FICO: Evaluating Vision-Language Models under Visual Fidelity and Compression at Scale

Jianhong Tu[1], Nicholas Crispino[1], Kyle Montgomery[1], Chenguang Wang[1]*, Dawn Song[2]*
[1]University of California, Santa Cruz [2]University of California, Berkeley
{jtu22, ncrispin, kylemontgomery, chenguangwang}@ucsc.edu, dawnsong@berkeley.edu

Abstract

Visual text compression is an emerging paradigm for rendering text as images for processing by vision-language models (VLMs), enabling higher information density per context token. However, the robustness of VLMs under dense, text-based visual inputs remains unevaluated. We introduce FICO, a benchmark designed to assess VLM robustness across seven controlled variants of visual fidelity and information density. FICO spans documents of 8k to 64k tokens and includes three tasks of increasing semantic granularity: optical character recognition (OCR), needle-in-a-haystack (NIAH) retrieval, and visual question answering (VQA). Evaluating 13 general-purpose VLMs and 3 OCR-specialized models reveals three consistent trends: performance drops sharply under increased density or reduced resolution; cross-task transfer between OCR, NIAH, and VQA is limited; and VQA is comparatively robust, suggesting that low-level details are lost before semantics. By exposing failure modes that remain invisible under conventional VLM evaluations, FICO establishes a rigorous testbed for visual text compression. Data and code are available at https://github.com/wang-research-lab/fico-bench.

1 Introduction

Despite the strong performance of large language models (LLMs) across a wide range of complex tasks, the quadratic $O(n^2)$ computational cost of the self-attention mechanism (Vaswani et al., 2017) remains a fundamental bottleneck for long-context processing (Zhao et al., 2025; Tay et al., 2023). To mitigate these issues, prior work has explored various self-attention alternatives (Yuan et al., 2025a; Katharopoulos et al., 2020) and context compression methods (Ge et al., 2024; Jiang et al., 2023). More recently, visual text compression has been proposed to directly reduce token counts by representing text in the pixel space, as illustrated in Figure 2. Studies, including DeepSeek-OCR (Wei et al., 2025) and Glyph (Cheng et al., 2025), demonstrate that strong performance can be preserved at compression rates up to 10x (i.e., representing the same text with images using roughly one-tenth the tokens), showing promise toward general-purpose, efficient language models that process text visually.

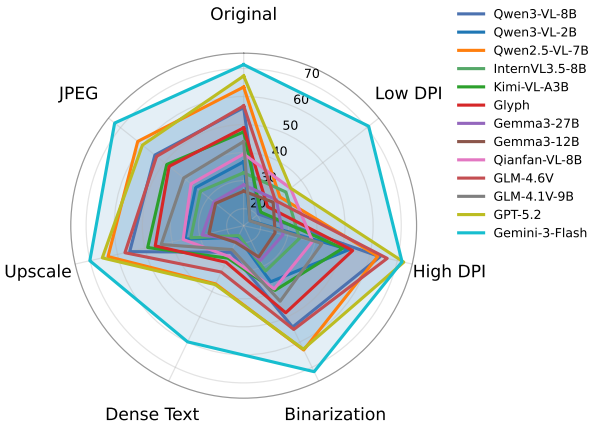


Figure 1: Accuracy of 13 VLMs across 7 rendering variations, averaged across OCR, NIAH, and VQA tasks.

While DeepSeek-OCR (Wei et al., 2025) and Glyph (Cheng et al., 2025) are explicitly trained for visual text compression, *any* vision-language model (VLM) is capable of visual text compression to various degrees of effectiveness. This raises a fundamental question: Are VLMs proficient and robust when reasoning over long, dense, text-centric image contexts? While recent benchmarks have evaluated VLMs' abilities to retrieve facts or answers to multi-hop questions from multimodal documents (Wang et al., 2025a; Ma et al., 2024; Zou et al., 2024; Mathew et al., 2021), rendering text as images introduces new challenges that can impact the reliability of VLMs (Cheng et al., 2025; Zhao et al., 2025). Deliberate choices between fi-

* Corresponding authors.

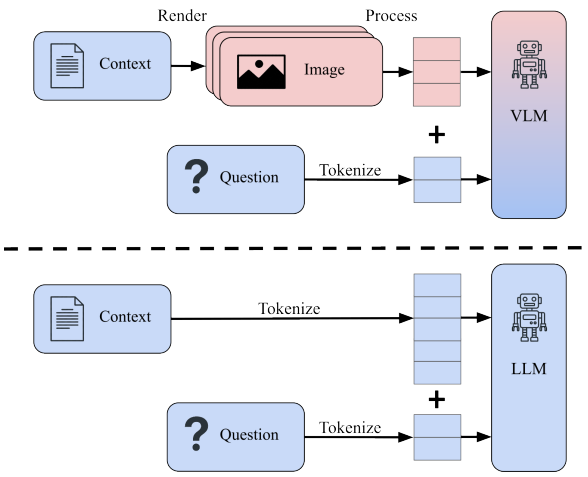


Figure 2: Comparison of visual text compression (top) and traditional text-based inference (bottom).

delity and efficiency, e.g., font size, color space, and dots per inch (DPI), are crucial in balancing compression and performance.

To systematically study these trade-offs, we introduce **Fi**delity-**co**mpression Bench (FICO), the first benchmark designed to evaluate the fidelity-scale trade-off in long-context vision-language models. FICO studies three tasks of increasing abstraction, namely optical character recognition (OCR), needle-in-a-haystack (NIAH), and visual question answering (VQA), under various levels of visual fidelity and compression. In doing so, we evaluate models' abilities to recognize low-level characters and reason over high-level semantics under two orthogonal perturbation axes: information density, which alters the effective compression ratio through font size and DPI changes, and fidelity, which introduces visual degradation via binarization, JPEG compression, and upscaling blur. Building on an existing long-context language benchmark, we construct a vision-language dataset with 7 rendering variants that systematically cover both dense and sparse text layouts, as well as high- and low-fidelity visual conditions.

Through extensive experiments with 16 VLMs (including 3 specialized OCR models), we obtain the following key findings: (1) Current VLMs exhibit limited robustness to long, dense image inputs, with consistently high error rates in OCR and NIAH, and pronounced performance degradation when visual context is used as a substitute for textual context in VQA. (2) Each task relies on distinct vision-language capabilities with limited inter-task generalization. For instance, Kimi-VL-A3B (Du et al., 2025) and Glyph (Cheng et al., 2025) achieve strong performance on NIAH and VQA tasks, yet exhibit high OCR error rates. (3) Across tasks, performance is most sensitive to information density. Dynamic native-resolution models, such as Qwen2.5-VL (Bai et al., 2025a), achieve strong OCR and NIAH results under high-resolution inputs, but are highly vulnerable to changes in resolution and text-density, particularly in the low-DPI and dense-text regimes. In contrast, tile-based and static-resolution models generally attain lower peak performance but degrade more gracefully across resolution and density shifts, demonstrating greater robustness. (4) For VQA, models remain largely robust to both density and fidelity perturbations, with only minor performance variations. This suggests that high-level semantic information can be efficiently compressed and preserved, whereas low-level literal recognition (e.g., characters and exact spans) requires substantially larger token budgets.

The key contributions of this paper are as follows:

- We introduce FICO, a benchmark for evaluating the fidelity-compression trade-off in long-context vision-language models across three complementary tasks of increasing abstraction: OCR, NIAH, and VQA.
- We construct a large-scale image-rendered dataset with seven rendering variants with differing levels of information density and visual fidelity, enabling controlled analysis of robustness under various compression ratios.
- We benchmark 13 general-purpose and 3 OCR-specialized VLMs, revealing limited robustness to dense contexts and weak cross-task transfer.
- Our study uncovers a task-dependent efficiency-fidelity trade-off: high-level semantics remain stable under compression, whereas low-level literal recognition is highly sensitive to text density.

2 Related Works

Long-context modeling. The quadratic complexity of self-attention is a key bottleneck for scaling language models to long contexts (Vaswani et al., 2017; Tay et al., 2023), which is necessary in long-horizon or deep-reasoning tasks such as agentic coding or mathematical problem solving (Jimenez

et al., 2024; Rando et al., 2025; Chen et al., 2025; Shao et al., 2025). Autoregressive inference becomes prohibitive at long contexts due to the growing key-value cache. Many techniques have been proposed to improve the efficiency of language models on long contexts, including architectural solutions (Yuan et al., 2025a; Katharopoulos et al., 2020; Yang et al., 2025; Gu and Dao, 2023; Poli et al., 2023; Hutchins et al., 2022), context compression methods (Ge et al., 2024; Jiang et al., 2023), and selective context management techniques (Yu et al., 2025; Yuan et al., 2025b; Sun et al., 2025). VIST (Xing et al., 2025b) is an exploration of compressing text contexts using visual tokens with a learned perceiver-resampler module that encodes high-signal context information. DeepSeek-OCR (Wei et al., 2025) systematically studies the trade-off between OCR performance and compression rate, showing that a 96% accuracy is achievable using 10% of equivalent tokens. Concurrently, Glyph (Cheng et al., 2025) and SeeTok (Xing et al., 2025a) build an instruction-following model with up to 4x token savings. A separate line compresses the visual tokens that already encode image content via merging (Shang et al., 2024), pruning (Chen et al., 2024), or in-context aggregation (Gao et al., 2025), targeting image redundancy rather than the textual density we study.

