VCR: Visual Caption Restoration

Anonymous Author(s)
Affiliation
Address
email

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 is intrinsically different 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. Solving VCR likely requires complex system-2 level reasoning capability, which existing models struggle with, while humans excel. We release VCR-WIKI and the construction code to promote further research in this area.

1 Introduction

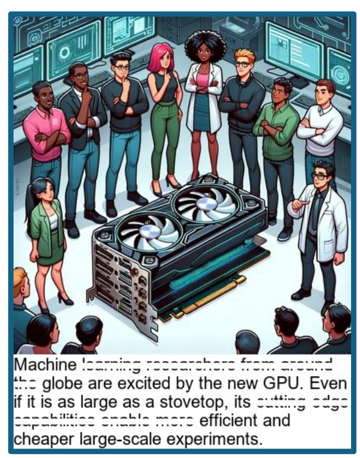

Figure 1: An example of the VCR task.

Recent advances in large language models, such as ChatGPT [39, 38] and Llama [48], have spurred significant interest and progress in the field of vision-language models. With models like GPT-4V [38] and LLaVA [26, 27, 28] 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.

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, a task that 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 [47, 40, 49, 13, 24]. 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 [3, 51, 35, 43] 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 introduces a distinct challenge: restoring occluded text in images. This task taps into system-2 reasoning, which involves complex cognitive processes that go beyond the quick, reflexive responses typical of system-1 reasoning. System-2 reasoning requires deep thinking, logical analysis, and integration of multiple types of information, similar to the capabilities needed to solve the VCR task. Besides, 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 [43, 7]. 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 that need complex System-2 level reasoning.
- **C2** Develop a pipeline for generating synthetic images with embedded text that allows for adjusting 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 3 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 [13, 24]. 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.

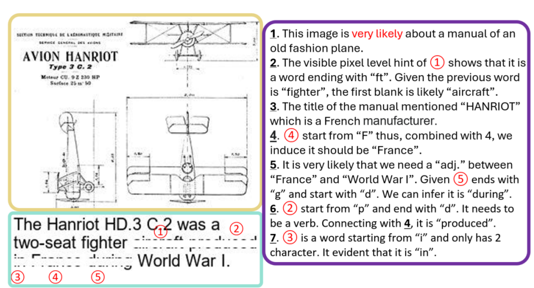

Figure 2: An example of how humans would solve the VCR task.

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

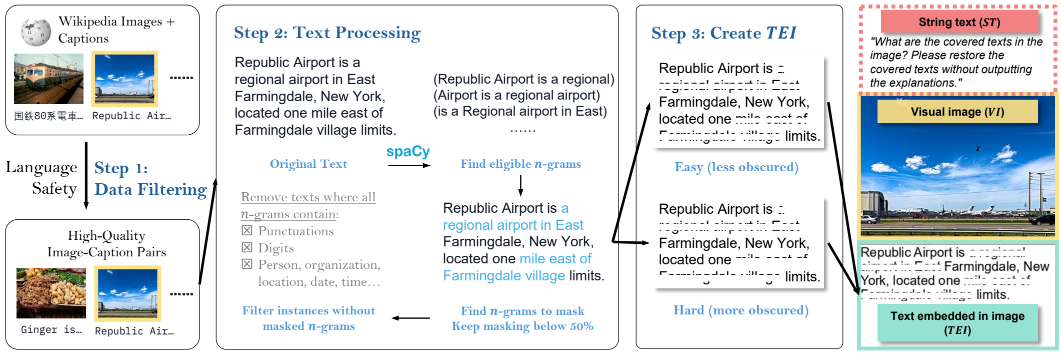

Figure 3: Illustration of the dataset creation pipeline for VCR-WIKI. visual image (VI), text embedded in image (TEI) and string text (ST) in an example of the English Hard configuration of VCR-WIKI. The solid line-enclosed contents (VI and TEI) are part of the image, whereas the dotted line-enclosed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

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 and System-2 level reasoning skill, akin to human cognitive processes where context and visual clues guide understanding and response. Below we show an example of how humans would solve this task in 'hard' difficulty in Figure 2.

3 Dataset Creation

The VCR task aligns visual images (VI) with text embedded in images (ET) by using highly correlated image-text pairs. We create VCR-WIKI, a Wikipedia-based VCR dataset, using images and captions from Wikipedia[1], filtering out sensitive content such as NSFW and crime-related terms. Each instance includes a stacked VI+ET image and a question-answer pair, mimicking a VQA format. The VI+ET images are resized to 300 pixels wide, with ET truncated to five lines to avoid excessive height. We exclude instances where VI+ET exceeds 900 pixels in height.

For masking within ET, we randomly select 5-grams using spaCy, excluding terms like numbers, names, locations, etc. The 5-grams are partially obscured to vary task difficulty, but the total masked tokens don't exceed 50% of the caption. Images without an eligible 5-gram are excluded. An ablation version retains only the ET portion to assess the impact of VI on model performance.

The task involves a question prompting the model to restore the obscured text, with ground truth corresponding to the visible caption. The dataset supports both English and Chinese, offering two difficulty levels: an easy version where OCR models fail but native speakers succeed, and a hard version with minimal visible text. The dataset is released under CC BY-SA 4.0 but is not linked due to anonymity. Please refer to Appendix C for more details.

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) [39, 38], Claude 3 Opus (claude-3-opus-20240229), Claude 3.5 Sonnet (claude-3-5-sonnet-20240620) [2], Gemini 1.5 pro (gemini-1.5-pro-001) [45], Reka Core (reka-core-20240501) [46], and Qwen-VL-Max (tested on May 2024) [4].

