about filling-in: A guide to perceptual completion for visual science and the philosophy of perception. *Behavioral and Brain Sciences*, 21 (6):723–748, 1998. doi: 10.1017/S0140525X98001757.

Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*, 2021.

Faisal Shafait, Daniel Keysers, and Thomas Breuel. Performance evaluation and benchmarking of six-page segmentation algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(6):941–954, 2008.

Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11):2298–2304, 2017. doi: 10.1109/TPAMI.2016.2646371.

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*, pp. 8317–8326, 2019.

Ray Smith. A simple and efficient skew detection algorithm via text row accumulation. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 2, pp. 1145–1148. IEEE, 1995.

Zeyi Sun, Ye Fang, Tong Wu, Pan Zhang, Yuhang Zang, Shu Kong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Alpha-clip: A clip model focusing on wherever you want. *arXiv preprint arXiv: 2312.03818*, 2023.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, Andrea Tacchetti, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv: 2403.05530*, 2024a.

Reka Team, Aitor Ormazabal, Che Zheng, Cyprien de Masson d'Autume, Dani Yogatama, Deyu Fu, Donovan Ong, Eric Chen, Eugenie Lamprecht, Hai Pham, Isaac Ong, Kaloyan Aleksiev, Lei Li, Matthew Henderson, Max Bain, Mikel Artetxe, Nishant Relan, Piotr Padlewski, Qi Liu, Ren Chen, Samuel Phua, Yazheng Yang, Yi Tay, Yuqi Wang, Zhongkai Zhu, and Zhihui Xie. Reka core, flash, and edge: A series of powerful multimodal language models. *arXiv preprint arXiv: 2404.12387*, 2024b.

G. Thinés, A. Costall, and G. Butterworth. *Michotte's Experimental Phenomenology of Perception*. Routledge Library Editions: Phenomenology. Taylor & Francis, 2013. ISBN 9781134506897. URL `https://books.google.com/books?id=nbREAQAAQBAJ`.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs. *arXiv preprint arXiv:2406.16860*, 2024.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv: 2307.09288*, 2023.

Rob van Lier and Walter Gerbino. Perceptual completions. In *The Oxford Handbook of Perceptual Organization*. Oxford University Press, 08 2015. ISBN 9780199686858. doi: 10.1093/oxfordhb/9780199686858.013.040. URL https://doi.org/10.1093/oxfordhb/9780199686858.013.040.

Dongsheng Wang, Natraj Raman, Mathieu Sibue, Zhiqiang Ma, Petr Babkin, Simerjot Kaur, Yulong Pei, Armineh Nourbakhsh, and Xiaomo Liu. Docllm: A layout-aware generative language model for multimodal document understanding. *arXiv preprint arXiv: 2401.00908*, 2023a.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton van den Hengel. Fvqa: Fact-based visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40 (10):2413–2427, 2018. doi: 10.1109/TPAMI.2017.2754246.

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.

Tao Wang, David J. Wu, Adam Coates, and Andrew Y. Ng. End-to-end text recognition with convolutional neural networks. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pp. 3304–3308, 2012.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models, 2023b.

Xinyu Wang, Yuliang Liu, Chunhua Shen, Chun Chet Ng, Canjie Luo, Lianwen Jin, Chee Seng Chan, Anton van den Hengel, and Liangwei Wang. On the general value of evidence, and bilingual scene-text visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Shilian Wu, Yongrui Li, and Zengfu Wang. Chinese text recognition enhanced by glyph and character semantic information. *International Journal on Document Analysis and Recognition (IJDAR)*, 27(1):45–56, March 2024. ISSN 1433-2825. doi: 10.1007/s10032-023-00444-9. URL https://doi.org/10.1007/s10032-023-00444-9.

Yusong Wu, K. Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and S. Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2022. doi: 10.1109/ICASSP49357.2023.10095969.

Ruyi Xu, Yuan Yao, Zonghao Guo, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, and Gao Huang. LLaVA-UHD: an lmm perceiving any aspect ratio and high-resolution images. *arXiv preprint arXiv:2403.11703*, 2024.

Zhengyuan Yang, Yijuan Lu, Jianfeng Wang, Xi Yin, Dinei Florencio, Lijuan Wang, Cha Zhang, Lei Zhang, and Jiebo Luo. Tap: Text-aware pre-training for text-vqa and text-caption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8751–8761, June 2021.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. *arXiv preprint arXiv:2312.00849*, 2023.

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

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. *IEEE International Conference on Computer Vision*, 2023. doi: 10.1109/ICCV51070.2023.01100.

Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, Songyang Zhang, Wenwei Zhang, Yining Li, Yang Gao, Peng Sun, Xinyue Zhang, Wei Li, Jingwen Li, Wenhai Wang, Hang Yan, Conghui He, Xingcheng Zhang, Kai Chen, Jifeng Dai, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*, 2024.

Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and Yang: Balancing and answering binary visual questions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv: 2302.00923*, 2023.

Yuliang Zhao, Xinyue Zhang, Boya Fu, Zhikun Zhan, Hui Sun, Lianjiang Li, and Guanglie Zhang. Evaluation and recognition of handwritten chinese characters based on similarities. *Applied Sciences*, 12(17), 2022. ISSN 2076-3417. doi: 10.3390/app12178521. URL `https://www.mdpi.com/2076-3417/12/17/8521`.

Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. East: An efficient and accurate scene text detector. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

A Additional evaluation results on first 100 and 500 test cases

Table 4: Results of various open-source and closed-source vision language models on the VCR task using the first 100 test cases. FT = finetuned on 16,000 samples from the VCR-WIKI training set. The best results among the finetuned models are <u>underlined</u> while the best results among the models without finetuning are highlighted in **bold**. Subscripts are standard deviations obtained from Bootstrap.