Multimodal long-context & robustness benchmarks. Needle-in-a-haystack (NIAH) (Kamradt, 2023; Hsieh et al., 2024) and long-context question answering (Bai et al., 2024, 2025b) evaluate the effectiveness of language models on long texts. In the multimodal setting, MM-NIAH (Wang et al., 2025a) evaluates a model's ability to retrieve, count, and reason with long multimodal documents, while Wang et al. (2024) requires a model to extract the relevant image needle. Moreover, MMLongBench-Doc (Ma et al., 2024) is a question answering (QA) benchmark with long PDF documents featuring interleaved image and text paragraphs, and MMLongBench (Bai et al., 2025b) expands the scope by designing a diverse problem set, including summarization, visual retrieval augmented generation, and in-context learning tasks. MLLM-IC (Qiu et al., 2025) and MMC-Bench (Zhang et al., 2024) are image benchmarks that test VLMs under common corruptions, including color space shift, occlusion, and patch artifacts. However, these benchmarks are not tailored to test long-context abilities, and the perturbation types are unlikely in the text rendering pipeline.

3 FICO

Visual text compression offers a new perspective on long-context modeling, particularly in scenarios where large external documents are provided as auxiliary context. Formally, given a long text-only document $T = [t_1, \ldots, t_{N_T}]$, where N_T denotes the total number of text tokens, we apply a text rendering pipeline R, implemented using the ReportLab library (ReportLab Inc., 2024), to render the text into images: $I = R(T) = [I_1, \ldots, I_M]$, where M denotes the total number of rendered images. The vision encoder ϕ_V of a VLM transforms and encodes each image into a set of visual tokens: $N_V = \sum_{I_i} |\phi_V(I_i)|$, where N_V denotes the sum of all vision tokens for document T. We thereby define the compression ratio $R_c = \frac{N_T}{N_V}$, which measures the effective reduction in context length. Notably, the number of visual tokens ultimately consumed by a model can vary substantially across architectures due to differences in image preprocessing pipelines and vision tokenizers (Du et al., 2025; Wei et al., 2025; Team, 2025b,a).

Parameter	Value
Font style	NotoSans
Font size	11
Line height	12
DPI	96
Margin	10
Aspect ratio	A4

Table 1: Default rendering configuration.

Task	Documents	Instances	Images
OCR	76	5,312	5,312
NIAH	185	3,885	19,672
VQA	185	185	19,672

Table 2: Dataset statistics for FICO. Documents denotes the number of source documents, instances denotes the number of evaluated samples, and images denotes the total number of rendered images consumed by the models.

To evaluate general performance, we adopt a reasonable default rendering configuration (detailed in Table 1) following prior studies (Zhao et al., 2025; Cheng et al., 2025), balancing the compression ratio and visual fidelity. We then construct a set of perturbed variants by modifying the rendering configuration or applying image post-processing operations. Overall, we curate a seed

text-only dataset derived from LongBenchv2 (Bai et al., 2025b), containing documents of 8k-64k tokens, with extreme-length cases removed to accommodate a broader range of VLMs. Under the default rendering configuration, each document is rendered into 5-42 images, with only 3-20 images in the densest setting.

3.1 Visual Text Compression Tasks

To provide a holistic analysis of VLMs, we deliberately design 3 tasks spanning different levels of abstraction, ranging from low-level character recognition to high-level semantic reasoning.

OCR The OCR task represents the most fine-grained evaluation, challenging models to faithfully reconstruct the source text from rendered images. This task tests low-level visual recognition, requiring models to attend to fine-grained details while avoiding hallucination or paraphrasing.

Following OmniDocBench (Ouyang et al., 2025) and DocVQA (Mathew et al., 2021), we adopt the Normalized Edit Distance (Levenshtein, 1965) and report the Character Error Rates (CER). These metrics are widely used in OCR evaluation because they directly measure character-level fidelity and remain robust to length variations. To isolate the impact of visual text compression and image-level perturbations, we restrict the OCR task to single-image inputs. Despite the absence of multi-image context, each image contains, on average, 1,600 and 3,600 tokens for the original and dense setting, respectively, ensuring that the task remains challenging.

NIAH The NIAH (Kamradt, 2023; Hsieh et al., 2024) task is a standard evaluation for understanding models' effective context length, measuring whether a model can retrieve a small but precise piece of information embedded sparsely within a long context dominated by irrelevant content. Concretely, a model is required to report a target value associated with a queried key, which is surrounded by substantial distractor text.

Following previous works (Hsieh et al., 2024; Zhao et al., 2025), we formulate a multi-key NIAH retrieval task by programmatically injecting text needles into the original document before rendering. To control needle depth, we insert five needles per document, each placed uniformly at random within one of five non-overlapping context intervals (0-20%, 20-40%, 40-60%, 60-80%, and 80-100%). Each needle follows a fixed key-value template, "`The secret {key} is {value}.`" We ensure that all keys and values are unique and do not appear elsewhere in the context. This setting explicitly evaluates a model's ability to isolate the correct information in the presence of multiple distractors, requiring both fine-grained recognition and semantic grounding. We evaluate performance using Exact Match (EM), in which a prediction is only assigned credit if the target value is correctly identified in the parsed model output.

VQA This task challenges a model to perform multi-hop reasoning over long documents, requiring high-level semantic understanding and complex reasoning to deduce the correct answer. We reuse the questions from LongBenchv2 (Bai et al., 2025b) but render the input context as images. Each question takes the form of a multiple-choice question, spanning 5 subtasks: single-document QA, multi-document QA, codebase and structured data understanding, multi-turn history understanding, and long in-context learning. We measure accuracy, computed by parsing the predicted answer and comparing it against the ground-truth answer.

3.2 Perturbation Variants

Previous work (Cheng et al., 2025; Zhao et al., 2025) has shown that VLMs can be sensitive to rendering and preprocessing choices. To comprehensively evaluate model capability and robustness under common and practical rendering conditions, we construct seven splits in total for each of the three tasks introduced above: one original (unperturbed) split and six perturbed splits. The configuration of the original split is detailed in Table 1, while representative examples of the six perturbation splits are shown in Figure 3.

1. Binarization: We convert grayscale images to 1-bit black-and-white using a global threshold of 128 (8-bit scale). This enforces clear foreground-background separation but perturbs fine character strokes and contours.

2. JPEG: We apply JPEG encoding with a quality factor of 10, inducing severe quantization of Discrete Cosine Transform (DCT) coefficients, resulting in characteristic blocking artifacts and ringing around high-contrast edges.

3. Dense: We decrease the font size from 11 to 7 and the line height from 12 to 8 to increase the character density, forcibly increasing the compression ratio for all models.

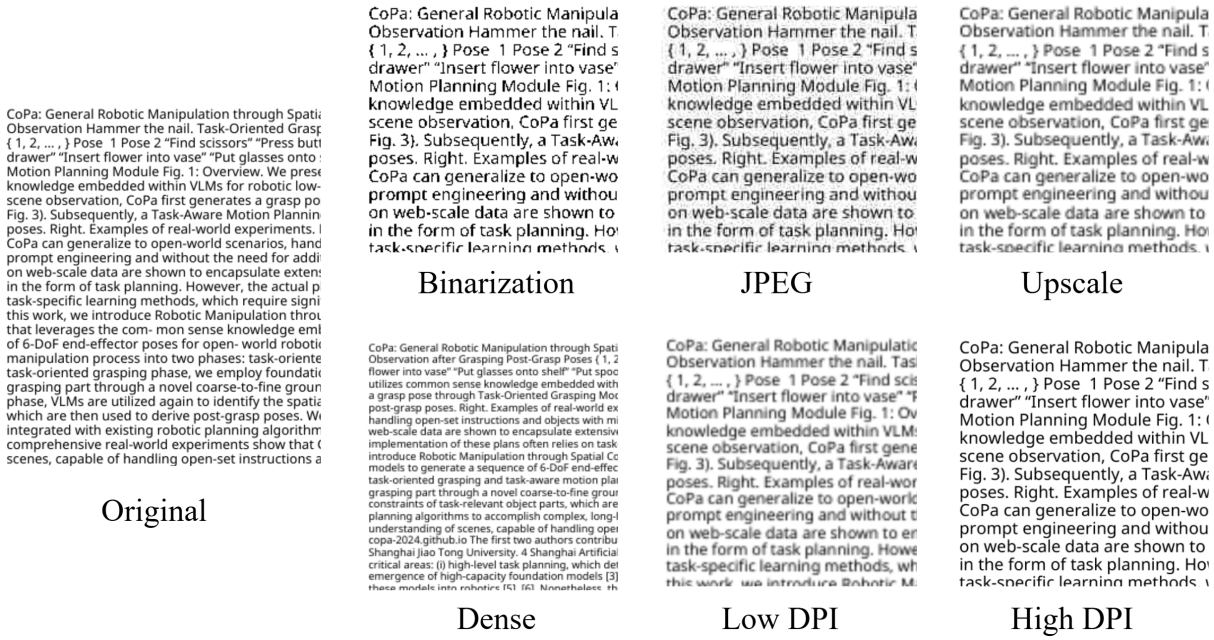


Figure 3: Examples of the default rendering and 6 perturbation variants.

