# VCR: A Task for Pixel-Level Complex Reasoning in Vision Language Models via Restoring Occluded Text

Tianyu Zhang*, Suyuchen Wang*, Lu Li, Ge Zhang,
Perouz Taslakian, Sai Rajeswar, Jie Fu, Bang Liu, Yoshua Bengio

ICLR 2025

# Visual Caption Restoration (VCR) Overview

In this paper, we introduce:

1. A Task: *Visual Caption Restoration (VCR)*
2. A Dataset + Benchmark: *VCR-Wiki*
3. A Hidden Test: *VCR-Hidden*

*"What are the covered texts in the image?"*


The *VCR Task*

Easy Hard

English


Chinese


The *VCR-Wiki Dataset*

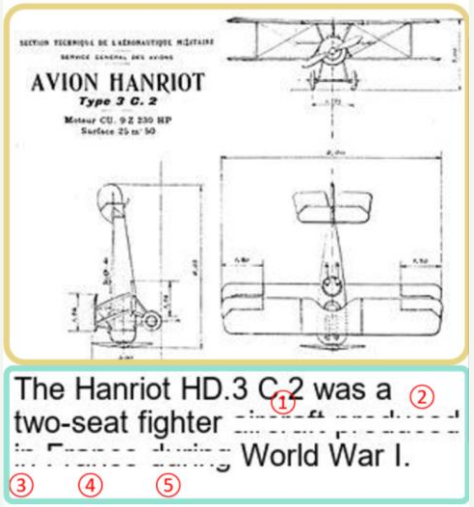
Latest data source, multi-lingual, multi-fonts, multi-resolution, multi-layout


The *VCR-Hidden Test*

Mila

# The Visual Caption Restoration (VCR) Task

1. A task: *Visual Caption Restoration (VCR)*
   - Task: given an image with two parts: a visual element and partly-occluded text-in-image, restore the covered text
   - A typical thinking process for this task involves using pixel-level hints for texts and multi-step reasoning




- Characteristics:
  - Unsolvable by OCR: the remaining pixels of the covered texts are crucial and hard to use
  - Have deterministic answer: the remaining pixels gives definitive answer; No LLM-as-a-judge needed
  - Flexibility: over covered area size, text to cover...

**1**. This image is very likely about a manual of an old fashion plane.
**2**. The visible pixel level hint of ① shows that it is a word ending with "ft". Given the previous word is "fighter", the first blank is likely "aircraft".
**3**. The title of the manual mentioned "HANRIOT" which is a French manufacturer.
**4**. ④ start from "F" thus, combined with 4, we induce it should be "France".
**5**. It is very likely that we need a "adj." between "France" and "World War I". Given ⑤ ends with "g" and start with "d". We can infer it is "during".
**6**. ② start from "p" and end with "d". It needs to be a verb. Connecting with **4**, it is "produced".
**7**. ③ is a word starting from "i" and only has 2 character. It evident that it is "in".

# The VCR-Wiki Dataset

2. A Dataset + Benchmark: *VCR-Wiki*
   - Sourced from Wikipedia dump
   - Multilingual: Contains En and Zh splits
   - Multi-difficulty: Contains easy and hard splits:
     - Easy: The text cover causes OCR to completely fail
     - Hard: Only 1-2 pixels left on the top and bottom
     - The easy / hard performance are directly comparable, with exact same splits and texts and only pixel-level difference on the cover
   - Contains train / val / test sets
   - SFT on the training set can improve real-world occluded text recognition


Mila

# The VCR-Hidden Hidden Test

3. A Hidden Test: *VCR-Hidden*
   - To avoid data contamination on VCR-Wiki's test set, we created a hidden test for VCR:
     - Up-to-date source: we use latest news to update the data source
     - More flexible data: the data contains multi-lingual, multi-fonts, multi-resolution, multi-layout VCR images for maximum flexibility


# Huge Gap Between SoTA VLM and Human

Table 4: 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 |

Native speakers can very easily achieve
~100% restoration accuracy on VCR-Wiki!

English Hard

| Language | Mode | Model name | Model size | Exact match (%) ↑ | | | Jaccard index (%) ↑ | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | *VI+TEI* | *TEI* | Δ | *VI+TEI* | *TEI* | Δ |
| | | *Closed-source models* | | | | | | | |
| English | Hard | 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-source models* | | | | | | | |
| | | CogVLM2 | 19B | $\mathbf{37.98_{0.18}}$ | $\mathbf{17.68_{0.06}}$ | 20.3 | $\mathbf{59.99_{0.05}}$ | $\mathbf{39.69_{0.03}}$ | 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 |
| | | 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 |
| | | 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 |
| | | MiniCPM-V2.5 | 8B | $1.41_{0.03}$ | $1.96_{0.02}$ | -0.55 | $11.94_{0.02}$ | $13.37_{0.04}$ | -1.43 |
| | | 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 |

Chinese Hard

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | *Closed-source models* | | | | | | | |
| Hard | 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-source models* | | | | | | | |
| | 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 |
| | 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 |
| | 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 |

... Yet a lot of SoTA VLMs achieve
nearly 0 restoration accuracy on VCR-Wiki

More interesting findings
(and VLM development guidelines) in the paper

Mila

# See You in Poster Session 2!

arXiv

GitHub

Hugging Face

OpenReview

**Test with EvolvingLMMs-Lab/lmms-eval**

EvolvingLMMs-Lab/lmms-eval: Accelerating the development of large multimodal models (LMMs) with lmms-eval (github.com)

**Test with open-compass/VLMEvalKit**

open-compass/VLMEvalKit: Open-source evaluation toolkit of large vision-language models (LVLMs), support GPT-4v, Gemini, QwenVLPlus, 50+ HF models, 20+ benchmarks (github.com)

Mila