Language	Mode	Open/closed source	Model name	Model size	Exact match (%) ↑ $VI+TEI$	Exact match (%) ↑ TEI	Exact match (%) ↑ Δ	Jaccard index (%) ↑ $VI+TEI$	Jaccard index (%) ↑ TEI	Jaccard index (%) ↑ Δ
English	Easy	Closed	Claude 3 Opus	-	$62.0_{0.76}$	$82.0_{0.63}$	-20	$78.06_{0.24}$	$91.12_{0.13}$	-13.06
			Claude 3.5 Sonnet	-	$70.41_{3.46}$	$75.15_{3.36}$	-4.73	$78.1_{2.85}$	$86.5_{2.18}$	-8.4
			Gemini 1.5 Pro	-	$71.01_{3.4}$	$86.98_{2.67}$	-15.98	$82.89_{2.27}$	$94.21_{1.32}$	-11.32
			GPT-4 Turbo	-	$78.47_{0.22}$	$86.6_{0.79}$	-8.13	$88.08_{0.25}$	$94.15_{0.2}$	-6.07
			GPT-4o	-	$\mathbf{90.91_{0.36}}$	$\mathbf{95.69_{0.23}}$	-4.78	$\mathbf{96.77_{0.16}}$	$\mathbf{98.45_{0.06}}$	-1.68
			GPT-4V	-	$25.36_{0.5}$	$18.18_{0.54}$	7.18	$35.64_{0.22}$	$28.49_{0.23}$	7.15
			Qwen-VL-Max	-	$82.3_{0.19}$	$88.04_{0.43}$	-5.74	$89.73_{0.32}$	$92.55_{0.17}$	-2.82
			Reka Core	-	$65.68_{3.78}$	$78.11_{3.19}$	-12.43	$83.14_{2.04}$	$90.43_{1.49}$	-7.29
		Open	Cambrian-1	34B	$78.11_{3.16}$	$82.84_{2.86}$	-4.73	$87.88_{1.97}$	$93.12_{1.26}$	-5.24
			CogVLM2	19B	$86.39_{0.66}$	$84.62_{0.92}$	1.78	$91.39_{0.11}$	$91.63_{0.11}$	-0.24
			CogVLM2-FT	19B	$\underline{94.08_{0.2}}$	$\underline{94.67_{0.26}}$	-0.59	$\underline{98.03_{0.07}}$	$\underline{98.22_{0.03}}$	-0.2
			DeepSeek-VL	1.3B	$19.53_{0.69}$	$26.04_{1.47}$	-6.51	$43.73_{0.18}$	$48.03_{0.16}$	-4.3
			DeepSeek-VL	7B	$36.09_{1.36}$	$44.97_{0.79}$	-8.88	$57.81_{0.18}$	$61.83_{0.33}$	-4.01
			DocOwl-1.5-Omni	8B	$0.59_{0.14}$	$1.18_{0.14}$	-0.59	$12.69_{0.04}$	$13.3_{0.06}$	-0.61
			Idefics3	8B	$26.63_{3.35}$	$32.54_{3.63}$	-5.92	$48.83_{2.95}$	$55.11_{2.78}$	-6.29
			InternLM-XComposer2-VL	7B	$47.93_{0.69}$	$47.34_{0.57}$	0.59	$73.88_{0.22}$	$74.58_{0.16}$	-0.7
			InternLM-XComposer2-VL-4K	7B	$4.14_{1.54}$	$3.55_{1.49}$	0.59	$21.91_{1.81}$	$21.85_{1.86}$	0.06
			InternLM-XComposer2.5-VL	7B	$45.56_{3.83}$	$28.99_{3.50}$	16.57	$67.70_{2.79}$	$54.25_{2.70}$	13.45
			InternVL-V2	40B	$86.39_{2.56}$	$86.98_{2.60}$	-0.59	$93.51_{1.40}$	$94.35_{1.24}$	-0.84
			Llama-3.2	11B	$79.88_{2.96}$	$68.64_{3.75}$	11.24	$90.88_{1.56}$	$82.91_{2.11}$	7.97
			Llama-3.2	90B	$79.29_{3.14}$	$71.01_{3.36}$	8.28	$87.81_{2.00}$	$83.17_{2.30}$	4.64
			MiniCPM-V2.5	8B	$30.18_{0.66}$	$36.09_{0.34}$	-5.92	$53.1_{0.18}$	$59.06_{0.14}$	-5.96
			MiniCPM-V2.5-FT	8B	$39.05_{0.69}$	$46.75_{0.59}$	-7.69	$63.05_{0.28}$	$69.89_{0.33}$	-6.84
			Monkey	7B	$46.75_{0.44}$	$48.52_{0.41}$	-1.78	$67.82_{0.22}$	$68.59_{0.13}$	-0.76
			Pixtral	12B	$14.79_{2.65}$	$13.02_{2.62}$	1.78	$39.00_{2.49}$	$33.16_{2.53}$	5.84
			Qwen-VL	7B	$47.34_{0.44}$	$46.75_{0.57}$	0.59	$69.02_{0.35}$	$69.19_{0.37}$	-0.17
			Qwen2-VL	7B	$\mathbf{90.53_{2.25}}$	$\mathbf{96.45_{1.39}}$	-5.92	$\mathbf{94.28_{1.54}}$	$\mathbf{98.82_{0.49}}$	-4.54
			Yi-VL	34B	$1.78_{0.16}$	$1.18_{0.11}$	0.59	$6.21_{0.06}$	$7.5_{0.08}$	-1.3
			Yi-VL	6B	$2.37_{0.13}$	$1.78_{0.22}$	0.59	$6.24_{0.07}$	$8.05_{0.11}$	-1.81
	Hard	Closed	Claude 3 Opus	-	$34.0_{1.12}$	$51.0_{0.5}$	-17	$57.02_{0.24}$	$70.32_{0.15}$	-13.31
			Claude 3.5 Sonnet	-	$46.75_{3.58}$	$43.2_{3.83}$	3.55	$57.74_{3.33}$	$54.13_{3.51}$	3.61
			Gemini 1.5 Pro	-	$33.73_{3.69}$	$43.79_{3.74}$	-10.06	$57.09_{2.67}$	$62.34_{2.76}$	-5.25
			GPT-4 Turbo	-	$53.11_{0.46}$	$57.42_{0.5}$	-4.31	$71.75_{0.19}$	$73.82_{0.24}$	-2.07
			GPT-4o	-	$\mathbf{74.16_{0.31}}$	$\mathbf{84.69_{0.31}}$	-10.53	$\mathbf{86.99_{0.09}}$	$\mathbf{93.19_{0.07}}$	-6.21
			GPT-4V	-	$28.71_{0.49}$	$16.27_{0.73}$	12.44	$49.89_{0.15}$	$33.64_{0.16}$	16.25
			Qwen-VL-Max	-	$40.67_{0.38}$	$55.02_{0.46}$	-14.35	$61.8_{0.19}$	$72.46_{0.15}$	-10.66
			Reka Core	-	$7.1_{2.01}$	$10.65_{2.38}$	-3.55	$25.49_{1.99}$	$36.78_{2.19}$	-11.29
		Open	Cambrian-1	34B	$27.81_{3.29}$	$29.59_{3.54}$	-1.78	$51.39_{2.79}$	$54.00_{2.76}$	-2.61
			CogVLM2	19B	$44.97_{0.83}$	$21.3_{0.47}$	23.67	$65.39_{0.2}$	$43.86_{0.27}$	21.53
			CogVLM2-FT	19B	$\underline{75.74_{0.72}}$	$\underline{67.46_{0.64}}$	8.28	$\underline{90.6_{0.13}}$	$\underline{84.26_{0.08}}$	6.34
			DeepSeek-VL	1.3B	$0.0_{0.0}$	$0.0_{0.0}$	0	$11.17_{0.03}$	$10.88_{0.06}$	0.29
			DeepSeek-VL	7B	$0.59_{0.09}$	$1.78_{0.17}$	-1.18	$16.71_{0.11}$	$18.09_{0.13}$	-1.38
			DocOwl-1.5-Omni	8B	$0.0_{0.0}$	$0.0_{0.0}$	0	$7.89_{0.05}$	$8.28_{0.05}$	-0.4
			Idefics3	8B	$1.18_{0.87}$	$0.59_{0.61}$	0.59	$11.62_{1.20}$	$10.28_{1.00}$	1.34
			InternLM-XComposer2-VL	7B	$0.0_{0.0}$	$0.59_{0.09}$	-0.59	$12.69_{0.08}$	$14.05_{0.11}$	-1.35
			InternLM-XComposer2-VL-4K	7B	$0.00_{0.00}$	$0.59_{0.59}$	-0.59	$9.67_{0.90}$	$8.83_{0.95}$	0.84
			InternLM-XComposer2.5-VL	7B	$0.59_{0.58}$	$1.78_{1.01}$	-1.18	$14.09_{1.04}$	$16.57_{1.25}$	-2.48
			InternVL-V2	40B	$12.43_{2.54}$	$16.57_{2.89}$	-4.14	$33.74_{2.40}$	$39.51_{2.69}$	-5.76
			Llama-3.2	11B	$10.65_{2.33}$	$7.69_{2.04}$	2.96	$33.50_{2.28}$	$26.80_{2.03}$	6.70
			Llama-3.2	90B	$13.02_{2.59}$	$15.38_{2.75}$	-2.37	$36.80_{2.36}$	$39.85_{2.52}$	-3.06