4. Low DPI: We decrease the DPI from 96 to 56 to downscale the image resolution, increasing character density for native-resolution models while having minimal effect on models that rescale inputs to a fixed size.

5. High DPI: We double the DPI from 96 to 192 to increase the resolution, yielding more visual tokens for dynamic-resolution models and a lower effective compression ratio.

6. Upscale: We bilinearly upscale images rendered at 56 DPI to the 96 DPI image size. Although the spatial dimensions match the default setting, high-frequency details are lost, resulting in blur.

The Binarization, JPEG, and Upscale perturbations don't impact the compression ratio relative to the default rendering configuration. Per-model estimated compression ratios for the remaining variants are reported in Appendix E, where we also analyze their compute implications.

4 Experiments

4.1 Setup

We evaluate 11 open-source general-purpose VLMs, including Qwen3-VL-8B and Qwen3-VL-2B (Team, 2025b), Qwen2.5-VL-7B (Bai et al., 2025a), InternVL3.5-8B (Wang et al., 2025b), Kimi-VL-A3B (Du et al., 2025), Glyph (Cheng et al., 2025), Gemma3-27B and Gemma3-12B (Team, 2025a), Qianfan-VL-8B (Dong et al., 2025), GLM-4.6V and GLM-4.1V-9B (Hong et al., 2025), 3 OCR-specialized models, namely PaddleOCR-VL (Cui et al., 2025), MinerU2.5 (Niu et al., 2025), and DeepSeek-OCR (Wei et al., 2025), as well as 2 proprietary frontier models, GPT-5.2 (high image detail, no reasoning effort, low verbosity) (OpenAI, 2025) and Gemini-3-Flash (medium media resolution, minimal thinking level) (Google DeepMind, 2025), covering a diverse range of model architectures and dynamic resolution strategies. If applicable, we select their "instruct" variants or disable thinking through prompting. GLM-4.6V, Gemini-3-Flash, and GPT-5.2 are accessed through APIs. All other models are served on NVIDIA RTX 6000 Pro Blackwell GPUs using the vLLM backend (Kwon et al., 2023). Additional hyperparameters are provided in Appendix D.

4.2 Main Results

We report the OCR, NIAH, and VQA scores of each model under the original and the 6 perturbed rendering strategies in Table 3.

OCR The OCR results show that optical character recognition remains challenging across most settings: even OCR-specialized models achieve character error rates (CER) no lower than 10% using the default rendering configuration. Among open-source models, Qwen2.5-VL, Qwen3-VL, and GLM-4.6V perform competitively and, in the HighDPI setting, can even surpass specialized

Task	Model	Orig.	Perturbations					
			LowDPI	HighDPI	Bin.	Dense	Upscale	JPEG
OCR (Err ↓)	GPT-5.2	6.54	36.37	4.50	7.55	37.94	7.14	9.83
	Gemini-3-Flash	**2.26**	**2.08**	**2.49**	**2.18**	**9.24**	**1.70**	**1.92**
	Qwen3-VL-8B	24.02	75.11	13.18	26.44	69.76	22.08	24.97
	Qwen3-VL-2B	33.88	78.46	25.06	44.03	76.98	34.77	37.49
	Qwen2.5-VL-7B	11.07	60.29	8.26	13.61	51.14	12.46	14.29
	InternVL3.5-8B	36.15	36.97	34.46	36.87	76.44	36.59	36.64
	Kimi-VL-A3B	47.41	75.89	45.69	47.58	76.39	45.72	46.01
	Glyph	47.46	72.47	35.70	47.42	73.53	47.91	47.23
	Gemma3-27B	53.28	53.41	53.70	53.50	72.66	53.41	53.46
	Gemma3-12B	57.53	58.14	58.00	58.23	72.73	58.57	58.11
	Qianfan-VL-8B	32.05	42.32	34.27	34.16	69.65	39.18	33.01
	GLM-4.6V	20.09	71.55	7.03	22.17	67.82	20.78	21.05
	GLM-4.1V-9B	50.22	82.56	29.22	51.30	78.52	49.92	58.01
	PaddleOCR-VL†	17.33	73.38	12.80	23.78	72.48	21.66	40.09
	MinerU2.5†	17.72	63.58	14.76	18.83	72.79	18.82	22.39
	DeepSeek-OCR†	21.21	22.31	20.92	21.14	41.01	24.09	30.99
NIAH (Acc ↑)	GPT-5.2	84.50	18.01	**96.04**	74.77	21.62	83.42	68.29
	Gemini-3-Flash	**93.15**	**94.95**	92.43	**93.15**	**59.81**	**95.85**	**93.30**
	Qwen3-VL-8B	74.59	9.91	88.29	71.53	15.68	69.37	67.21
	Qwen3-VL-2B	23.96	1.08	63.42	32.07	4.14	14.41	27.03
	Qwen2.5-VL-7B	85.59	25.41	82.88	83.60	38.20	83.42	84.14
	InternVL3.5-8B	27.21	24.86	26.13	25.41	11.71	23.60	25.59
	Kimi-VL-A3B	68.65	19.10	75.50	42.70	31.71	70.63	70.81
	Glyph	81.08	27.03	87.75	81.26	36.94	78.20	80.90
	Gemma3-27B	19.64	19.64	20.54	18.20	8.29	20.00	18.56
	Gemma3-12B	18.02	19.64	16.76	18.92	8.29	18.56	19.10
	Qianfan-VL-8B	42.70	40.54	43.96	41.44	27.57	39.82	39.28
	GLM-4.6V	77.66	30.45	91.89	77.48	40.90	77.66	70.99
	GLM-4.1V-9B	71.89	21.62	55.32	74.23	31.17	72.43	74.95
VQA (Acc ↑)	GPT-5.2	46.48	36.21	**49.72**	42.16	40.54	**42.16**	41.62
	Gemini-3-Flash	**47.03**	**43.24**	49.19	**48.11**	**49.18**	39.46	**50.81**
	Qwen3-VL-8B	36.22	34.05	37.30	35.14	35.68	37.84	38.91
	Qwen3-VL-2B	32.43	33.51	36.22	32.43	34.59	38.38	29.73
	Qwen2.5-VL-7B	36.76	36.22	35.14	40.54	36.22	40.54	37.30
	InternVL3.5-8B	18.38	22.70	22.70	17.84	24.32	23.78	23.24
	Kimi-VL-A3B	36.76	29.19	37.30	36.76	34.05	38.38	38.38
	Glyph	29.19	28.11	27.03	27.57	31.35	23.78	25.95
	Gemma3-27B	28.65	29.73	27.57	27.57	29.19	30.27	26.49
	Gemma3-12B	27.03	29.73	25.95	27.03	32.97	28.65	28.11
	Qianfan-VL-8B	21.08	19.46	20.54	22.70	29.19	19.46	20.00
	GLM-4.6V	31.35	35.68	36.21	28.11	34.59	34.05	28.65
	GLM-4.1V-9B	24.86	16.76	15.68	23.24	25.41	24.32	20.54

Table 3: Main results for OCR (Err ↓), NIAH (Acc ↑), and VQA (Acc ↑) across image perturbations. **Bold**: Best score for each task and perturbation. †: Dedicated OCR-specialized models.

OCR systems. Both proprietary models demonstrate strong performance overall. Gemini-3-Flash outperforms the best general-purpose and OCR-specialized models across all settings, while GPT-5.2 achieves competitive results across most splits.