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

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 InternVL-Chat-V1.5[10], MiniCPM-Llama3-V2.5 [18], InternLM-XComposer2-VL-7B [12], CogVLM2-Llama3-19B-Chat [52], Idefics2-8B [23], Yi-VL-34B [1], Yi-VL-6B [1], Qwen-VL-Chat [4], DeepSeek-VL-7B-Chat [30], DeepSeek-VL-1.3B-Chat [30], Monkey [29, 25] and DocOwl-1.5 [15]. Out of these models, Idefics2-8B is an English-only model, 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 two models from the open-sourced models, CogVLM2-Llama3-19B-Chat and MiniCPM-Llama3-V2.5, and fine-tune them on a subset of VCR's training set.

More specifically, we fine-tune CogVLM2-Llama3-19B-Chat and MiniCPM-Llama3-V2.5 in the English Hard configuration, and CogVLM2-Llama3-19B-Chinese-Chat and MiniCPM-Llama3-V2.5 on the Chinese Hard configuration. The models are finetuned using LoRA [16] with $r = 8$ and $\alpha = 32$. We adopt the schedule-free AdamW optimizer [11] 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 C). 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), which measures whether the restored n-gram $\hat{m}$ totally matches the ground-truth m; and **Jaccard Index** (J), which measures the similarity of $\hat{m}$ and m as bag-of-words.

- **Exact Match** (EM), which measures whether the restored n-gram $\hat{m}$ totally matches the ground-truth m;

$$EM(m, \hat{m}) = \begin{cases} 1 & \text{if } m = \hat{m}, \\ 0 & \text{otherwise} \end{cases}.$$

- **Jaccard Index** (J), which is a more relaxed metric that measures the similarity of $\hat{m}$ and m as bag-of-words.

$$J(m, \hat{m}) = \frac{|S(m) \cap S(\hat{m})|}{|S(m) \cup S(\hat{m})|},$$

  where $S(m)$ represents the set of tokens in m.

4.3 Experimental Results

Please refer to the exact match score and the Jaccard-index of the evaluation in Table 2.

Open-Source Models. We evaluate each open-source model based on the whole 5,000 examples in the test set. Note that Idefics2-8B only supports the English task, hence it has no evaluation score on the Chinese task.

Although achieving state-of-the-art performance on the Open VLM leaderboard, almost all the tested models achieve a low exact match accuracy in the English Easy configuration and fail on the other settings. The best open-source model across the 4 configurations (English Easy, English Hard, Chinese Easy, and Chinese Hard) is CogVLM2-Llama3-Chat. This might be attributed to

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

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

its pretraining process and the special architecture. We also notice that VI has a negative impact for most models on the exact match scores ($\Delta < 0$), which means that the image information is not properly utilized. The best performed open-source model on average, CogVLM2-Llama3-Chat and CogVLM2-Llama3-Chinese-Chat, and its fine-tuned version have positive Δ, except for the Chinese Hard configuration. This indicates that information from VI could help improve the model performance on VCR.

For different languages, we noticed a large performance drop when testing the model in Chinese configurations, despite the fact that all models claim to have basic English-Chinese duolingual capabilities. This is somehow surprising, since Chinese characters, due to their logographic nature, may exhibit a higher degree of recognizability compared to languages that use alphabetic scripts in one order [54, 62].

Moreover, we found that models, such as internlm-xcomposer2, are good at OCR and understanding image documents (as demonstrated by DocOwl 1.5 and Monkey) still have the potential to be improved in the VCR task. This highlights the unique and indispensable role of VCR in the current suite of benchmarks. Excelling in other document-related benchmarks does not guarantee similar performance in VCR tasks, emphasizing VCR's distinct challenges and value.

Closed-Source Models. We evaluate every closed-source model with the first 500 examples in the test set. In English tasks, GPT-4o scores the best among the models that have not been finetuned. Even though GPT-4 series support Chinese, we found that GPT-4V (gpt-4-1106-vision-preview) is not able to recognize Chinese characters embedded in the image even without any occlusion. Thus, we do not test GPT-4V on Chinese tasks.

In English configurations, closed-source models outperform all open-source models except CogVLM2, which indicates that model scaling might help improve performance on the VCR task. However, compared with the human evaluation results in Section 4.4, we notice a large performance gap, especially in the English Hard configuration. This shows significant room for improvement in the current state-of-the-art models.

Please refer to Table 4 to compare open and closed source models using the same 500 test cases.

4.4 Human Evaluation

We recruited 7 volunteers to perform human evaluation on a subset of the samples of 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 volunteer the following instructions: (1) We ask 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 ask 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 count errors not related to date and person names (Filtered).

Table 1: Human evaluation results on the VCR task for in terms of exact matches. N is the number of puzzles in each language.

	EN Easy (N = 169)		**EN Hard (N = 169)**		**ZH Easy (N = 188)**		**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

Refer to Table 3 to compare all models with human evaluation results using the same test cases.