			MiniCPM-V2.5	8B	$1.18_{0.12}$	$1.78_{0.12}$	-0.59	$12.02_{0.12}$	$12.41_{0.07}$	-0.39
			MiniCPM-V2.5-FT	8B	$10.06_{0.43}$	$13.02_{0.54}$	-2.96	$34.67_{0.2}$	$36.43_{0.19}$	-1.76
			Monkey	7B	$1.18_{0.22}$	$3.55_{0.18}$	-2.37	$12.66_{0.21}$	$15.97_{0.08}$	-3.31
			Pixtral	12B	$0.00_{0.00}$	$0.59_{0.61}$	-0.59	$9.90_{0.79}$	$12.56_{1.09}$	-2.66
			Qwen-VL	7B	$1.78_{0.21}$	$2.96_{0.12}$	-1.18	$15.7_{0.14}$	$15.06_{0.19}$	0.63
			Qwen2-VL	7B	$\mathbf{75.74_{3.32}}$	$\mathbf{73.96_{3.55}}$	1.78	$\mathbf{85.91_{2.17}}$	$\mathbf{85.83_{2.04}}$	0.08
			Yi-VL	34B	$0.59_{0.09}$	$0.0_{0.0}$	0.59	$4.39_{0.07}$	$5.49_{0.08}$	-1.1
			Yi-VL	6B	$0.59_{0.13}$	$0.0_{0.0}$	0.59	$5.12_{0.03}$	$5.5_{0.06}$	-0.38
Chinese	Easy	Closed	Claude 3 Opus	-	$0.53_{0.51}$	$0.53_{0.55}$	0	$11.34_{1.07}$	$9.14_{0.93}$	2.2
			Claude 3.5 Sonnet	-	$1.6_{0.91}$	$2.13_{1.05}$	-0.53	$8.07_{1.29}$	$9.9_{1.48}$	-1.84
			Gemini 1.5 Pro	-	$0.53_{0.56}$	$0.0_{0.0}$	0.53	$12.94_{1.26}$	$12.77_{1.17}$	0.16
			GPT-4o	-	$\mathbf{14.89_{2.51}}$	$\mathbf{21.81_{2.98}}$	-6.91	$\mathbf{38.57_{2.46}}$	$\mathbf{48.29_{2.43}}$	-9.72
			GPT-4 Turbo	-	$0.53_{0.55}$	$0.0_{0.0}$	0.53	$11.09_{1.05}$	$7.51_{0.65}$	3.58
			Qwen-VL-Max	-	$5.93_{0.19}$	$8.7_{0.37}$	-2.77	$13.53_{0.11}$	$18.5_{0.1}$	-4.97
			Reka Core	-	$0.0_{0.0}$	$0.0_{0.0}$	0	$3.04_{0.53}$	$2.42_{0.45}$	0.61
		Open	CogVLM2-Chinese	19B	$34.57_{0.66}$	$34.04_{1.01}$	0.53	$58.78_{0.13}$	$57.26_{0.12}$	1.52
			CogVLM2-Chinese-FT	19B	$\underline{66.49_{0.74}}$	$\underline{67.55_{0.73}}$	-1.06	$\underline{79.48_{0.17}}$	$\underline{81.78_{0.09}}$	-2.3
			DeepSeek-VL	1.3B	$0.0_{0.0}$	$0.0_{0.0}$	0	$6.69_{0.07}$	$2.92_{0.02}$	3.78
			DeepSeek-VL	7B	$0.0_{0.0}$	$0.0_{0.0}$	0	$3.99_{0.07}$	$6.71_{0.02}$	-2.72
			DocOwl-1.5-Omni	8B	$0.0_{0.0}$	$0.0_{0.0}$	0	$1.23_{0.04}$	$2.97_{0.02}$	-1.75
			InternLM-XComposer2-VL	7B	$1.06_{0.09}$	$0.53_{0.07}$	0.53	$13.1_{0.03}$	$13.26_{0.03}$	-0.16
			InternLM-XComposer2-VL-4K	7B	$0.00_{0.00}$	$0.00_{0.00}$	0.00	$13.49_{1.01}$	$14.56_{0.98}$	-1.08
			InternLM-XComposer2.5-VL	7B	$0.00_{0.00}$	$1.60_{0.91}$	-1.60	$11.94_{0.88}$	$16.12_{1.24}$	-4.18
			InternVL-V2	40B	$26.06_{3.17}$	$19.15_{2.88}$	6.91	$48.98_{2.61}$	$41.25_{2.57}$	7.72
			MiniCPM-V2.5	8B	$4.79_{0.16}$	$7.45_{0.35}$	-2.66	$20.58_{0.11}$	$25.38_{0.13}$	-4.81
			MiniCPM-V2.5-FT	8B	$6.91_{0.33}$	$7.98_{0.4}$	-1.06	$30.8_{0.07}$	$31.46_{0.52}$	-0.66
			Monkey	7B	$1.06_{0.12}$	$0.53_{0.06}$	0.53	$9.23_{0.08}$	$12.29_{0.13}$	-3.06
			Qwen-VL	7B	$0.0_{0.0}$	$0.0_{0.0}$	0	$1.41_{0.02}$	$0.66_{0.03}$	0.76
			Qwen2-VL	7B	$\mathbf{67.55_{3.34}}$	$\mathbf{73.40_{3.21}}$	-5.85	$\mathbf{84.63_{1.80}}$	$\mathbf{87.12_{1.76}}$	-2.49
			Yi-VL	34B	$0.0_{0.0}$	$0.0_{0.0}$	0	$4.53_{0.03}$	$1.84_{0.05}$	2.69
			Yi-VL	6B	$0.0_{0.0}$	$0.0_{0.0}$	0	$4.73_{0.02}$	$1.55_{0.02}$	3.18
	Hard	Closed	Claude 3 Opus	-	$1.06_{0.77}$	$0.53_{0.54}$	0.53	$9.23_{1.04}$	$7.77_{0.83}$	1.45
			Claude 3.5 Sonnet	-	$0.53_{0.51}$	$0.0_{0.0}$	0.53	$4.11_{0.84}$	$3.32_{0.71}$	0.79
			Gemini 1.5 Pro	-	$1.06_{0.71}$	$1.06_{0.77}$	0	$11.58_{1.14}$	$13.34_{1.2}$	-1.76
			GPT-4o	-	$\mathbf{2.66_{1.16}}$	$\mathbf{1.6_{0.92}}$	1.06	$\mathbf{23.69_{1.65}}$	$\mathbf{23.69_{1.48}}$	0
			GPT-4 Turbo	-	$0.0_{0.0}$	$0.53_{0.53}$	-0.53	$8.51_{0.7}$	$8.02_{0.78}$	0.49
			Qwen-VL-Max	-	$1.19_{0.12}$	$1.98_{0.09}$	-0.79	$6.19_{0.1}$	$11.09_{0.11}$	-4.9
			Reka Core	-	$0.0_{0.0}$	$0.0_{0.0}$	0	$3.22_{0.51}$	$3.62_{0.57}$	-0.4
		Open	CogVLM2-Chinese	19B	$3.19_{0.19}$	$3.19_{0.32}$	0	$18.33_{0.14}$	$21.38_{0.09}$	-3.05
			CogVLM2-Chinese-FT	19B	$\underline{46.81_{0.32}}$	$\underline{46.28_{0.49}}$	0.53	$\underline{66.85_{0.39}}$	$\underline{69.79_{0.12}}$	-2.95
			DeepSeek-VL	1.3B	$0.0_{0.0}$	$0.0_{0.0}$	0	$6.5_{0.03}$	$4.16_{0.03}$	2.34
			DeepSeek-VL	7B	$0.0_{0.0}$	$0.0_{0.0}$	0	$5.22_{0.04}$	$7.45_{0.06}$	-2.23
			DocOwl-1.5-Omni	8B	$0.0_{0.0}$	$0.0_{0.0}$	0	$1.35_{0.02}$	$3.57_{0.04}$	-2.23
			InternLM-XComposer2-VL	7B	$0.0_{0.0}$	$0.0_{0.0}$	0	$8.17_{0.03}$	$7.99_{0.03}$	0.18
			InternLM-XComposer2-VL-4K	7B	$0.00_{0.00}$	$0.00_{0.00}$	0.00	$7.73_{0.68}$	$8.12_{0.79}$	-0.39
			InternLM-XComposer2.5-VL	7B	$0.00_{0.00}$	$0.00_{0.00}$	0.00	$10.87_{0.82}$	$10.54_{0.84}$	0.32
			InternVL-V2	40B	$0.53_{0.50}$	$1.06_{0.72}$	-0.53	$12.26_{1.01}$	$13.58_{1.20}$	-1.32
			MiniCPM-V2.5	8B	$0.53_{0.07}$	$0.53_{0.07}$	0	$7.28_{0.06}$	$7.71_{0.06}$	-0.43
			MiniCPM-V2.5-FT	8B	$1.06_{0.08}$	$2.13_{0.19}$	-1.06	$18.46_{0.1}$	$16.42_{0.22}$	2.03
			Monkey	7B	$0.0_{0.0}$	$0.0_{0.0}$	0	$6.15_{0.11}$	$6.62_{0.11}$	-0.47
			Qwen-VL	7B	$0.0_{0.0}$	$0.0_{0.0}$	0	$1.1_{0.04}$	$0.06_{0.01}$	1.04
			Qwen2-VL	7B	$\mathbf{17.55_{2.81}}$	$\mathbf{27.66_{3.21}}$	-10.11	$\mathbf{43.87_{2.48}}$	$\mathbf{51.99_{2.61}}$	-8.12
			Yi-VL	34B	$0.0_{0.0}$	$0.0_{0.0}$	0	$4.17_{0.04}$	$2.02_{0.04}$	2.15
			Yi-VL	6B	$0.0_{0.0}$	$0.0_{0.0}$	0	$4.15_{0.06}$	$2.38_{0.04}$	1.77