Despite their strong peak performance, the open native-resolution models Qwen2.5-VL, Qwen3-VL, and GLM-4.6V are highly sensitive to changes in compression ratio, and GPT-5.2 exhibits the same sensitivity. Their OCR error rate grows significantly under LowDPI and Dense conditions and only partially recovers in the Upscale setting, indicating that scale alone does not resolve this failure mode. In contrast, models that operate on fixed input resolutions (e.g., the Gemma3 family) or employ dynamic tiling (e.g., Qianfan-

VL-8B) exhibit greater robustness, as intermediate resizing mitigates extreme density shifts. Among OCR-specialized models, PaddleOCR-VL and MinerU2.5 are also sensitive to increased text density and further degrade under JPEG compression. DeepSeek-OCR is the most consistent open model across perturbations, reflecting the benefits of its compression-targeted training, while Gemini-3-Flash is the only evaluated system that combines frontier peak accuracy with uniform robustness across density and artifact splits.

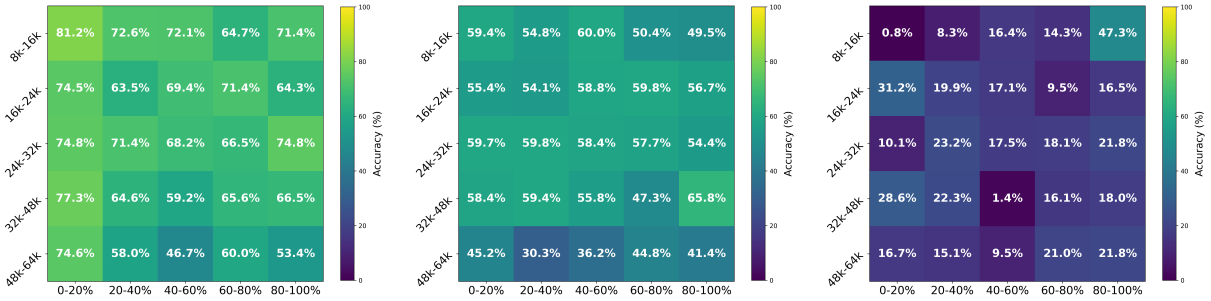
NIAH The retrieval (NIAH) results exhibit patterns broadly consistent with those observed in OCR. Gemini-3-Flash attains the highest accuracy on every split except HighDPI, where GPT-5.2 edges it out at 96.04%. Among open-source models, native-resolution architectures again set the ceiling, with Qwen2.5-VL-7B reaching ∼85%. However, these peak gains come at the cost of robustness: for both open native-resolution models and GPT-5.2, accuracy degrades sharply as text density increases, while Gemini-3-Flash maintains strong LowDPI performance. Notably, despite high error rates in OCR, Glyph, Kimi-VL-A3B, and GLM-4.1V-9B perform competitively on NIAH and, in some cases, can even outperform the Qwen models. This discrepancy further supports the conclusion that OCR fidelity and long-context retrieval rely on distinct capabilities. Moreover, when comparing Glyph and its counterpart GLM-4.1V-9B, we find a significant performance boost likely due to Glyph's focused training.

Across models, tolerance to visual artifacts such as binarization, upscaling blur, and JPEG compression varies considerably. Most models experience modest performance drops under these perturbations. For example, relative to the default configurations, InternVL3.5-8B degrades by 3.5 points in the upscale setting, and Qwen3-VL-8B drops by 7 points under JPEG compression. In contrast, Kimi-VL-A3B shows limited robustness, suffering a severe 26-point drop under binarization. Surprisingly, while seeing occasionally large performance drops, GLM-4.1V-9B also slightly benefits from certain perturbations, exhibiting gains of up to 3 points. Overall, these results reinforce that the effective compression ratio is the primary determinant of retrieval performance, while also revealing substantial variation in models' resilience to fidelity perturbations.

To examine how retrieval performance varies with needle depth and context length, we present a detailed analysis of Glyph, Kimi-VL-A3B, and Gemma3-27B in Figure 4, and report scores for the other models in Appendix H. For Glyph and Kimi-VL-A3B, we observe a consistent trend: retrieval accuracy is highest for shorter contexts and shallow needle placements, and degrades progressively as both context length and needle depth increase. In Glyph, the gap between the best- and worst-performing regions approaches 30 percentage points, indicating that even models with specific training for visual text compression exhibit uneven information retention across context positions. In contrast, Gemma3-27B demonstrates poor long-context proficiency and shows no depth-dependent structure. Instead, performance appears largely uncorrelated with either needle depth or context length, with accuracies remaining below 20% across most regions and isolated outliers ranging from 47% to as low as 0.8%. While VLMs demonstrate promising potential for long-context retrieval, their ability to attend to fine-grained information is highly non-uniform across context positions.

VQA Unlike OCR and NIAH, VQA cannot be solved by literal string matching and instead requires multi-hop reasoning over the context. Across both proprietary and open-source models, we observe markedly stronger robustness on VQA than on the other two tasks. The two proprietary models outperform the best open-source baselines on every split, and, echoing the OCR and NIAH trends, GPT-5.2 is weaker under LowDPI and Dense settings while Gemini-3-Flash is more uniform. Among open models, native-resolution systems such as Qwen2.5-VL and Qwen3-VL that collapsed under text-density shifts in OCR and NIAH retain their VQA accuracy, and Qwen3-VL-8B still reaches 38% under JPEG compression. Taken together, these results suggest that for models capable of leveraging native-resolution inputs, input modality is not the bottleneck for context understanding: high-level semantic information is preserved under visual text compression even when fine-grained character details are lost.

To directly compare the VQA performance between visual text compression (using the default rendering configuration) and the traditional text-based baseline, we provide each model the context as text instead of as rendered images. Figure 5 plots the performance for each model. While most



(a) Glyph (b) Kimi-VL-A3B (c) Gemma3-27B

Figure 4: NIAH heatmaps across different OCR-capable VLMs.

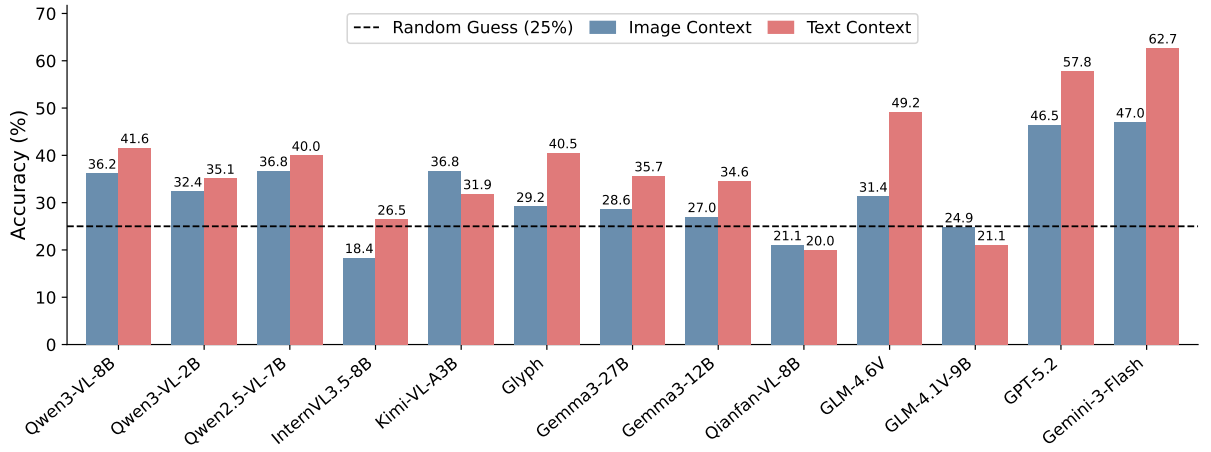


Figure 5: Accuracies of VLMs on VQA with image context using FICO default rendering (blue) and direct LongBenchV2 text context without any images (pink).

models perform better with the text-based context, both Glyph and GLM-4.1V-9B exhibit significant performance degradation relative to their text-only baselines. While the former can likely be attributed to its specialized visual text compression training, GLM-4.1V-9B's weaker performance likely stems from limited text-only capabilities, which can constrain its reasoning ability regardless of input modality. This limitation has been noted by the community and acknowledged by the GLM team in the GLM-4.6V release.[1]

To analyze performance as a function of document length, we aggregate results across all rendering variants into five context-length intervals and provide results in Table 4. Across most open-source models, performance peaks at intermediate context lengths (24–32k) and degrades noticeably as the context extends to 48k and 64k tokens. This pattern suggests these models fail to utilize very long contexts, likely due to the increased presence of noisy information. In contrast, GPT-5.2 and Gemini-3-Flash sustain their accuracy through 48–64k, indicating that long-context degradation on VQA is not intrinsic to the visual-text-compression setting but reflects limits of current open models. Models with shorter maximum context windows exhibit more pronounced failures. In particular, InternVL3.5-8B and Qianfan-VL-8B show substantially reduced accuracy in the 48–64k interval due to their 32k context limitation.