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

Table 2: Performance of vision language models on VCR task in English (EN) and Chinese (ZH), for easy and hard modes. FT indicates finetuning on 16,000 VCR-wiki samples. Best results: underlined (finetuned), bold (non-finetuned). Subscripts show bootstrap 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
			CogVLM2-FT	19B	$\underline{93.27_{0.03}}$	$\underline{92.63_{0.07}}$	0.64	$\underline{97.62_{0.02}}$	$\underline{97.4_{0.01}}$	0.22
			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
			Monkey	7B	$50.66_{0.1}$	$56.2_{0.08}$	-5.54	$67.6_{0.09}$	$72.82_{0.08}$	-5.22
			Idefics2	8B	$15.75_{0.11}$	$27.77_{0.11}$	-12.02	$31.97_{0.02}$	$51.0_{0.03}$	-19.03
			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-XComposer2.5-VL	7B	$41.35_{0.55}$	$25.37_{0.51}$	15.97	$63.04_{0.42}$	$49.95_{0.41}$	13.09
			InternVL-V1.5	25.5B	$14.65_{0.13}$	$75.06_{0.1}$	-60.41	$51.42_{0.04}$	$87.1_{0.03}$	-35.68
			InternVL-V2	25.5B	$74.51_{0.48}$	$77.79_{0.47}$	-3.27	$86.74_{0.28}$	$89.02_{0.26}$	-2.28
			InternVL-V2	40B	$\mathbf{84.67_{0.40}}$	$\mathbf{87.71_{0.37}}$	-3.04	$\mathbf{92.64_{0.22}}$	$\mathbf{95.10_{0.16}}$	-2.47
			MiniCPM-V2.5	8B	$31.81_{0.08}$	$40.05_{0.09}$	-8.25	$53.24_{0.1}$	$63.2_{0.1}$	-9.96
			MiniCPM-V2.5-FT	8B	$40.96_{0.14}$	$44.62_{0.07}$	-3.67	$64.4_{0.05}$	$67.62_{0.1}$	-3.22
			Qwen-VL	7B	$49.71_{0.17}$	$52.15_{0.15}$	-2.44	$69.94_{0.07}$	$72.28_{0.08}$	-2.34
			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	-	$\mathbf{73.2_{0.16}}$	$\mathbf{82.43_{0.17}}$	-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	$27.20_{0.48}$	$\mathbf{29.68_{0.50}}$	-2.48	$50.04_{0.40}$	$\mathbf{55.66_{0.39}}$	-5.62
			CogVLM2	19B	$\mathbf{37.98_{0.18}}$	$17.68_{0.06}$	20.3	$\mathbf{59.99_{0.05}}$	$39.69_{0.03}$	20.3
			CogVLM2-FT	19B	$\underline{77.44_{0.05}}$	$\underline{66.07_{0.13}}$	11.38	$\underline{90.17_{0.03}}$	$\underline{83.41_{0.07}}$	6.76
			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
			Monkey	7B	$1.96_{0.04}$	$2.43_{0.03}$	-0.48	$14.02_{0.03}$	$14.11_{0.03}$	-0.09
			Idefics2	8B	$0.65_{0.01}$	$0.94_{0.02}$	-0.29	$9.93_{0.05}$	$12.57_{0.02}$	-2.64
			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-XComposer2.5-VL	7B	$0.93_{0.11}$	$1.11_{0.11}$	-0.18	$13.82_{0.16}$	$14.72_{0.18}$	-0.89
			InternVL-V1.5	25.5B	$1.99_{0.02}$	$6.49_{0.04}$	-4.5	$16.73_{0.06}$	$26.4_{0.03}$	-9.67
			InternVL-V2	25.5B	$6.18_{0.27}$	$6.38_{0.27}$	-0.20	$24.52_{0.29}$	$24.37_{0.30}$	0.16
			InternVL-V2	40B	$13.10_{0.37}$	$19.16_{0.44}$	-6.06	$33.64_{0.36}$	$41.35_{0.39}$	-7.71
			MiniCPM-V2.5	8B	$1.41_{0.03}$	$1.96_{0.02}$	-0.55	$11.94_{0.02}$	$13.37_{0.04}$	-1.43
			MiniCPM-V2.5-FT	8B	$13.86_{0.1}$	$13.73_{0.05}$	0.12	$36.89_{0.06}$	$36.51_{0.06}$	0.38
			Qwen-VL	7B	$2.0_{0.03}$	$2.32_{0.03}$	-0.32	$15.04_{0.05}$	$14.27_{0.05}$	0.77
			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	-	$\mathbf{14.87_{1.14}}$	$\mathbf{22.46_{1.35}}$	-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	$\mathbf{33.24_{0.04}}$	$\mathbf{30.7_{0.07}}$	2.54	$\mathbf{57.57_{0.06}}$	$\mathbf{53.66_{0.04}}$	3.91
			CogVLM2-Chinese-FT	19B	$\underline{61.69_{0.05}}$	$\underline{59.85_{0.08}}$	1.84	$\underline{78.14_{0.05}}$	$\underline{77.12_{0.04}}$	1.02
			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
			Monkey	7B	$0.62_{0.01}$	$1.44_{0.01}$	-0.82	$8.34_{0.06}$	$10.95_{0.03}$	-2.61
			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-XComposer2.5-VL	7B	$0.46_{0.07}$	$0.58_{0.08}$	-0.12	$12.97_{0.16}$	$14.99_{0.17}$	-2.01
			InternVL-V1.5	25.5B	$4.78_{0.02}$	$5.32_{0.02}$	-0.55	$26.43_{0.03}$	$21.7_{0.04}$	4.72
			InternVL-V2	25.5B	$9.02_{0.28}$	$7.92_{0.26}$	1.10	$32.50_{0.29}$	$26.90_{0.28}$	5.60
			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
			MiniCPM-V2.5-FT	8B	$7.44_{0.03}$	$7.92_{0.03}$	-0.49	$29.87_{0.04}$	$31.32_{0.03}$	-1.45
			Qwen-VL	7B	$0.04_{0.01}$	$0.0_{0.0}$	0.04	$1.5_{0.01}$	$0.34_{0.01}$	1.15
			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	-	$\mathbf{2.2_{0.47}}$	$\mathbf{1.8_{0.4}}$	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	$\mathbf{1.34_{0.03}}$	$\mathbf{2.67_{0.02}}$	-1.32	$\mathbf{17.35_{0.03}}$	$\mathbf{19.51_{0.03}}$	-2.16
			CogVLM2-Chinese-FT	19B	$\underline{42.11_{0.09}}$	$\underline{45.63_{0.06}}$	-3.51	$\underline{65.67_{0.15}}$	$\underline{69.28_{0.04}}$	-3.61
			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
			Monkey	7B	$0.12_{0.01}$	$0.07_{0.0}$	0.05	$6.36_{0.01}$	$6.68_{0.03}$	-0.32
			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-XComposer2.5-VL	7B	$0.11_{0.04}$	$0.12_{0.04}$	-0.01	$10.95_{0.11}$	$11.43_{0.12}$	-0.48
			InternVL-V1.5	25.5B	$0.03_{0.0}$	$0.1_{0.01}$	-0.07	$8.46_{0.01}$	$6.27_{0.04}$	2.19
			InternVL-V2	25.5B	$0.05_{0.02}$	$0.22_{0.05}$	-0.18	$9.49_{0.10}$	$9.90_{0.12}$	-0.41
			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
			MiniCPM-V2.5-FT	8B	$1.53_{0.01}$	$1.11_{0.02}$	0.42	$18.0_{0.03}$	$15.35_{0.02}$	2.65
			Qwen-VL	7B	$0.01_{0.0}$	$0.01_{0.0}$	0	$1.17_{0.01}$	$0.12_{0.0}$	1.06
			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