Table 5: Results of various open-source and closed-source vision language models on the VCR task using the first 500 test cases. FT = finetuned on 16,000 samples from the VCR-WIKI training set. The best results among the finetuned models are <u>underlined</u> while the best results among the models without finetuning are highlighted in **bold**. Subscripts are standard deviations obtained from Bootstrap.

Language	Mode	Open/closed source	Model name	Model size	Exact match (%) ↑			Jaccard index (%) ↑		
					$VI+TEI$	TEI	Δ	$VI+TEI$	TEI	Δ
English	Easy	Closed	Claude 3 Opus	-	$62.0_{0.13}$	$77.0_{0.5}$	-15	$77.67_{0.32}$	$88.41_{0.39}$	-10.74
			Claude 3.5 Sonnet	-	$63.85_{1.71}$	$72.8_{1.56}$	-8.94	$74.65_{1.33}$	$83.48_{1.14}$	-8.83
			Gemini 1.5 Pro	-	$62.73_{1.66}$	$82.98_{1.3}$	-20.25	$77.71_{1.21}$	$91.56_{0.76}$	-13.85
			GPT-4 Turbo	-	$78.74_{0.13}$	$81.94_{0.25}$	-3.2	$88.54_{0.24}$	$92.18_{0.3}$	-3.65
			GPT-4o	-	$\mathbf{91.55_{0.29}}$	$\mathbf{94.56_{0.13}}$	-3.01	$\mathbf{96.44_{0.11}}$	$\mathbf{97.76_{0.06}}$	-1.32
			GPT-4V	-	$52.04_{0.24}$	$37.86_{0.22}$	14.17	$65.36_{0.39}$	$54.13_{0.41}$	11.23
			Qwen-VL-Max	-	$76.8_{0.5}$	$85.53_{0.19}$	-8.74	$85.71_{0.28}$	$91.45_{0.29}$	-5.74
			Reka Core	-	$66.46_{1.64}$	$78.51_{1.42}$	-12.05	$84.23_{0.86}$	$90.45_{0.7}$	-6.22
		Open	Cambrian-1	34B	$76.89_{1.52}$	$80.25_{1.36}$	-3.35	$87.66_{0.90}$	$92.42_{0.60}$	-4.76
			CogVLM2	19B	$83.11_{0.28}$	$79.63_{0.33}$	3.48	$89.43_{0.27}$	$88.65_{0.26}$	0.79
			CogVLM2-FT	19B	$\underline{92.8_{0.06}}$	$\underline{92.67_{0.13}}$	0.12	$\underline{97.51_{0.24}}$	$\underline{97.45_{0.07}}$	0.06
			DeepSeek-VL	1.3B	$21.86_{0.17}$	$30.68_{0.3}$	-8.82	$45.4_{0.33}$	$52.02_{0.73}$	-6.62
			DeepSeek-VL	7B	$37.76_{0.42}$	$45.47_{0.21}$	-7.7	$59.07_{0.43}$	$64.26_{0.57}$	-5.2
			DocOwl-1.5-Omni	8B	$0.62_{0.06}$	$1.86_{0.06}$	-1.24	$12.65_{0.3}$	$14.09_{0.12}$	-1.44
			Idefics3	8B	$26.71_{1.57}$	$29.81_{1.55}$	-3.11	$46.91_{1.40}$	$51.84_{1.30}$	-4.93
			InternLM-XComposer2-VL	7B	$46.09_{0.35}$	$46.34_{0.25}$	-0.25	$71.11_{0.2}$	$71.76_{0.67}$	-0.65
			InternLM-XComposer2-VL-4K	7B	$5.22_{0.80}$	$3.23_{0.63}$	1.99	$22.70_{0.89}$	$18.67_{0.79}$	4.03
			InternLM-XComposer2.5-VL	7B	$42.48_{1.73}$	$25.84_{1.53}$	16.65	$63.03_{1.32}$	$50.75_{1.21}$	12.28
			InternVL-V2	40B	$84.84_{1.21}$	$87.08_{1.19}$	-2.24	$93.13_{0.69}$	$94.83_{0.50}$	-1.71
			Llama-3.2	11B	$79.25_{1.40}$	$66.46_{1.63}$	12.80	$89.98_{0.77}$	$80.91_{1.06}$	9.08
			Llama-3.2	90B	$80.87_{1.37}$	$71.55_{1.54}$	9.32	$89.63_{0.85}$	$84.34_{0.98}$	5.29
			MiniCPM-V2.5	8B	$32.8_{0.16}$	$36.77_{0.25}$	-3.98	$52.56_{0.25}$	$60.89_{0.19}$	-8.32
			MiniCPM-V2.5-FT	8B	$42.36_{0.3}$	$45.34_{0.35}$	-2.98	$65.39_{0.6}$	$67.85_{0.43}$	-2.46
			Monkey	7B	$47.2_{0.2}$	$54.16_{0.41}$	-6.96	$65.7_{0.4}$	$71.17_{0.72}$	-5.47
			Pixtral	12B	$16.65_{1.31}$	$11.80_{1.13}$	4.84	$39.81_{1.16}$	$31.47_{1.11}$	8.34
			Qwen-VL	7B	$45.47_{0.35}$	$52.17_{0.33}$	-6.71	$66.81_{0.74}$	$71.73_{0.59}$	-4.93
			Qwen2-VL	7B	$\mathbf{90.06_{1.07}}$	$\mathbf{94.53_{0.82}}$	-4.47	$\mathbf{93.77_{0.76}}$	$\mathbf{97.80_{0.34}}$	-4.03
			Yi-VL	34B	$0.87_{0.06}$	$1.24_{0.04}$	-0.37	$5.61_{0.28}$	$7.63_{0.42}$	-2.02
			Yi-VL	6B	$1.12_{0.03}$	$1.37_{0.14}$	-0.25	$5.93_{0.16}$	$7.33_{0.23}$	-1.39
	Hard	Closed	Claude 3 Opus	-	$37.8_{0.28}$	$50.0_{0.33}$	-12.2	$57.68_{0.8}$	$70.16_{0.64}$	-12.48
			Claude 3.5 Sonnet	-	$41.74_{1.69}$	$44.72_{1.78}$	-2.98	$56.15_{1.46}$	$58.54_{1.6}$	-2.4
			Gemini 1.5 Pro	-	$28.07_{1.58}$	$38.76_{1.68}$	-10.68	$51.9_{1.22}$	$59.62_{1.27}$	-7.72
			GPT-4 Turbo	-	$45.15_{0.28}$	$48.64_{0.57}$	-3.5	$65.72_{0.25}$	$67.86_{0.2}$	-2.14
			GPT-4o	-	$\mathbf{73.2_{0.16}}$	$\mathbf{82.43_{0.17}}$	-9.22	$\mathbf{86.17_{0.21}}$	$\mathbf{92.01_{0.2}}$	-5.84
			GPT-4V	-	$25.83_{0.44}$	$14.95_{0.3}$	10.87	$44.63_{0.48}$	$30.08_{0.67}$	14.56
			Qwen-VL-Max	-	$41.65_{0.32}$	$52.72_{0.2}$	-11.07	$61.18_{0.35}$	$70.19_{0.37}$	-9.01
			Reka Core	-	$6.71_{0.89}$	$11.18_{1.15}$	-4.47	$25.84_{0.95}$	$35.83_{1.05}$	-9.99
		Open	Cambrian-1	34B	$27.20_{1.59}$	$30.19_{1.55}$	-2.98	$49.96_{1.36}$	$55.93_{1.23}$	-5.97
			CogVLM2	19B	$41.74_{0.25}$	$16.77_{0.22}$	24.97	$62.56_{0.33}$	$38.41_{0.44}$	24.15
			CogVLM2-FT	19B	$\underline{75.9_{0.13}}$	$\underline{65.22_{0.18}}$	10.68	$\underline{89.75_{0.14}}$	$\underline{82.71_{0.27}}$	7.04
			DeepSeek-VL	1.3B	$0.37_{0.02}$	$0.12_{0.01}$	0.25	$11.42_{0.09}$	$11.41_{0.22}$	0.01
			DeepSeek-VL	7B	$0.75_{0.02}$	$1.61_{0.1}$	-0.87	$15.8_{0.29}$	$17.18_{0.41}$	-1.38
			DocOwl-1.5-Omni	8B	$0.0_{0.0}$	$0.0_{0.0}$	0	$7.34_{0.06}$	$7.61_{0.16}$	-0.27
			Idefics3	8B	$0.75_{0.30}$	$0.50_{0.25}$	0.25	$10.44_{0.49}$	$9.17_{0.43}$	1.27
			InternLM-XComposer2-VL	7B	$0.5_{0.04}$	$0.37_{0.05}$	0.12	$12.38_{0.13}$	$13.22_{0.11}$	-0.83
			InternLM-XComposer2-VL-4K	7B	$0.00_{0.00}$	$0.12_{0.12}$	-0.12	$9.55_{0.38}$	$9.18_{0.38}$	0.37
			InternLM-XComposer2.5-VL	7B	$0.75_{0.31}$	$1.24_{0.39}$	-0.50	$13.67_{0.51}$	$14.92_{0.56}$	-1.25
			InternVL-V2	40B	$14.16_{1.22}$	$18.51_{1.36}$	-4.35	$35.01_{1.18}$	$41.02_{1.22}$	-6.02