Model	8-16k	16-24k	24-32k	32-48k	48-64k
GPT-5.2	37.04	41.82	**52.94**	51.22	**50.00**
Gemini-3-Flash	37.04	**47.27**	44.12	**58.54**	42.86
Qwen3-VL-8B	40.21	39.48	44.53	31.70	23.97
Qwen3-VL-2B	29.62	34.80	36.55	37.63	27.55
Qwen2.5-VL-7B	**40.74**	36.10	42.43	42.85	23.46
InternVL3.5-8B	16.40	26.75	33.61	19.86	6.12
Kimi-VL-A3B	37.56	35.84	42.43	31.35	32.65
Glyph	29.10	30.64	26.89	26.82	21.93
Gemma3-27B	19.57	34.02	30.67	28.91	22.95
Gemma3-12B	36.50	29.61	26.05	25.08	26.53
Qianfan-VL-8B	14.28	36.88	37.39	5.92	3.57
GLM-4.6V	21.69	36.10	39.91	40.76	15.81
GLM-4.1V-9B	5.29	27.01	18.06	14.98	12.24

Table 4: VQA Accuracy (%) by context-length interval.

[1] https://huggingface.co/zai-org/GLM-4.6V

4.3 Reasoning Variants

To assess whether reasoning capability affects robustness under visual text compression, we evaluate three additional variants: Qwen3-VL-8B-Thinking and Kimi-VL-A3B-Thinking, which are reinforcement-learning post-trained for extended reasoning, and GLM-4.6V, which exposes reasoning via a prompt-level toggle. Table 5 shows that OCR worsens for all three models, as reflected by consistently higher OCR error. NIAH also declines for the two RL-trained variants, though GLM-4.6V is an exception with a small gain. VQA is mixed: GLM-4.6V improves substantially, Qwen3-VL-8B improves slightly, and Kimi-VL-A3B declines. In contrast, Text VQA improves consistently across all three variants. Overall, these results suggest that reasoning can help higher-level question answering, but often at the cost of low-level visual recognition under compression.

Model	OCR Err ↓	NIAH Acc ↑	VQA Acc ↑	Text VQA ↑
Qwen3-VL-8B	46.65 (+10.14)	51.32 (−5.33)	38.06 (+1.61)	52.43 (+10.81)
Kimi-VL-A3B	59.96 (+5.00)	47.23 (−6.92)	33.90 (−1.93)	37.84 (+5.95)
GLM-4.6V	34.17 (+1.24)	68.72 (+2.00)	44.00 (+11.34)	57.30 (+8.11)

Table 5: Averaged results for reasoning variants, with deltas from the non-thinking baselines. Text VQA denotes scores with text-only inputs.

4.4 Error Analysis

To provide a qualitative analysis, we summarize common patterns of model failures for each task and use Gemini-3-Flash as an automated classifier to assign each error to a predefined category. Full error definitions and distributions are provided in Appendix A. In OCR, Gemma3-12B is primarily hallucination-dominant, whereas Qwen3-VL-8B-Instruct and Qianfan-VL-8B are prone to degenerate repetition loops or heavily distorted outputs. In contrast, NIAH and VQA failures across all models are dominated by grounding errors. For NIAH, the most prominent failure mode is Incorrect Value, where models such as Qwen3-VL-8B-Thinking and Kimi-VL-A3B-Thinking often misspell the target string or produce a plausible substitute rather than recovering the exact answer. For VQA, the most frequent error mode is Incorrect Recall, indicating that failures arise primarily from the retrieval of the relevant textual evidence. Overall, the results indicate that OCR failures are often driven by model-specific generation pathologies, whereas NIAH and VQA failures more often reflect breakdowns in evidence retrieval and use.

5 Efficiency Analysis

We present an efficiency analysis by estimating the compression ratio of selected open-source models and the prefill cost of Qwen3-VL-8B as a function of input context length. Table 15 shows that compression is highly model- and rendering-dependent: dynamic-resolution models achieve substantial compression under Original, Dense, and LowDPI rendering, but often lose this advantage under HighDPI, whereas fixed-resolution models exhibit more stable but less adaptive behavior across splits. Consistent with this pattern, Table 6 verifies the trend: Original, Dense, and LowDPI consistently incur lower prefill cost than the text-only baseline, while HighDPI is more expensive. Moreover, the cost advantage of these compressed visual inputs becomes larger as context length increases. Table 16 demonstrates that the fraction of total computation attributable to the vision tower steadily decreases with longer inputs, indicating that the benefit of shortening the LM sequence increasingly dominates in the long-context regime. Overall, visual text compression is compute-favorable under reasonable rendering settings.

Ctx.	Original	Dense	LowDPI	HighDPI	Text
8K	88.29	35.82	27.29	495.14	130.01
16K	187.72	73.61	55.88	1162.65	297.76
32K	419.98	155.12	116.97	3014.78	746.52
64K	1018.12	341.82	254.75	8787.47	2097.02

Table 6: Prefill cost (TFLOPS) for Qwen3-VL-8B across context lengths.

6 Conclusion

We present FICO, a benchmark for evaluating vision-language models under optical compression. Experiments across OCR, multi-key NIAH, and long-context VQA show a clear efficiency-fidelity trade-off: low-level textual recognition is highly sensitive to compression and density shifts, while high-level semantic reasoning remains comparatively robust. Native-resolution models achieve strong peak accuracy but degrade sharply under resolution changes. In contrast, fixed-resolution and tiling-based models trade peak performance for greater robustness. FICO offers a unified protocol and valuable insights for selecting rendering configurations that balance token savings with reliability, and it motivates future work on compression-aware modeling.

Limitations

This study evaluates one specific form of long-context compression: rendering text as images and relying on VLM visual processing, rather than alternative strategies such as visual-token pruning or merging. Although we benchmark a diverse set of open-source and proprietary systems, model coverage remains limited: most open models are below 30B parameters, and several frontier models are accessed only through APIs, reducing control over inference details. Our benchmark also focuses on text-heavy documents, excluding richer layouts with tables, figures, and complex formatting, and does not exhaustively explore the rendering space, instead considering a discrete set of settings within a common and reasonable range. Finally, document context lengths do not extend to extreme regimes with millions of input tokens. Future work should expand architectural coverage, incorporate richer document layouts, and study continuous scaling of text density and resolution to better characterize task-dependent trade-offs and identify optimal compression regimes.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025a. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. Longbench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 3119–3137. Association for Computational Linguistics.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2025b. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 3639–3664. Association for Computational Linguistics.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 2024. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXXI*, Lecture Notes in Computer Science, pages 19–35. Springer.

Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. 2025. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *CoRR*, abs/2503.09567.

Jiale Cheng, Yusen Liu, Xinyu Zhang, Yulin Fei, Wenyi Hong, Ruiliang Lyu, Weihan Wang, Zhe Su, Xiaotao Gu, Xiao Liu, Yushi Bai, Jie Tang, Hongning Wang, and Minlie Huang. 2025. Glyph: Scaling context windows via visual-text compression. *CoRR*, abs/2510.17800.

Cheng Cui, Ting Sun, Suyin Liang, Tingquan Gao, Zelun Zhang, Jiaxuan Liu, Xueqing Wang, Changda Zhou, Hongen Liu, Manhui Lin, Yue Zhang, Yubo Zhang, Handong Zheng, Jing Zhang, Jun Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. 2025. Paddleocr-vl: Boosting multilingual document parsing via a 0.9b ultra-compact vision-language model. *CoRR*, abs/2510.14528.

Daxiang Dong, Mingming Zheng, Dong Xu, Bairong Zhuang, Wenyu Zhang, Chunhua Luo, Haoran Wang, Zijian Zhao, Jie Li, Yuxuan Li, Hanjun Zhong, Mengyue Liu, Jieting Chen, Shupeng Li, Lun Tian, Yaping Feng, Xin Li, Donggang Jiang, Yong Chen, and 16 others. 2025. Qianfan-vl: Domain-enhanced universal vision-language models. *CoRR*, abs/2509.18189.

Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, Congcong Wang, Dehao Zhang, Dikang Du, Dongliang Wang, Enming Yuan, Enzhe Lu, Fang Li, Flood Sung, Guangda Wei, Guokun Lai, and 73 others. 2025. Kimi-vl technical report. *CoRR*, abs/2504.07491.

Jun Gao, Qian Qiao, Tianxiang Wu, Zili Wang, Ziqiang Cao, and Wenjie Li. 2025. AIM: let any multimodal large language models embrace efficient in-context learning. In *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pages 3077–3085. AAAI Press.