5 Related Work

Complex Reasoning in Vision Language Models. In the emerging field of complex reasoning in vision-language models, several significant contributions have been made to enhance multimodal reasoning capabilities. The Visual CoT dataset [42] is a noteworthy development, introducing a comprehensive dataset for chain-of-thought reasoning across visual contexts, aiming to improve interpretability and precision in multimodal large language models (MLLMs) by annotating key regions in images that inform VQA processes. Similarly, the Zhang et al. [61] extends the chain-of-thought framework to incorporate both visual and textual data, demonstrating improvements in reasoning and inference accuracy on complex multimodal datasets. Further, the benchmark MathVista [31] is put forward as a challenge of mathematical reasoning in visual contexts by evaluating large models on tasks that require both deep visual understanding and mathematical computation, marking a significant step towards models performing complex, real-world tasks.

Visual Question Answering (VQA) and Optical Character Recognition (OCR). Visual Question Answering (VQA) involves datasets designed for answering questions based on images, such as FVQA [51] and OK-VQA [32], which require external knowledge. CLEVR [21] focuses on visual reasoning, while Text-VQA [43, 6, 36, 53] targets understanding embedded text in images. Various datasets support the Text-VQA task, including TextVQA [43], ST-VQA [6], OCR-VQA [35], InfographicVQA [33], and DocVQA [34]. Optical Character Recognition (OCR) [37] has been widely studied, though classical methods struggle with unconstrained images. Advances in scene-text recognition [5, 14, 19, 20] have improved OCR in the wild, and OCR is integral to Text-VQA tasks. Models like LoRRA [43] and TAP [57] enhance VQA performance by integrating OCR to improve text recognition in images.

Vision Language Model. Vision-language models are designed for tasks that involve understanding and generating content from images and text [44, 28, 22, 23]. For example, models have been developed to combine Llama3 with advanced vision-language processing capabilities to handle complex multimodal tasks [59, 56, 17, 58, 52, 12]. Qwen-VL [4] enhances visual-linguistic representations for more accurate contextual interpretations, while OpenGVLab-InternVL-Chat [10, 9] merges the InternVL framework with interactive chat capabilities. These studies typically employ a multimodal encoder [41, 60, 55] 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 [39, 38], the Claude series models [2], the Gemini series models [45] and the Reka series models [46] have also been adapted for vision-language tasks, demonstrating strong performance in multimodal tasks. Finally, DocLLM [50] 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.

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. We invite the community to utilize our dataset and develop innovative strategies to boost the performance of vision-language models.

References

[1] 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.

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

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

[4] 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.

[5] 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.

[6] 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.

[7] 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, page 71–79, New York, NY, USA, 2018. Association for Computing Machinery.

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

[9] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024.

[10] 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.

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

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

[13] 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.

[14] 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.

[15] 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*, 2024.

[16] 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.

[17] 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.

[18] 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, 2024.

[19] 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, editors, *Computer Vision – ECCV 2014*, pages 497–511, Cham, 2014. Springer International Publishing.

[20] Max Jaderberg, Andrea Vedaldi, and Andrew Zisserman. Deep features for text spotting. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 512–528, Cham, 2014. Springer International Publishing.

[21] 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.

[22] 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.

[23] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models?, 2024.

[24] 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 2023.

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

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

[27] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024.

[28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc., 2023.

[29] Yuliang Liu, Biao Yang, Qiang Liu, Zhang Li, Zhiyin Ma, Shuo Zhang, and Xiang Bai. Textmonkey: An ocr-free large multimodal model for understanding document. *arXiv preprint arXiv: 2403.04473*, 2024.

[30] 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.

[31] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

[32] 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.

[33] 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)*, pages 1697–1706, January 2022.

[34] 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)*, pages 2200–2209, January 2021.