			Llama-3.2	11B	$13.91_{1.25}$	$7.33_{0.94}$	6.58	$35.78_{1.14}$	$26.14_{0.94}$	9.64
			Llama-3.2	90B	$15.16_{1.28}$	$12.17_{1.13}$	2.98	$37.57_{1.13}$	$35.14_{1.04}$	2.43
			MiniCPM-V2.5	8B	$1.74_{0.08}$	$1.61_{0.08}$	0.12	$11.55_{0.24}$	$11.69_{0.38}$	-0.15
			MiniCPM-V2.5-FT	8B	$11.43_{0.11}$	$14.29_{0.16}$	-2.86	$35.13_{0.19}$	$36.65_{0.68}$	-1.52
			Monkey	7B	$1.37_{0.05}$	$2.24_{0.15}$	-0.87	$13.16_{0.18}$	$14.45_{0.24}$	-1.29
			Pixtral	12B	$0.25_{0.19}$	$0.62_{0.28}$	-0.37	$10.04_{0.41}$	$11.21_{0.45}$	-1.17
			Qwen-VL	7B	$1.61_{0.03}$	$1.74_{0.03}$	-0.12	$15.28_{0.13}$	$14.43_{0.54}$	0.85
			Qwen2-VL	7B	$\mathbf{76.27_{1.49}}$	$\mathbf{75.65_{1.45}}$	0.62	$\mathbf{86.56_{1.00}}$	$\mathbf{86.77_{0.93}}$	-0.22
			Yi-VL	34B	$0.12_{0.01}$	$0.0_{0.0}$	0.12	$4.31_{0.08}$	$5.45_{0.13}$	-1.14
			Yi-VL	6B	$0.12_{0.02}$	$0.0_{0.0}$	0.12	$4.49_{0.05}$	$5.7_{0.12}$	-1.21
Chinese	Easy	Closed	Claude 3 Opus	-	$0.9_{0.3}$	$1.0_{0.31}$	-0.1	$11.5_{0.48}$	$10.0_{0.49}$	1.49
			Claude 3.5 Sonnet	-	$1.0_{0.31}$	$0.8_{0.28}$	0.2	$7.54_{0.54}$	$7.5_{0.51}$	0.03
			Gemini 1.5 Pro	-	$1.1_{0.32}$	$0.5_{0.22}$	0.6	$11.1_{0.56}$	$11.47_{0.48}$	-0.37
			GPT-4o	-	$\mathbf{14.87_{1.14}}$	$\mathbf{22.46_{1.35}}$	-7.58	$\mathbf{39.05_{0.99}}$	$\mathbf{48.24_{1.09}}$	-9.19
			GPT-4 Turbo	-	$0.2_{0.14}$	$0.1_{0.1}$	0.1	$8.42_{0.36}$	$6.97_{0.29}$	1.45
			Qwen-VL-Max	-	$6.34_{0.08}$	$9.92_{0.09}$	-3.58	$13.45_{0.41}$	$22.86_{0.46}$	-9.42
			Reka Core	-	$0.0_{0.0}$	$0.0_{0.0}$	0	$3.43_{0.26}$	$3.15_{0.2}$	0.28
		Open	CogVLM2-Chinese	19B	$33.63_{0.15}$	$31.44_{0.19}$	2.2	$57.97_{0.56}$	$54.05_{0.54}$	3.92
			CogVLM2-Chinese-FT	19B	$\underline{63.97_{0.55}}$	$\underline{62.67_{0.17}}$	1.3	$\underline{79.71_{0.41}}$	$\underline{79.22_{0.47}}$	0.49
			DeepSeek-VL	1.3B	$0.0_{0.0}$	$0.0_{0.0}$	0	$6.1_{0.1}$	$3.25_{0.05}$	2.85
			DeepSeek-VL	7B	$0.0_{0.0}$	$0.0_{0.0}$	0	$4.28_{0.07}$	$7.3_{0.05}$	-3.02
			DocOwl-1.5-Omni	8B	$0.0_{0.0}$	$0.0_{0.0}$	0	$1.19_{0.05}$	$3.83_{0.06}$	-2.63
			InternLM-XComposer2-VL	7B	$0.6_{0.05}$	$0.2_{0.04}$	0.4	$12.34_{0.25}$	$12.52_{0.14}$	-0.18
			InternLM-XComposer2-VL-4K	7B	$0.20_{0.14}$	$0.10_{0.10}$	0.10	$11.93_{0.42}$	$13.68_{0.41}$	-1.74
			InternLM-XComposer2.5-VL	7B	$0.30_{0.17}$	$0.40_{0.20}$	-0.10	$12.76_{0.42}$	$14.99_{0.43}$	-2.23
			InternVL-V2	40B	$22.75_{1.36}$	$16.67_{1.14}$	6.09	$49.51_{1.06}$	$39.46_{1.10}$	10.05
			MiniCPM-V2.5	8B	$4.59_{0.11}$	$4.89_{0.09}$	-0.3	$18.12_{0.33}$	$22.28_{0.18}$	-4.17
			MiniCPM-V2.5-FT	8B	$7.29_{0.14}$	$7.09_{0.12}$	0.2	$29.36_{0.39}$	$30.67_{0.38}$	-1.31
			Monkey	7B	$0.2_{0.01}$	$1.4_{0.05}$	-1.2	$7.89_{0.3}$	$10.26_{0.24}$	-2.37
			Qwen-VL	7B	$0.0_{0.0}$	$0.0_{0.0}$	0	$1.25_{0.03}$	$0.43_{0.06}$	0.82
			Qwen2-VL	7B	$\mathbf{61.08_{1.52}}$	$\mathbf{68.16_{1.46}}$	-7.09	$\mathbf{78.38_{0.96}}$	$\mathbf{83.48_{0.84}}$	-5.10
			Yi-VL	34B	$0.0_{0.0}$	$0.0_{0.0}$	0	$4.69_{0.09}$	$1.71_{0.06}$	2.98
			Yi-VL	6B	$0.0_{0.0}$	$0.0_{0.0}$	0	$4.28_{0.06}$	$1.66_{0.04}$	2.62
	Hard	Closed	Claude 3 Opus	-	$0.3_{0.18}$	$0.1_{0.1}$	0.2	$9.22_{0.38}$	$8.09_{0.33}$	1.13
			Claude 3.5 Sonnet	-	$0.2_{0.15}$	$0.0_{0.0}$	0.2	$4.0_{0.33}$	$2.37_{0.23}$	1.63
			Gemini 1.5 Pro	-	$0.7_{0.26}$	$0.5_{0.23}$	0.2	$11.82_{0.51}$	$11.75_{0.44}$	0.07
			GPT-4o	-	$\mathbf{2.2_{0.47}}$	$\mathbf{1.8_{0.4}}$	0.4	$\mathbf{22.72_{0.67}}$	$\mathbf{22.89_{0.65}}$	-0.17
			GPT-4 Turbo	-	$0.0_{0.0}$	$0.2_{0.13}$	-0.2	$8.58_{0.3}$	$6.87_{0.28}$	1.72
			Qwen-VL-Max	-	$0.89_{0.06}$	$1.38_{0.1}$	-0.49	$5.4_{0.19}$	$12.29_{0.18}$	-6.89
			Reka Core	-	$0.0_{0.0}$	$0.0_{0.0}$	0	$3.35_{0.23}$	$2.97_{0.2}$	0.38
		Open	CogVLM2-Chinese	19B	$1.2_{0.07}$	$2.3_{0.09}$	-1.1	$16.83_{0.22}$	$19.86_{0.23}$	-3.04
			CogVLM2-Chinese-FT	19B	$\underline{42.51_{0.32}}$	$\underline{45.91_{0.23}}$	-3.39	$\underline{65.79_{0.24}}$	$\underline{69.46_{0.46}}$	-3.68
			DeepSeek-VL	1.3B	$0.0_{0.0}$	$0.0_{0.0}$	0	$6.87_{0.09}$	$3.53_{0.07}$	3.33
			DeepSeek-VL	7B	$0.0_{0.0}$	$0.0_{0.0}$	0	$5.49_{0.07}$	$7.57_{0.05}$	-2.08
			DocOwl-1.5-Omni	8B	$0.0_{0.0}$	$0.0_{0.0}$	0	$1.68_{0.04}$	$4.42_{0.07}$	-2.73
			InternLM-XComposer2-VL	7B	$0.0_{0.0}$	$0.0_{0.0}$	0	$8.36_{0.09}$	$7.92_{0.09}$	0.44
			InternLM-XComposer2-VL-4K	7B	$0.00_{0.00}$	$0.00_{0.00}$	0.00	$7.49_{0.31}$	$7.25_{0.30}$	0.25
			InternLM-XComposer2.5-VL	7B	$0.00_{0.00}$	$0.00_{0.00}$	0.00	$10.83_{0.31}$	$10.81_{0.31}$	0.02
			InternVL-V2	40B	$0.40_{0.20}$	$0.90_{0.29}$	-0.50	$12.30_{0.42}$	$13.80_{0.48}$	-1.50
			MiniCPM-V2.5	8B	$0.2_{0.03}$	$0.2_{0.01}$	0	$7.23_{0.18}$	$7.6_{0.13}$	-0.37
			MiniCPM-V2.5-FT	8B	$1.2_{0.03}$	$1.4_{0.06}$	-0.2	$18.01_{0.35}$	$15.25_{0.25}$	2.76
			Monkey	7B	$0.0_{0.0}$	$0.0_{0.0}$	0	$5.69_{0.15}$	$6.3_{0.13}$	-0.61
			Qwen-VL	7B	$0.0_{0.0}$	$0.0_{0.0}$	0	$1.1_{0.07}$	$0.15_{0.01}$	0.94
			Qwen2-VL	7B	$\mathbf{18.76_{1.22}}$	$\mathbf{26.75_{1.40}}$	-7.98	$\mathbf{43.84_{1.10}}$	$\mathbf{53.56_{1.09}}$	-9.72
			Yi-VL	34B	$0.0_{0.0}$	$0.0_{0.0}$	0	$4.49_{0.09}$	$1.73_{0.1}$	2.76
			Yi-VL	6B	$0.0_{0.0}$	$0.0_{0.0}$	0	$3.95_{0.05}$	$2.08_{0.09}$	1.87