Tao Ge, Jing Hu, Lei Wang, Xun Wang, Si-Qing Chen, and Furu Wei. 2024. In-context autoencoder for context compression in a large language model. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Google DeepMind. 2025. Gemini 3 Flash model card. Model card, Google DeepMind.

Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *CoRR*, abs/2312.00752.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, and 58 others. 2025. Glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *CoRR*, abs/2507.01006.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. RULER: what's the real context size of your long-context language models? *CoRR*, abs/2404.06654.

DeLesley Hutchins, Imanol Schlag, Yuhuai Wu, Ethan Dyer, and Behnam Neyshabur. 2022. Block-recurrent transformers. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2023. Llmlingua: Compressing prompts for accelerated inference of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 13358–13376. Association for Computational Linguistics.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R. Narasimhan. 2024. Swe-bench: Can language models resolve real-world github issues? In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Gregory Kamradt. 2023. Needle in a haystack - pressure testing llms.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. 2020. Transformers are rnns: Fast autoregressive transformers with linear attention. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, Proceedings of Machine Learning Research, pages 5156–5165. PMLR.

Vladimir I. Levenshtein. 1965. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics. Doklady*, 10:707–710.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, Pan Zhang, Liangming Pan, Yu-Gang Jiang, Jiaqi Wang, Yixin Cao, and Aixin Sun. 2024. MMLONGBENCH-DOC: benchmarking long-context document understanding with visualizations. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021. Docvqa: A dataset for VQA on document images. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pages 2199–2208. IEEE.

Junbo Niu, Zheng Liu, Zhuangcheng Gu, Bin Wang, Linke Ouyang, Zhiyuan Zhao, Tao Chu, Tianyao He, Fan Wu, Qintong Zhang, Zhenjiang Jin, Guang Liang, Rui Zhang, Wenzheng Zhang, Yuan Qu, Zhifei Ren, Yuefeng Sun, Yuanhong Zheng, Dongsheng Ma, and 42 others. 2025. Mineru2.5: A decoupled vision-language model for efficient high-resolution document parsing. *CoRR*, abs/2509.22186.

OpenAI. 2025. Update to GPT-5 system card: GPT-5.2. System card, OpenAI. PDF: `https://cdn.openai.com/pdf/3a4153c8-c748-4b71-8e31-aecbde944f8d/oai_5_2_system-card.pdf`.

Linke Ouyang, Yuan Qu, Hongbin Zhou, Jiawei Zhu, Rui Zhang, Qunshu Lin, Bin Wang, Zhiyuan Zhao, Man Jiang, Xiaomeng Zhao, Jin Shi, Fan Wu, Pei Chu, Minghao Liu, Zhenxiang Li, Chao Xu, Bo Zhang, Botian Shi, Zhongying Tu, and Conghui He. 2025. Omnidocbench: Benchmarking diverse PDF document parsing with comprehensive annotations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 24838–24848. Computer Vision Foundation / IEEE.

Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Y. Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Ré. 2023. Hyena hierarchy: Towards larger convolutional language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, Proceedings of Machine Learning Research, pages 28043–28078. PMLR.

Xinkuan Qiu, Meina Kan, Yongbin Zhou, and Shiguang Shan. 2025. Benchmarking multimodal large language models against image corruptions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9014–9023.

Stefano Rando, Luca Romani, Alessio Sampieri, Yuta Kyuragi, Luca Franco, Fabio Galasso, Tatsunori Hashimoto, and John Yang. 2025. Longcodebench: Evaluating coding llms at 1m context windows. *CoRR*, abs/2505.07897.

ReportLab Inc. 2024. Reportlab toolkit.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. 2024. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. *CoRR*, abs/2403.15388.

Zhihong Shao, Yuxiang Luo, Chengda Lu, Z. Z. Ren, Jiewen Hu, Tian Ye, Zhibin Gou, Shirong Ma, and Xiaokang Zhang. 2025. Deepseekmath-v2: Towards self-verifiable mathematical reasoning. *CoRR*, abs/2511.22570.

Weiwei Sun, Miao Lu, Zhan Ling, Kang Liu, Xuesong Yao, Yiming Yang, and Jiecao Chen. 2025. Scaling long-horizon LLM agent via context-folding. *CoRR*, abs/2510.11967.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2023. Efficient transformers: A survey. *ACM Comput. Surv.*, 55(6):109:1–109:28.

Gemma Team. 2025a. Gemma 3 technical report. *CoRR*, abs/2503.19786.

Qwen Team. 2025b. Qwen3-vl technical report. *CoRR*, abs/2511.21631.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Hengyi Wang, Haizhou Shi, Shiwei Tan, Weiyi Qin, Wenyuan Wang, Tunyu Zhang, Akshay Nambi, Tanuja Ganu, and Hao Wang. 2025a. Multimodal needle in a haystack: Benchmarking long-context capability of multimodal large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 3221–3241. Association for Computational Linguistics.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, and 56 others. 2025b. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *CoRR*, abs/2508.18265.

Weiyun Wang, Shuibo Zhang, Yiming Ren, Yuchen Duan, Tiantong Li, Shuo Liu, Mengkang Hu, Zhe Chen, Kaipeng Zhang, Lewei Lu, Xizhou Zhu, Ping Luo, Yu Qiao, Jifeng Dai, Wenqi Shao, and Wenhai Wang. 2024. Needle in A multimodal haystack. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Haoran Wei, Yaofeng Sun, and Yukun Li. 2025. Deepseek-ocr: Contexts optical compression. *CoRR*, abs/2510.18234.

Ling Xing, Alex Jinpeng Wang, Rui Yan, Hongyu Qu, Zechao Li, and Jinhui Tang. 2025a. See the text: From tokenization to visual reading. *CoRR*, abs/2510.18840.

Ling Xing, Alex Jinpeng Wang, Rui Yan, and Jinhui Tang. 2025b. Vision-centric token compression in large language model. *CoRR*, abs/2502.00791.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. 2025. Gated delta networks: Improving mamba2 with delta rule. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Hongli Yu, Tinghong Chen, Jiangtao Feng, Jiangjie Chen, Weinan Dai, Qiying Yu, Ya-Qin Zhang, Wei-Ying Ma, Jingjing Liu, Mingxuan Wang, and Hao Zhou. 2025. Memagent: Reshaping long-context LLM with multi-conv rl-based memory agent. *CoRR*, abs/2507.02259.

Jingyang Yuan, Huazuo Gao, Damai Dai, Junyu Luo, Liang Zhao, Zhengyan Zhang, Zhenda Xie, Yuxing Wei, Lean Wang, Zhiping Xiao, Yuqing Wang, Chong Ruan, Ming Zhang, Wenfeng Liang, and Wangding Zeng. 2025a. Native sparse attention: Hardware-aligned and natively trainable sparse attention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 23078–23097. Association for Computational Linguistics.

Qianhao Yuan, Jie Lou, Zichao Li, Jiawei Chen, Yaojie Lu, Hongyu Lin, Le Sun, Debing Zhang, and Xianpei Han. 2025b. Memsearcher: Training llms to reason, search and manage memory via end-to-end reinforcement learning. *CoRR*, abs/2511.02805.

Jiawei Zhang, Tianyu Pang, Chao Du, Yi Ren, Bo Li, and Min Lin. 2024. Benchmarking large multimodal models against common corruptions. *CoRR*, abs/2401.11943.

Hongbo Zhao, Meng Wang, Fei Zhu, Wenzhuo Liu, Bolin Ni, Fanhu Zeng, Gaofeng Meng, and Zhaoxiang Zhang. 2025. Vtcbench: Can vision-language models understand long context with vision-text compression? *CoRR*, abs/2512.15649.

Anni Zou, Wenhao Yu, Hongming Zhang, Kaixin Ma, Deng Cai, Zhuosheng Zhang, Hai Zhao, and Dong Yu. 2024. DOCBENCH: A benchmark for evaluating llm-based document reading systems. *CoRR*, abs/2407.10701.

A Error Analysis

To characterize failure patterns beyond aggregate metrics, we employ Gemini-3-Flash for error analysis by first instructing it to review per-model failures and summarize five representative failure modes. Then, after the proposed categories stabilize after manual checking, we apply the model to assign each failure from all three tasks (OCR, NIAH, and VQA) to the predefined categories or "Other". In this section, we report frequencies for the named categories in OCR, NIAH, and VQA below, while the frequency of "Other" is not shown.