[35] 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)*, pages 947–952, 2019.

[36] 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)*, pages 947–952, 2019.

[37] G. Nagy. Twenty years of document image analysis in pami. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(01):38–62, jan 2000.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *arXiv preprint arXiv: 2403.16999*, 2024.

[43] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019.

[44] 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.

[45] 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*, 2024.

[46] 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*, 2024.

[47] G. Thinés, A. Costall, and G. Butterworth. *Michotte's Experimental Phenomenology of Perception*. Routledge Library Editions: Phenomenology. Taylor & Francis, 2013.

[48] 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.

[49] Rob van Lier and Walter Gerbino. Perceptual completions. In *The Oxford Handbook of Perceptual Organization*. Oxford University Press, 08 2015.

[50] 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*, 2023.

[51] 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.

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

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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)*, pages 8751–8761, June 2021.

[58] 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.

[59] 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.

[60] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. *IEEE International Conference on Computer Vision*, 2023.

[61] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, G. Karypis, and Alexander J. Smola. Multimodal chain-of-thought reasoning in language models. *Trans. Mach. Learn. Res.*, 2023.

[62] 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.

A Additional evaluation results on first 100 and 500 test cases

Table 3: Results of various open-source and closed-source vision language models on the VCR task using the first 100 test cases. Each test case includes one or more puzzles. FT means that the model is finetuned on 16,000 samples from the VCR-wiki train dataset. The best results among the finetuned models are underlined while the best results among the models without finetuning are highlighted in bold. Subscripts provide the standard deviation 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	-	$62.0_{0.76}$	$82.0_{0.63}$	-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	$\mathbf{86.39_{0.66}}$	$\mathbf{84.62_{0.92}}$	1.78	$\mathbf{91.39_{0.11}}$	$\mathbf{91.63_{0.11}}$	-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
			Monkey	7B	$46.75_{0.44}$	$48.52_{0.41}$	-1.78	$67.82_{0.22}$	$68.59_{0.13}$	-0.76
			Idefics2	8B	$14.79_{0.72}$	$26.63_{0.37}$	-11.83	$34.2_{0.37}$	$51.96_{0.1}$	-17.76
			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-XComposer2.5-VL	7B	$45.56_{3.83}$	$28.99_{3.50}$	16.57	$67.70_{2.79}$	$54.25_{2.70}$	13.45
			InternVL-V1.5	25.5B	$15.38_{0.29}$	$75.15_{0.7}$	-59.76	$52.21_{0.16}$	$85.87_{0.29}$	-33.66
			InternVL-V2	25.5B	$76.92_{3.15}$	$78.70_{3.22}$	-1.78	$88.29_{1.85}$	$89.40_{1.83}$	-1.11
			InternVL-V2	40B	$86.39_{2.56}$	$86.98_{2.60}$	-0.59	$93.51_{1.40}$	$94.35_{1.24}$	-0.84
			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
			Qwen-VL	7B	$47.34_{0.44}$	$46.75_{0.57}$	0.59	$69.02_{0.35}$	$69.19_{0.37}$	-0.17
			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	$\mathbf{44.97_{0.83}}$	$\mathbf{21.3_{0.47}}$	23.67	$\mathbf{65.39_{0.2}}$	$\mathbf{43.86_{0.27}}$	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
			Monkey	7B	$1.18_{0.22}$	$3.55_{0.18}$	-2.37	$12.66_{0.21}$	$15.97_{0.08}$	-3.31
			Idefics2	8B	$1.18_{0.2}$	$0.59_{0.1}$	0.59	$10.81_{0.08}$	$11.34_{0.12}$	-0.53
			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-XComposer2.5-VL	7B	$0.59_{0.58}$	$1.78_{1.01}$	-1.18	$14.09_{1.04}$	$16.57_{1.25}$	-2.48
			InternVL-V1.5	25.5B	$1.78_{0.21}$	$7.1_{0.22}$	-5.33	$16.28_{0.09}$	$26.6_{0.14}$	-10.32
			InternVL-V2	25.5B	$4.73_{1.62}$	$7.10_{2.03}$	-2.37	$24.16_{1.69}$	$26.34_{1.97}$	-2.19
			InternVL-V2	40B	$12.43_{2.54}$	$16.57_{2.89}$	-4.14	$33.74_{2.40}$	$39.51_{2.69}$	-5.76
			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
			Qwen-VL	7B	$1.78_{0.21}$	$2.96_{0.12}$	-1.18	$15.7_{0.14}$	$15.06_{0.19}$	0.63
			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	$\mathbf{34.57_{0.66}}$	$\mathbf{34.04_{1.01}}$	0.53	$\mathbf{58.78_{0.13}}$	$\mathbf{57.26_{0.12}}$	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
			Monkey	7B	$1.06_{0.12}$	$0.53_{0.06}$	0.53	$9.23_{0.08}$	$12.29_{0.13}$	-3.06
			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-XComposer2.5-VL	7B	$0.00_{0.00}$	$1.60_{0.91}$	-1.60	$11.94_{0.88}$	$16.12_{1.24}$	-4.18
			InternVL-V1.5	25.5B	$4.26_{0.28}$	$3.19_{0.38}$	1.06	$26.9_{0.23}$	$16.31_{0.14}$	10.59