B MORE RELATIONSHIP BETWEEN VCR AND OTHER BENCHMARKS

The heatmap shown in Figure 6 provides a detailed view of the pairwise correlation between 23 different benchmarks used to evaluate 38 VLMs. The scores of these models on each benchmark were utilized to compute this correlation matrix. The color intensity and numerical values represent the degree of correlation, ranging from -1 (perfect negative correlation) to 1 (perfect positive correlation), with warmer colors indicating higher positive correlations and cooler colors indicating weaker or negative correlations.

We observe that $\text{VCR}_{\text{EN, HARD}}$ is markedly different from other benchmarks in the evaluation set. It demonstrates minimal and even sometimes negative correlation with other benchmarks. This suggests that the skill set required for $\text{VCR}_{\text{EN, HARD}}$ is largely unrelated to those tested by other popular tasks emphasizing more straightforward image-to-text associations or OCR capabilities. Specifically, $\text{VCR}_{\text{EN, HARD}}$ challenges models with tasks that involve high-level reasoning and minimal reliance on pixel-level information, focusing instead on understanding context, commonsense reasoning, and visual narrative interpretation. These features are less critical in other benchmarks, which explains the weak correlation across tasks. $\text{VCR}_{\text{EN, EASY}}$ exhibits a slightly stronger correlation with a few other benchmarks but remains moderately independent of most others. Like $\text{VCR}_{\text{EN, HARD}}$, $\text{VCR}_{\text{EN, EASY}}$ also evaluates visual commonsense reasoning, but with less stringent requirements, offering models more cues and simpler connections between visual elements and textual understanding. This leads to moderate overlap with benchmarks like TextVQA, which similarly focus on text understanding in a visual context, but $\text{VCR}_{\text{EN, EASY}}$ still emphasizes a higher level of interpretative reasoning than standard text-based vision benchmarks.

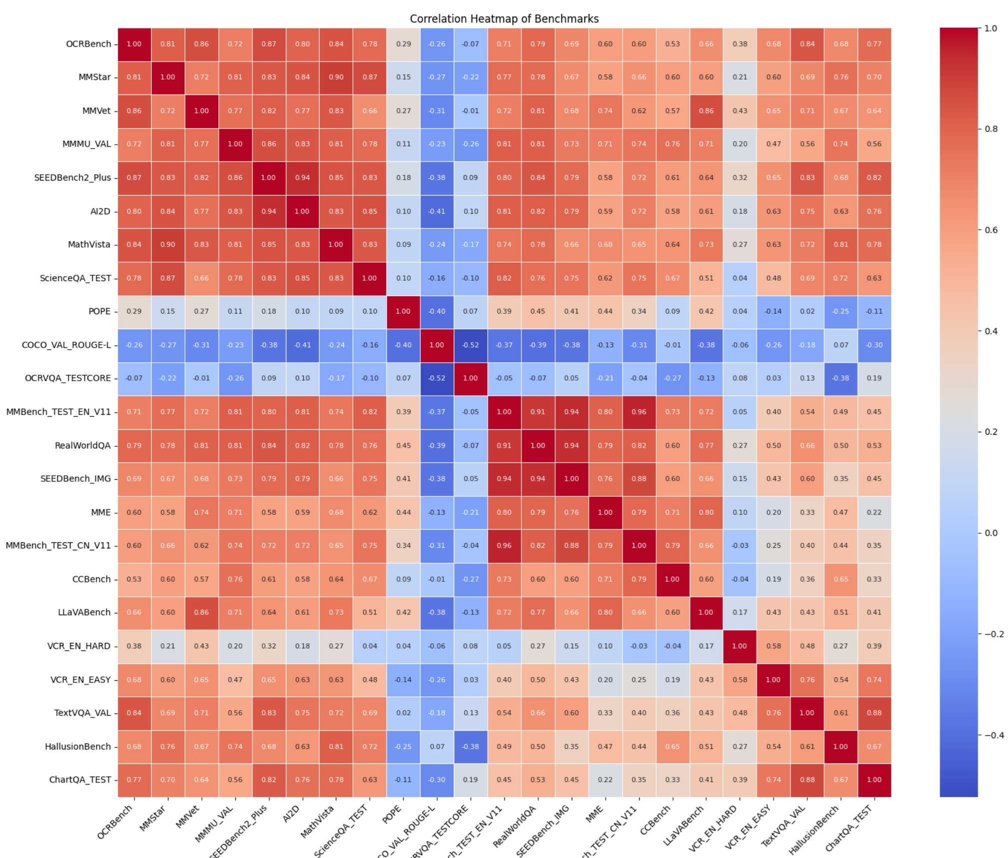

Figure 6: The heat map of benchmarks displays the correlation between the metric scores of 38 models for each benchmark pair.