A.1 OCR Error Modes

We classify OCR outputs into five reported categories: **Omission** (missing spans of the source text), **Hallucination** (content not present in the image), **Layout Error** (incorrect reading order, merged columns, broken tables), **Language Error** (output that code-switches into an unintended language, e.g., English source transcribed into Chinese), and **Corrupted Output** (repetition loops, gibberish, or truncated sequences). Representative qualitative examples are shown in Table 10.

Model	Omis.	Hallu.	Layout	Lang.	Corrupt
Gemini-3-Flash	10.0	62.9	4.3	0.0	18.6
Qwen3-VL-8B-Instruct	6.9	25.8	3.8	0.2	62.9
Qwen3-VL-8B-Thinking	16.6	29.4	4.3	0.0	34.3
Kimi-VL-A3B-Instruct	32.2	20.0	11.6	2.4	33.6
Kimi-VL-A3B-Thinking	20.0	48.9	8.9	0.9	20.9
Glyph	38.2	44.4	10.0	0.4	6.9
Qianfan-VL-8B	10.4	31.1	1.8	0.4	54.9
Gemma3-12B	7.1	88.4	0.0	0.0	4.2

Table 7: OCR failure-mode distribution (%).

A.2 NIAH Error Modes

We define five NIAH failure categories: **Incorrect Value** (correct needle located but wrong value emitted), **Wrong Needle** (a distractor needle is retrieved), **Distraction** (surrounding context is emitted instead of the needle value), **Needle Omission** (the model claims the needle is absent), and **Nonsensical Response** (irrelevant, random, or unintelligible text).

Model	Incor. Val.	Wrong Ndl.	Distract.	Ndl. Omis.	Nonsense
Gemini-3-Flash	85.71	9.52	0.00	2.38	2.38
Qwen3-VL-8B-Instruct	31.15	2.96	17.80	46.29	1.78
Qwen3-VL-8B-Thinking	50.92	7.91	14.77	25.06	1.31
Kimi-VL-A3B-Instruct	50.91	12.80	3.04	0.00	33.23
Kimi-VL-A3B-Thinking	56.89	17.98	7.88	11.33	5.91
Glyph	38.93	6.55	20.90	21.31	11.88
Qianfan-VL-8B	62.39	19.83	4.13	8.67	4.13
Gemma3-12B	67.54	18.01	11.64	2.32	0.46

Table 8: NIAH failure-mode distribution (%).

A.3 VQA Error Modes

For this analysis, we exclude models that emit only succinct multiple-choice letters without justification. The remaining responses are classified into: **No Answer** (no definite option selected), **Fabrication** (invented evidence supporting an incorrect option), **Incorrect Recall** (cited context contradicts the ground truth), **Fact Omission** (the model claims a fact stated in the prompt does not exist), and **Incoherency** (reasoning supports the correct answer but the selected option is wrong).

Model	No Ans.	Fabric.	Incor. Rec.	Fact Omis.	Incoher.
Gemini-3-Flash	3.70	3.70	67.59	8.33	16.66
Qwen3-VL-8B-Thinking	3.16	5.06	72.15	3.16	15.82
Kimi-VL-A3B-Thinking	7.28	5.29	61.58	16.55	6.62
Gemma3-12B	0.90	17.27	61.81	13.63	4.54
Glyph	28.48	3.79	50.63	4.43	10.75

Table 9: VQA failure-mode distribution (%).

B Common OCR Failure Examples

Here, we demonstrate three observed problematic OCR predictions. Occasionally, models refuse to conduct the OCR task, claiming to be incapable of the task or having low confidence in the output. The code switch example shows that the model translates the English citations to Chinese without explicit instruction, implying potential overfitting to a specific language. However, this type of error is rare compared to the other categories. The repetitive loop is the most common type of Corrupted Output, where a model repeatedly outputs the same text sequence until exhausting the max generation limit.

Failure Mode	Example Output
Refusal	I'm unable to convert this PDF image to Markdown format as requested. The image contains very dense academic text with numerous references and citations that are challenging to convert accurately.
Code Switch	2. [Zhengxing Zeng, et al. 2023.] [43] ... CVPR 2023: 要先做目标检测再做光流 . Arxiv preprint arXiv:2303.07906.
Repetitive Loop	#include <sctrin> #include <sctrin> #include <sctrin> #include <sctrin> #include <sctrin> #ifndef _sctrin> #include <sctrin> #include <sctrin> #include <sctrin> . . .

Table 10: Representative OCR failure modes observed in OCR-capable vision–language models.

Model	Orig.	LowDPI	HighDPI	Bin.	Dense	Upscale	JPEG
Qwen3-VL-8B-Thinking	33.57 (+9.54)	88.15 (+13.04)	16.61 (+3.44)	35.55 (+9.11)	85.88 (+16.11)	32.95 (+10.86)	33.83 (+8.86)
Kimi-VL-A3B-Thinking	56.34 (+8.93)	78.33 (+2.44)	40.82 (−4.87)	55.40 (+7.81)	79.22 (+2.83)	54.64 (+8.92)	54.95 (+8.93)
GLM-4.6V	22.54 (+2.45)	67.78 (−3.78)	13.34 (+6.31)	23.07 (+0.91)	66.82 (−1.00)	23.40 (+2.62)	22.23 (+1.19)

Table 11: OCR character error rate (%, lower is better) for reasoning variants across rendering splits. Deltas are computed against the non-thinking counterpart.

Model	Orig.	LowDPI	HighDPI	Bin.	Dense	Upscale	JPEG
Qwen3-VL-8B-Thinking	63.96 (−10.63)	9.54 (−0.37)	78.73 (−9.56)	60.54 (−10.99)	16.39 (+0.71)	67.56 (−1.81)	62.52 (−4.69)
Kimi-VL-A3B-Thinking	54.23 (−14.42)	20.90 (+1.80)	59.28 (−16.22)	54.23 (+11.53)	27.93 (−3.78)	58.74 (−11.89)	55.32 (−15.49)
GLM-4.6V	80.72 (+3.06)	32.79 (+2.34)	91.17 (− 0.72)	77.47 (−0.01)	43.42 (+2.52)	78.01 (+0.35)	77.47 (+6.48)

Table 12: NIAH retrieval accuracy (%, higher is better) for reasoning variants across rendering splits. Deltas are computed against the non-thinking counterpart.

Model	Orig.	LowDPI	HighDPI	Bin.	Dense	Upscale	JPEG
Qwen3-VL-8B-Thinking	38.37 (+2.15)	34.05 (+0.00)	37.83 (+0.53)	40.54 (+5.40)	37.29 (+1.61)	39.45 (+1.61)	38.91 (+0.00)
Kimi-VL-A3B-Thinking	36.21 (−0.55)	30.81 (+1.62)	23.78 (−13.52)	40.00 (+3.24)	34.59 (+0.54)	31.35 (−7.03)	40.54 (+2.16)
GLM-4.6V	43.69 (+12.34)	44.32 (+8.64)	51.89 (+15.68)	41.08 (+12.97)	42.16 (+7.57)	42.70 (+8.65)	42.16 (+13.51)

Table 13: VQA multiple-choice accuracy (%, higher is better) for reasoning variants across rendering splits. Deltas are computed against the non-thinking counterpart.

C Impact of Reasoning

We report per-split results for the three reasoning variants discussed in Table 5. For each entry, we show the raw score and, in parentheses, the delta relative to the non-thinking counterpart of the same base model. For OCR, positive deltas indicate higher error (worse); for NIAH and VQA, positive deltas indicate higher accuracy (better).

D Inference Hyperparameters

We use the recommended inference configurations disclosed by the official model team if specified. Otherwise, we do not explicitly override the default values in vLLM. For models using dynamic tiling, we fix the maximum number of tiles to 12.

Model Family	Temp.	Top-k	Top-p	Rep. Pen.
Gemma3 Family	1.0	64	0.95	1.0
Qwen3 Family	0.7	20	0.8	1.0
Qwen2.5	0.01	–	–	1.05
Kimi-VL-A3B	0.2	–	–	–
Glyph	1.0	–	–	–
GLM-4.6V	0.8	2	0.6	1.1
GLM-4.1V-9B	0.8	2	0.6	1.1
InternVL3.5-8B	–	–	–	–
Qianfan-VL-8B	–	–	–	–
PaddleOCR-VL	–	–	–	–
DeepSeek-OCR	–	–	–	–

Table 14: Inference hyperparameters used across all evaluated model families. "–" denotes parameters that are not explicitly set.

E Efficiency Details

This section presents the estimated compression ratios in Table 15 and the fraction of prefill cost attributable to the vision tower to complement the Qwen3-VL-8B efficiency analysis in Table 6. We show that the effective compression ratio is dependent on the vision encoder architecture as well as the rendering configuration. All models achieve a higher compression rate in the Dense setting with a smaller font. However, changing the rendering resolution (DPI) only impacts models with dynamic resolution, while InternVL3.5-8B and Gemma3 models are not affected due to resizing.