			InternVL-V2	25.5B	$7.45_{1.91}$	$11.70_{2.26}$	-4.26	$34.61_{2.16}$	$31.38_{2.34}$	3.22
			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
			Qwen-VL	7B	$0.0_{0.0}$	$0.0_{0.0}$	0	$1.41_{0.02}$	$0.66_{0.03}$	0.76
			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	$\mathbf{3.19_{0.19}}$	$\mathbf{3.19_{0.32}}$	0	$\mathbf{18.33_{0.14}}$	$\mathbf{21.38_{0.09}}$	-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
			Monkey	7B	$0.0_{0.0}$	$0.0_{0.0}$	0	$6.15_{0.11}$	$6.62_{0.11}$	-0.47
			InternLM-XComposer2-VL	7B	$0.0_{0.0}$	$0.0_{0.0}$	0	$8.17_{0.03}$	$7.99_{0.03}$	0.18
			InternLM-XComposer2.5-VL	7B	$0.00_{0.00}$	$0.00_{0.00}$	0.00	$10.87_{0.82}$	$10.54_{0.84}$	0.32
			InternVL-V1.5	25.5B	$0.0_{0.0}$	$0.0_{0.0}$	0	$7.7_{0.08}$	$4.67_{0.04}$	3.03
			InternVL-V2	25.5B	$0.00_{0.00}$	$0.53_{0.52}$	-0.53	$9.85_{0.72}$	$11.97_{1.13}$	-2.11
			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
			Qwen-VL	7B	$0.0_{0.0}$	$0.0_{0.0}$	0	$1.1_{0.04}$	$0.06_{0.01}$	1.04
			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 4: Results of various open-source and closed-source vision language models on the VCR task using the first 500 test cases. Each test case includes one or more puzzles. FT means the model is finetuned on 16,000 samples from the VCR-wiki train dataset. The best results among the finetuned models are underlined while the best results among the models without finetuning are highlighted in bold. Subscripts provide the standard deviation 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	$\mathbf{83.11_{0.28}}$	$\mathbf{79.63_{0.33}}$	3.48	$\mathbf{89.43_{0.27}}$	$\mathbf{88.65_{0.26}}$	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
			Monkey	7B	$47.2_{0.2}$	$54.16_{0.41}$	-6.96	$65.7_{0.4}$	$71.17_{0.72}$	-5.47
			Idefics2	8B	$14.91_{0.14}$	$29.07_{0.2}$	-14.16	$31.63_{0.3}$	$51.5_{0.21}$	-19.87
			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-XComposer2.5-VL	7B	$42.48_{1.73}$	$25.84_{1.53}$	16.65	$63.03_{1.32}$	$50.75_{1.21}$	12.28
			InternVL-V1.5	25.5B	$15.78_{0.23}$	$74.91_{0.27}$	-59.13	$52.0_{0.31}$	$86.82_{0.47}$	-34.82
			InternVL-V2	25.5B	$76.15_{1.48}$	$79.13_{1.43}$	-2.98	$87.63_{0.89}$	$89.62_{0.80}$	-1.99
			InternVL-V2	40B	$84.84_{1.21}$	$87.08_{1.19}$	-2.24	$93.13_{0.69}$	$94.83_{0.50}$	-1.71
			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
			Qwen-VL	7B	$45.47_{0.35}$	$52.17_{0.33}$	-6.71	$66.81_{0.74}$	$71.73_{0.59}$	-4.93
			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	$\mathbf{41.74_{0.25}}$	$\mathbf{16.77_{0.22}}$	24.97	$\mathbf{62.56_{0.33}}$	$\mathbf{38.41_{0.44}}$	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
			Monkey	7B	$1.37_{0.05}$	$2.24_{0.15}$	-0.87	$13.16_{0.18}$	$14.45_{0.24}$	-1.29
			Idefics2	8B	$0.62_{0.02}$	$0.62_{0.06}$	0	$9.24_{0.11}$	$11.0_{0.16}$	-1.75
			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-XComposer2.5-VL	7B	$0.75_{0.31}$	$1.24_{0.39}$	-0.50	$13.67_{0.51}$	$14.92_{0.56}$	-1.25
			InternVL-V1.5	25.5B	$1.74_{0.13}$	$6.34_{0.13}$	-4.6	$16.85_{0.17}$	$26.11_{0.24}$	-9.26
			InternVL-V2	25.5B	$6.71_{0.87}$	$6.71_{0.89}$	0.00	$25.12_{0.94}$	$24.31_{0.96}$	0.80
			InternVL-V2	40B	$14.16_{1.22}$	$18.51_{1.36}$	-4.35	$35.01_{1.18}$	$41.02_{1.22}$	-6.02
			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
			Qwen-VL	7B	$1.61_{0.03}$	$1.74_{0.03}$	-0.12	$15.28_{0.13}$	$14.43_{0.54}$	0.85
			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	$\mathbf{33.63_{0.15}}$	$\mathbf{31.44_{0.19}}$	2.2	$\mathbf{57.97_{0.56}}$	$\mathbf{54.05_{0.54}}$	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
			Monkey	7B	$0.2_{0.01}$	$1.4_{0.05}$	-1.2	$7.89_{0.3}$	$10.26_{0.24}$	-2.37
			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-XComposer2.5-VL	7B	$0.30_{0.17}$	$0.40_{0.20}$	-0.10	$12.76_{0.42}$	$14.99_{0.43}$	-2.23
			InternVL-V1.5	25.5B	$3.99_{0.09}$	$4.69_{0.18}$	-0.7	$25.88_{0.45}$	$20.73_{0.53}$	5.15
			InternVL-V2	25.5B	$8.08_{0.86}$	$8.08_{0.92}$	0.00	$32.78_{0.89}$	$28.48_{0.91}$	4.31
			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
			Qwen-VL	7B	$0.0_{0.0}$	$0.0_{0.0}$	0	$1.25_{0.03}$	$0.43_{0.06}$	0.82
			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	$\mathbf{1.2_{0.07}}$	$\mathbf{2.3_{0.09}}$	-1.1	$\mathbf{16.83_{0.22}}$	$\mathbf{19.86_{0.23}}$	-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
			Monkey	7B	$0.0_{0.0}$	$0.0_{0.0}$	0	$5.69_{0.15}$	$6.3_{0.13}$	-0.61
			InternLM-XComposer2-VL	7B	$0.0_{0.0}$	$0.0_{0.0}$	0	$8.36_{0.09}$	$7.92_{0.09}$	0.44
			InternLM-XComposer2.5-VL	7B	$0.00_{0.00}$	$0.00_{0.00}$	0.00	$10.83_{0.31}$	$10.81_{0.31}$	0.02
			InternVL-V1.5	25.5B	$0.0_{0.0}$	$0.0_{0.0}$	0	$7.9_{0.12}$	$6.11_{0.26}$	1.79
			InternVL-V2	25.5B	$0.00_{0.00}$	$0.10_{0.09}$	-0.10	$9.59_{0.31}$	$10.15_{0.39}$	-0.57
			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
			Qwen-VL	7B	$0.0_{0.0}$	$0.0_{0.0}$	0	$1.1_{0.07}$	$0.15_{0.01}$	0.94
			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