C INFORMATION OF MODELS EVALUATED

We provide the specifications of the evaluated models and a link to each model's weights in Table 6.

Table 6: Model specifications

Model name	Model size	Open-sourced
Claude 3 Opus	-	×
Claude 3.5 Sonnet	-	×
Gemini 1.5 Pro	-	×
GPT-4 Turbo	-	×
GPT-4o	-	×
GPT-4V	-	×
Qwen-VL-Max	-	×
Reka Core	-	×
Cambrian-1 [5]	34B	✓
CogVLM2 [6]	19B	✓
CogVLM2-Chinese [7]	19B	✓
DeepSeek-VL [8]	1.3B	✓
DeepSeek-VL [9]	7B	✓
DocOwl-1.5-Omni[10]	8B	✓
Idefics3 [11]	8B	✓
InternLM-XComposer2-VL [12]	7B	✓
InternLM-XComposer2-VL-4KHD [13]	7B	✓
InternVL-V2 [14]	40B	✓
Llama-3.2-Vision [15]	11B	✓
Llama-3.2-Vision [16]	90B	✓
MiniCPM-V2.5 [17]	8B	✓
Monkey[18]	7B	✓
Pixtral [19]	12B	✓
Qwen-VL [20]	7B	✓
Qwen2-VL [21]	7B	✓
Yi-VL[22]	34B	✓
Yi-VL[23]	6B	✓

[5]https://huggingface.co/nyu-visionx/cambrian-34b
[6]https://huggingface.co/THUDM/CogVLM2-Llama3-chat-19B
[7]https://huggingface.co/THUDM/cogvlm2-llama3-Chinese-chat-19B
[8]https://huggingface.co/deepseek-ai/deepseek-vl-1.3b-chat
[9]https://huggingface.co/deepseek-ai/deepseek-vl-7b-chat
[10]https://huggingface.co/mPLUG/DocOwl1.5-Omni
[11]https://huggingface.co/HuggingFaceM4/Idefics3-8B
[12]https://huggingface.co/internlm/internlm-xcomposer2-vl-7b
[13]https://huggingface.co/internlm/internlm-xcomposer2-4khd-7b
[14]https://huggingface.co/OpenGVLab/InternVL2-40B
[15]https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct
[16]https://huggingface.co/meta-llama/Llama-3.2-90B-Vision-Instruct
[17]https://huggingface.co/OpenBMB/MiniCPM-Llama3-V-2_5
[18]https://huggingface.co/echo840/Monkey-Chat
[19]https://huggingface.co/mistralai/Pixtral-12B-2409
[20]https://huggingface.co/Qwen/Qwen-VL-Chat
[21]https://huggingface.co/Qwen/Qwen2-VL-7B-Instruct
[22]https://huggingface.co/01-ai/Yi-VL-34B
[23]https://huggingface.co/01-ai/Yi-VL-6B

D Potential QA

What could be the possible reason that CogVLM performs well in VCR-WIKI series benchmarks? Many models we tested (DocOwl-1.5, Monkey, MiniCPM-V2.5, InternLM series, InternVL series) follow a similar inference pipeline to adapt to high-resolution application scenarios:

1. An algorithm divides the input image into segments.
2. Each segment is encoded into tokens using a CILP-based image encoder.
3. A filtering mechanism (algorithm/resampler/abstractor) processes the visual tokens.
4. The filtered tokens are concatenated with language tokens and input to the LLM.

If, in step 3, pixel-level hints embedded in text within the image (TEL) are disregarded, the model cannot correctly answer the question. Consequently, some of these models may perform better on benchmarks emphasizing global features but struggle on the VCR-WIKI series benchmarks, particularly in the hard partitions. For example, while InternVL2-40B performs best on $\text{VCR}_{\text{EN, EASY}}$, it does not perform well on $\text{VCR}_{\text{EN, HARD}}$. As noted in the paper, the easy partition of the benchmark primarily verifies that the VCR task is feasible for the models. In contrast, the hard partition explores the boundaries of VCR capability for both models and human test-takers (who require more time and focus to solve the puzzles in the hard partition).

The CogVLM2 and Cambrian-1 series, by contrast, do not include step 3 in their inference pipelines. Instead, their image encoders operate at mid-to-high resolutions (1K level), and they resize the input image to match the supported resolution rather than dividing it into segments. The image encoder resolution for CogVLM2 is 1344×1344, while Cambrian-1 employs four image encoders, the largest supporting a resolution of 1024×1024. This approach may encounter challenges with extremely shaped input images (e.g., 8192×1024), but for VCR-WIKI, where images are mostly near-square (on average 300×360 for $\text{VCR}_{\text{ZH, EASY}}$/$\text{VCR}_{\text{ZH, HARD}}$ and 300×375 for $\text{VCR}_{\text{EN, EASY}}$/$\text{VCR}_{\text{EN, HARD}}$), high-resolution support is not necessary. For instance, InternLM-XComposer2-VL outperforms InternLM-XComposer2-VL-4KHD on this benchmark.

What could be the potential way to improve models' capability on VCR? To suggest potential avenues for improving VLM performance on VCR, we propose the following:

1. **Include VCR in VLM Pretraining**: Just as OCR parsing tasks are often included in pre-training to improve OCR performance, researchers could consider incorporating VCR tasks during pretraining. We will codebase to facilitate this process, making it as straightforward as data augmentation.
2. **Architectural Exploration**: CogVLM2 is the best-performing model on average across the four partitions, and we believe this is largely due to its vision expert architecture. We contacted the CogVLM2 team and learned that GLM-4 and CogVLM2 share the same training data, yet there is a significant performance gap between them on the VCR benchmarks.
3. **Chain-of-Thought (CoT) Methods**: Researchers could explore multi-modality pipelines based on CoT techniques to improve existing VLMs on VCR tasks Chen et al. (2024a); Zhang et al. (2023). Although a model might not initially focus on the correct visual area (e.g., pixel-level hints in the TEI), CoT-based techniques could help refine its focus over successive rounds.

How is the Transferability of VCR-WIKI Finetuning? In Table 7, we show the transferability of VCR-WIKI by finetuning multiple models on different finetuning datasets' training sets and testing their performance on a series of benchmarks.

The analysis of our experimental results highlights the strong transferability of the proposed VCR-WIKI dataset across various benchmarks. Notably, models fine-tuned on VCR-WIKI demonstrate significant performance improvements not only within the VCR-WIKI benchmarks themselves, but also across different language settings. For example, fine-tuning CogVLM2 on $\text{VCR}_{\text{EN, HARD}}$ leads to a substantial increase in performance on the $\text{VCR}_{\text{ZH, EASY}}$ benchmark, elevating the score from 9.15 to 42.55. Similarly, fine-tuning the Chinese version of CogVLM2 on $\text{VCR}_{\text{ZH, HARD}}$ enhances its performance on both the $\mathrm{VCR}_{\mathrm{EN,\ EASY}}$ and $\mathrm{VCR}_{\mathrm{EN,\ HARD}}$ benchmarks, with scores rising from 79.9 to 87.57 and from 25.13 to 44.97, respectively. These enhancements indicate that the VCR-wiki dataset facilitates the learning of effective transferable features even when the fine-tuning and evaluation involve different languages.

Additionally, the consistent achievement of the highest scores within each model's finetuning variations underscores the robustness of VCR-WIKI in improving model performance across diverse evaluation metrics. This evidence collectively demonstrates that VCR-WIKI serves as a versatile and powerful resource for enhancing model generalization and performance across multiple tasks and linguistic contexts.

Table 7: Performance Comparison of Base Models.