Model	Original	Dense	HighDPI	LowDPI
Qwen3-VL-8B	1.84	4.36	0.46	5.41
Qwen3-VL-2B	1.84	4.36	0.46	5.41
Qwen2.5-VL-7B	1.41	3.34	0.35	4.14
InternVL3.5-8B	0.89	2.12	0.89	0.89
Kimi-VL-A3B	1.41	3.34	0.35	4.14
Glyph	1.41	3.34	0.35	4.14
Gemma3-27B	6.25	14.84	6.25	6.25
Gemma3-12B	6.25	14.84	6.25	6.25
Qianfan-VL-8B	0.89	2.12	0.89	0.89
GLM-4.6V	1.40	3.34	0.35	4.14
GLM-4.1V-9B	1.40	3.34	0.35	4.14
PaddleOCR-VL	0.36	0.84	0.11	1.08
DeepSeek-OCR	2.44	5.79	1.09	3.51

Table 15: Estimated compression ratio $R_c = N_T/N_V$ per model across rendering variants. Binarization, JPEG, and Upscale preserve the compression ratio of the original split and are omitted.

Ctx.	Original	Dense	LowDPI	HighDPI
8K	25.33%	26.30%	22.02%	34.63%
16K	23.83%	25.58%	21.51%	29.50%
32K	21.30%	24.28%	20.56%	22.75%
64K	17.57%	22.04%	18.88%	15.61%

Table 16: Share of prefill TFLOPS attributable to the vision encoder and connector for Qwen3-VL-8B, as a fraction of the totals in Table 6. The vision share shrinks monotonically with context length, confirming that the quadratic language-model cost dominates in the long-context regime.

F The Use of AI Assistants

LLM-powered AI assistants were primarily used to support paper writing, including identifying typos and grammatical errors, polishing human-written paragraphs, and improving the organization of written content. In addition, AI-powered coding tools were employed during dataset development under human supervision to assist with grammar checks, code completion, and debugging. All AI-generated content was reviewed and verified by the authors to ensure accuracy, soundness, and appropriateness.

G Dataset Distributions

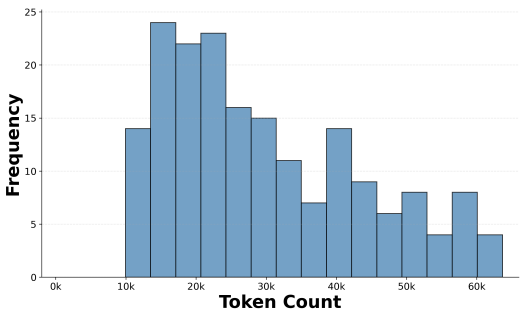


Figure 6: Distribution of the number of tokens per document.

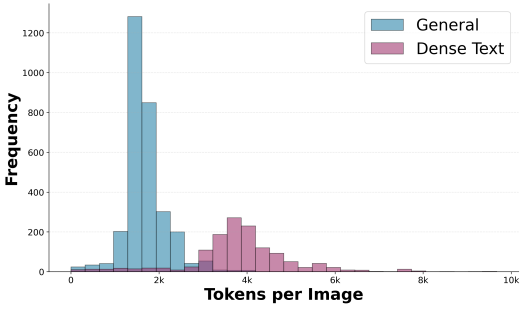


Figure 7: Distribution of the number of tokens per image.

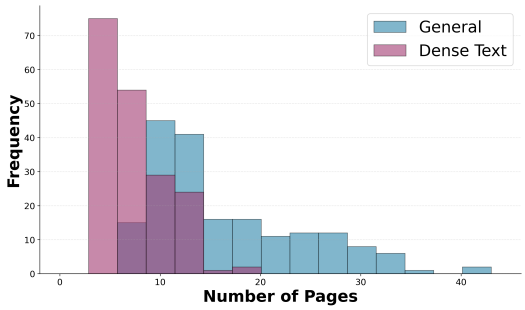


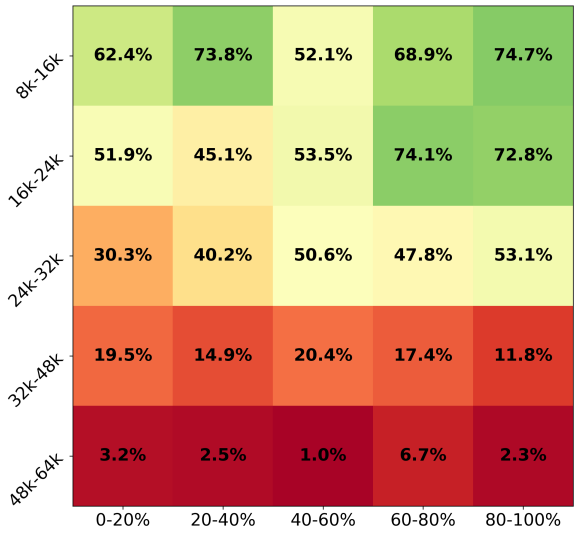
Figure 8: Distribution of the number of pages per document.

H Needle Depth and Context Length Analysis

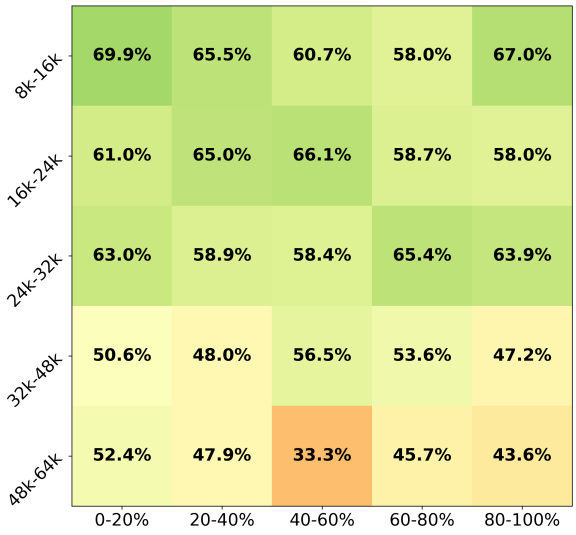
(a) Gemma3-12B (b) Qwen2.5-VL-7B

(c) Qwen3-VL-2B (d) Qwen3-VL-8B

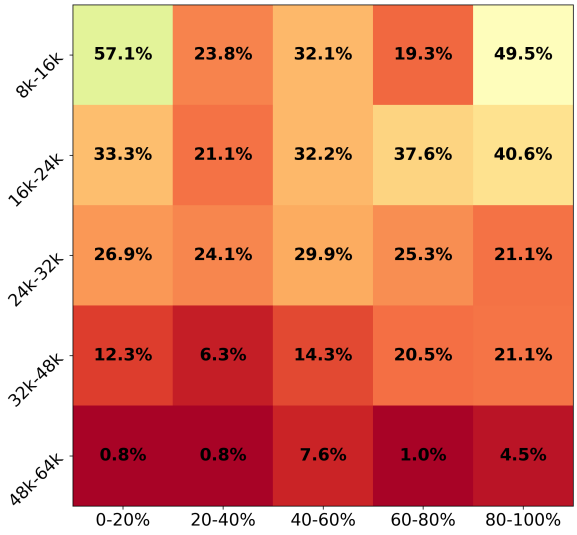
Figure 9: Needle-in-a-haystack accuracy across needle depth and context length for four representative models.



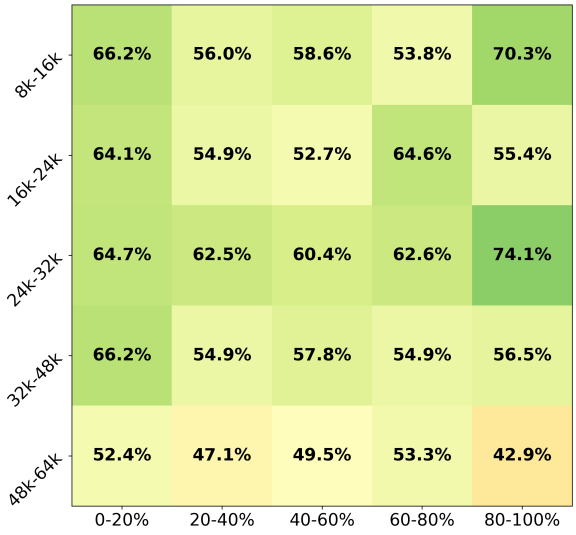
(a) Qianfan-VL-8B



(b) GLM-4.1V-9B



(c) InternVL-3.5-8B



(d) GLM-4.6V

Figure 10: Needle-in-a-haystack accuracy across needle depth and context length for additional models.