We show the table of evaluation results on first 100 and 500 test cases for better comparison with human evaluation results and closed-source models correspondingly.

B Relationship between VCR-wiki-en and Other Benchmarks

We evaluate 38 VLMs across 23 benchmarks, treating the VLM scores as features of the benchmarks to calculate a correlation matrix. The heatmap of this matrix is presented in Figure 4. Based on the heatmap, we performed K-Means clustering and visualized the results in 2D in figure 5, using the first two principal components derived from the rows of the correlation matrix for each benchmark. We did not test VCR-wiki-zh for these processes due to the limited availability of VLMs that support Chinese.

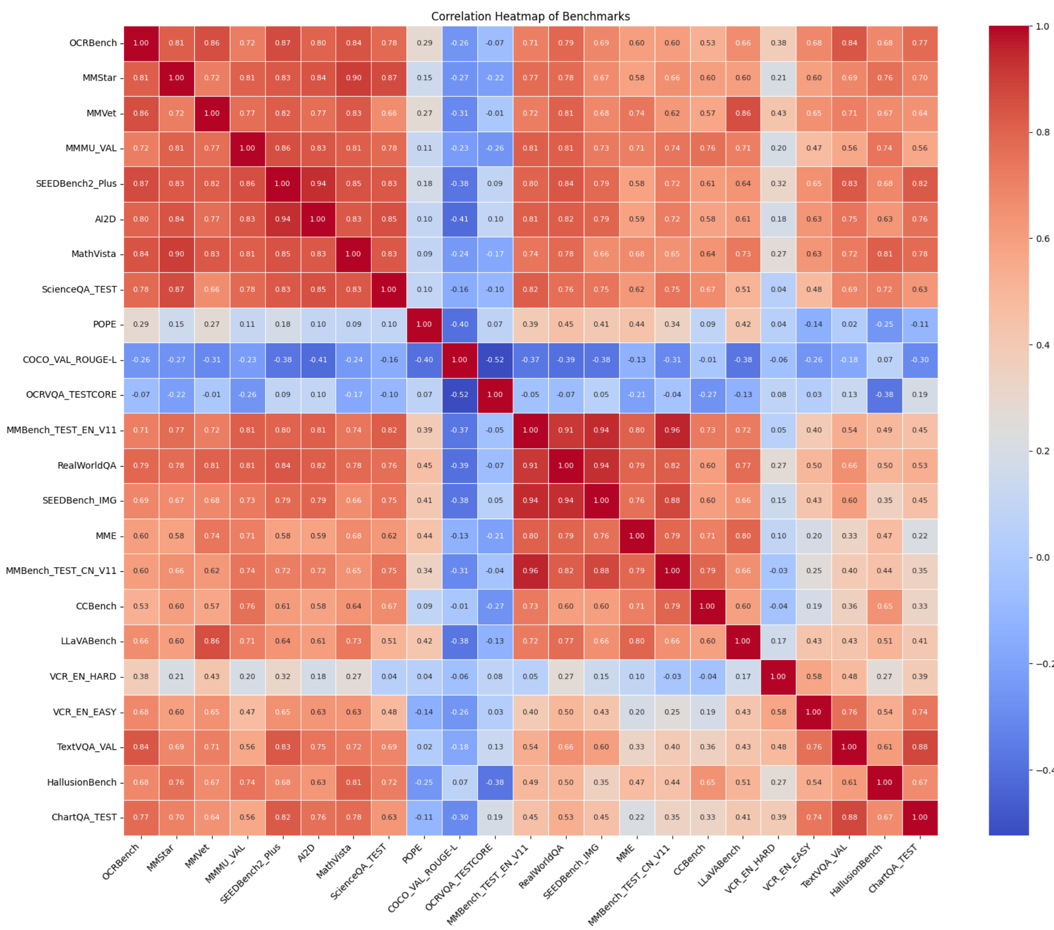

Figure 4: The heat map of benchmarks displays the correlation between the metric scores of 38 models for each benchmark pair.