Base Model	**MiniCPM-V2.5**	**MiniCPM-V2.5**	**MiniCPM-V2.5**	**MiniCPM-V2.5**	**CogVLM2**	**CogVLM2**	**CogVLM2**	**CogVLM2-Ch.**	**CogVLM2-Ch.**	**CogVLM2-Ch.**
Finetuning_dataset	None	OKVQA-Train	$\mathrm{VCR}_{\mathrm{EN,\ HARD}}$	$\mathrm{VCR}_{\mathrm{ZH,\ HARD}}$	None	OKVQA-Train	$\mathrm{VCR}_{\mathrm{EN,\ HARD}}$	None	OKVQA-Train	$\mathrm{VCR}_{\mathrm{ZH,\ HARD}}$
OKVQA	77.43	72.20	**77.09**	76.38	75.35	71.86	**75.45**	**74.16**	70.57	74.14
$\mathrm{VCR}_{\mathrm{EN,\ EASY}}$	31.81	19.53	**40.96**	28.40	83.25	79.29	**93.27**	79.90	30.18	**87.57**
$\mathrm{VCR}_{\mathrm{EN,\ HARD}}$	1.41	0.00	**13.86**	5.33	37.98	27.22	**77.44**	25.13	3.55	**44.97**
$\mathrm{VCR}_{\mathrm{ZH,\ EASY}}$	4.10	2.12	2.66	**7.44**	9.15	16.49	**42.55**	33.24	16.49	**61.69**
$\mathrm{VCR}_{\mathrm{ZH,\ HARD}}$	0.09	0.53	1.06	**1.53**	0.08	0.00	**1.60**	1.34	0.00	**42.11**
MMstar	50.20	**51.73**	50.40	50.27	50.50	**51.07**	50.20	52.73	50.87	**54.33**
MMBench_DEV_EN	**74.54**	74.46	**74.54**	74.30	72.70	**73.53**	72.60	**77.32**	77.09	76.78
MME	**2024.6**	1996.7	1923.65	1977.13	1869.5	1860.56	**1882.21**	**2040.7**	1898.42	1939.8
MMMU_VAL	45.89	**47.78**	46.11	46.56	**42.60**	38.67	38.67	42.44	41.56	**45.00**
AI2D_test	**78.04**	77.85	77.62	77.91	73.40	**74.97**	73.61	**72.64**	70.98	71.31
OCR_BENCH	**71.70**	71.60	71.50	71.30	75.40	72.80	**79.80**	77.30	75.20	**79.60**
MMVet	**53.12**	45.78	51.10	52.66	57.80	45.00	**59.31**	56.38	39.77	**56.88**
MathVista_MINI	**54.50**	53.70	52.80	52.70	**38.60**	38.30	35.40	37.80	37.60	**40.20**
ChartQA	71.80	72.12	71.92	**72.52**	72.80	**80.40**	79.56	63.16	58.92	**65.84**
OCRVQA	61.85	**62.70**	61.36	61.59	64.90	65.56	**66.11**	32.65	**34.41**	33.24

E CASE STUDY

This case study aims to assess the real-world applicability of models fine-tuned on VCR-WIKI for recognizing occluded text, a challenging task with significant practical implications. The setting involves evaluating the performance of three state-of-the-art models—MiniCPM-V2.5 8B, CogVLM2 19B, and Qwen2-VL 7B—on a curated dataset of eleven photographs featuring occluded text from real-world scenarios, such as street maps and collected images. Since no standard benchmark exists for real-world occluded text recognition, this dataset serves as a proxy to measure the efficacy of VCR fine-tuning in improving performance. We show whether the model completely recovers the occluded or distorted texts in the image with ✓ (correct) or × (partially correct or incorrect). This evaluation provides insight into how VCR fine-tuning translates to practical challenges and complements the quantitative analyses presented in the main paper.

Figure 7: Ground-truth: BANK OF AMERICA TWO BRYANT PARK

- MiniCPM: ✓
- MiniCPM-ft: ✓
- CogVLM2: ✓
- CogVLM2-ft: ✓
- Qwen2-VL-7B: ✓
- Qwen2-VL-7B-ft: ✓

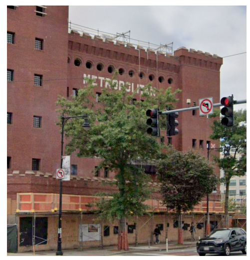

Figure 8: Ground-truth: METROPOLITAN

- MiniCPM: ✓
- MiniCPM-ft: ✓
- CogVLM2: ✓
- CogVLM2-ft: ✓
- Qwen2-VL-7B: ✓
- Qwen2-VL-7B-ft: ✓

Figure 9: Ground-truth: NOT IN SERVICE

- MiniCPM: ×
- MiniCPM-ft: ✓
- CogVLM2: ✓
- CogVLM2-ft: ✓
- Qwen2-VL-7B: ✓
- Qwen2-VL-7B-ft: ✓

Figure 10: Ground-truth: NO CYCLING

- MiniCPM: ×
- MiniCPM-ft: ×
- CogVLM2: ✓
- CogVLM2-ft: ✓
- Qwen2-VL-7B: ×
- Qwen2-VL-7B-ft: ✓

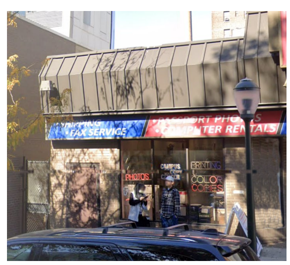

Figure 11: Ground-truth: SHIPPING FAX SERVICE PASSPORT PHOTOS COMPUTER RENTALS

- MiniCPM: ×
- MiniCPM-ft: ×
- CogVLM2: ✓
- CogVLM2-ft: ✓
- Qwen2-VL-7B: ✓
- Qwen2-VL-7B-ft: ✓

Figure 12: Ground-truth: Home of Peapack Private

- MiniCPM: ✓
- MiniCPM-ft: ✓
- CogVLM2: ✓
- CogVLM2-ft: ✓
- Qwen2-VL-7B: ✓
- Qwen2-VL-7B-ft: ✓

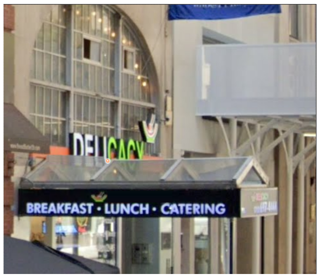

Figure 13: Ground-truth:
DELICACY BREAKFAST LUNCH CATERING

- MiniCPM: ✓
- MiniCPM-ft: ✓
- CogVLM2: ✓
- CogVLM2-ft: ✓
- Qwen2-VL-7B: ✓
- Qwen2-VL-7B-ft: ✓

Figure 14: Ground-truth:
**INSPECT UPON RECEIPT...
DO NOT SIGN...
VISIBLE DAMAGE?**

- MiniCPM: ×
- MiniCPM-ft: ×
- CogVLM2: ✓
- CogVLM2-ft: ✓
- Qwen2-VL-7B: ✓
- Qwen2-VL-7B-ft: ✓

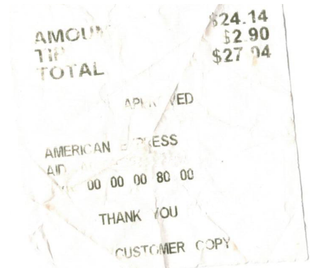

Figure 15: Ground-truth:
**AMOUNT TIP TOTAL APPROVED
AMERICAN EXPRESS AID
THANK YOU / MERCI
CUSTOMER COPY**

- MiniCPM: ×
- MiniCPM-ft: ✓
- CogVLM2: ✓
- CogVLM2-ft: ✓
- Qwen2-VL-7B: ✓
- Qwen2-VL-7B-ft: ✓

Figure 16: Ground-truth:
**INSPECT UPON RECEIPT...
DO NOT SIGN...
VISIBLE DAMAGE?**

- MiniCPM: ×
- MiniCPM-ft: ×
- CogVLM2: ✓
- CogVLM2-ft: ✓
- Qwen2-VL-7B: ✓
- Qwen2-VL-7B-ft: ✓

Figure 17: Ground-truth:
**AMOUNT TIP TOTAL APPROVED
AMERICAN EXPRESS AID
THANK YOU / MERCI
CUSTOMER COPY**

- MiniCPM: ×
- MiniCPM-ft: ✓
- CogVLM2: ✓
- CogVLM2-ft: ✓
- Qwen2-VL-7B: ✓
- Qwen2-VL-7B-ft: ✓