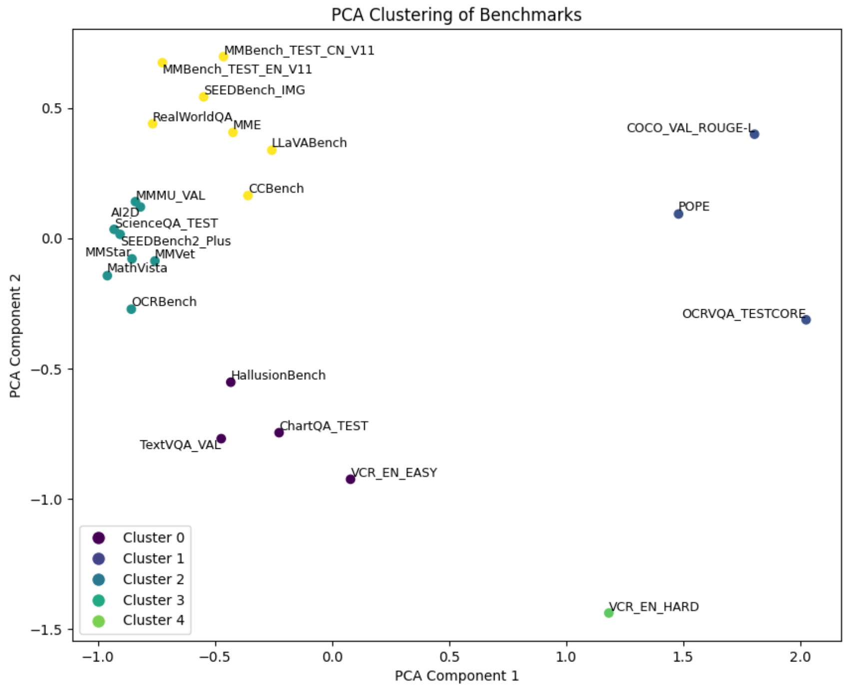

Figure 5: Each point in the figure represents the first 2 principal components of the metric score correlations between benchmarks.

C 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[5] 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.

Within TEI, we use spaCy to randomly select several 5-grams in the caption for masking. To ensure the restoration process is doable by a human without too much professional domain knowledge, the 5-grams do not contain numbers, person names, religious or political groups, facilities, organizations, locations, dates and time labeled by spaCy, and the total masked token does not exceed 50% of the tokens in the caption. We exclude all instances that do not have a single eligible 5-gram. 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 to adjust 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 vision-language models and a significant user base in both English and Chinese, we have chosen to develop the dataset in these two languages. For

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

each language, we meticulously select the covering proportion 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 of the language.

We release the dataset under the CC BY-SA 4.0 license. We do not include the link to the dataset due to anonymity.

C.1 Dataset Format and Statistics

Table 5: Basic statistics of the dataset. Note that the Easy and Hard configurations for each language 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 5.

D Information of models evaluated

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 [6]	34B	✓
CogVLM2 [7]	19B	✓
CogVLM2-Chinese [8]	19B	✓
DeepSeek-VL [9]	1.3B	✓
DeepSeek-VL [10]	7B	✓
Idefics2 [11]	8B	✓
InternLM-XComposer2-VL [12]	7B	✓
InternVL-V1.5 [13]	25.5B	✓
InternVL-V2 [14]	25.5B	✓
InternVL-V2 [15]	40B	✓
InternLM-XComposer2-VL [16]	7B	✓
MiniCPM-V2.5 [17]	8B	✓
Qwen-VL [18]	7B	✓
Yi-VL[19]	34B	✓
Yi-VL[20]	6B	✓
Monkey[21]	7B	✓
DocOwl-1.5-Omni[22]	8B	✓

E 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 VCR-wiki-en-easy, it does not perform well on VCR-wiki-en-hard. As noted in the paper, the easy partition of the benchmark primarily verifies that the VCR task is feasible for the models, while 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 VCR-wiki-zh and 300×375 for VCR-wiki-en), 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 pretraining to improve OCR performance, researchers could consider incorporating VCR tasks during pretraining. We have released 'vcr_transform.py' 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.

[6] https://huggingface.co/nyu-visionx/cambrian-34b
[7] https://huggingface.co/THUDM/CogVLM2-Llama3-chat-19B
[8] https://huggingface.co/THUDM/cogvlm2-llama3-Chinese-chat-19B
[9] https://huggingface.co/deepseek-ai/deepseek-vl-1.3b-chat
[10] https://huggingface.co/deepseek-ai/deepseek-vl-7b-chat
[11] https://huggingface.co/HuggingFaceM4/Idefics2-8B
[12] https://huggingface.co/internlm/internlm-xcomposer2-vl-7b
[13] https://huggingface.co/OpenGVLab/InternVL-Chat-V1-5
[14] https://huggingface.co/OpenGVLab/InternVL2-26B
[15] https://huggingface.co/OpenGVLab/InternVL2-40B
[16] https://huggingface.co/InternLM/InternLM-XComposer2-VL-7B
[17] https://huggingface.co/OpenBMB/MiniCPM-Llama3-V-2_5
[18] https://huggingface.co/Qwen/Qwen-VL-Chat
[19] https://huggingface.co/01-ai/Yi-VL-34B
[20] https://huggingface.co/01-ai/Yi-VL-6B
[21] https://huggingface.co/echo840/Monkey-Chat
[22] https://huggingface.co/mPLUG/DocOwl1.5-Omni

3. **Chain-of-Thought (CoT) Methods**: Researchers could explore multi-modality pipelines based on CoT techniques to improve existing VLMs on VCR tasks [8, 61]. 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.