When Visual Marks Help Text-Rich Prompts: A Paired Analysis of Annotation Modality

Anonymous ACL submission

Abstract

Multimodal models can be given evidence in two ways: as markup in text, or as visual marks over rendered text. We ask when these annotation channels change performance. In a paired design, each item keeps the same content, question, answer, and gold evidence while varying only the presentation: plain text, marked text, structured text evidence, plain image, or marked image. We evaluate synthetic reading, table, symbolic, form, and multi-hop tasks, together with DocVQA and FUNSD-style document tasks. Across Qwen2.5-VL-3B/7B/32B/72B and InternVL3-9B paired rows, image marks help more than text markup mainly in layout-centered document tasks: FUNSD gains 33.7 points from image marks versus 21.3 from text markup, and DocVQA gains 5.1 versus 2.2. In contrast, multi-hop retrieval gains 11.4 points from text markup but 5.1 from image marks; harder table lookup shows the same direction. Structured text evidence is often stronger than either human-like highlighting channel. Visual marks are therefore best understood as a spatial aid, not as a general replacement for symbolic evidence in text.

1 Introduction

Text-rich multimodal prompts often contain more information than the question needs. A user might highlight a sentence in a screenshot, box a form field, underline a definition, or tag the same evidence in text. These interfaces are semantically similar, but they enter different model pathways. Text markup is read as additional tokens; visual marks are perceived through the image encoder and then bound to text recovered from pixels.

This paper studies a narrow question: *when the evidence is held fixed, does annotating an image help more than annotating text?* The question is not whether rendered text is better than text tokens. Prior work suggests that it often is not: visualized-text evaluations report a text-to-pixel modality gap, sensitivity to rendering, and degraded long-context reasoning over dense visual text (Liu et al., 2026; Sun et al., 2026; Zhao et al., 2025). At the same time, visual prompting methods show that marks, boxes, and labels can improve spatial grounding in images (Yang et al., 2023). The open issue is how these forces trade off when the content is mostly text.

We answer with a paired behavioral analysis. Each canonical item is compiled into matched text and image conditions, and the main statistic compares the gain from text markup with the gain from image marking. The resulting pattern is not uniform. Visual marks help most on document and form tasks where spatial layout is part of the problem. Text markup is stronger on multi-hop and harder table settings where the model must compose discrete facts. A structured evidence sidecar is a strong diagnostic baseline, suggesting that many failures are not failures to notice evidence, but failures to convert marked regions into stable symbolic state.

2 Related Work

Visual prompting. Visual prompting modifies an image to guide a multimodal model toward relevant regions. Set-of-Mark prompting overlays masks, boxes, and alphanumeric marks on images so that a model can refer to regions more reliably (Yang et al., 2023). Graph-of-Mark extends this idea by explicitly representing relations among marked regions (Frisoni et al., 2026). This line of work motivates image-side annotation as an interface for spatial grounding. Our setting differs in that most of the content is text, so a visual mark must also survive OCR and text-region binding.

Text as pixels. Recent visualized-text work shows that equivalent semantic content can behave differently when it is tokenized as text versus ren-

dered as an image. VISTA-Bench contrasts pure text with rendered text under controlled rendering conditions and finds a substantial modality gap (Liu et al., 2026). Sun et al. diagnose this gap as strongly tied to reading and formatting errors, with rendering choices such as font and resolution acting as confounds (Sun et al., 2026). VTCBench studies long-context vision-text compression and finds that OCR ability alone does not imply robust long-context association over visual text (Zhao et al., 2025). Our analysis asks a complementary question: whether visual annotations compensate for this cost.

Documents and forms. Document QA and form understanding are natural domains for visual annotation because answers often depend on layout. DocVQA asks questions over document images where structure is part of the input (Mathew et al., 2021); FUNSD studies noisy scanned forms with entity labels and links (Jaume et al., 2019). We use these tasks to test whether visual marks help when the underlying data is already spatial.

3 Paired Annotation Setup

Conditions. A canonical item consists of unit-labeled content, a question, an answer, and gold evidence unit IDs. Consider a running example with three units: `[u1] Alice leads project Vega. [u2] Vega uses code K7. [u3] Orion uses code M2.` The question is: *What code is used by Alice's project?* The five conditions are:

- T_{plain}: plain unit-labeled text, with no evidence annotation.
- T_{mark}: the same text, but gold evidence units are wrapped in inline tags, e.g., `<hl id="u1">Alice leads project Vega</hl>` and `<hl id="u2">Vega uses code K7</hl>`.
- T_{struct}: the plain text plus a JSON sidecar such as `"evidence_unit_ids":["u1","u2"]`.
- I_{plain}: the same units rendered as a document image, with no visual marks.
- I_{ann}: the same image with the gold evidence units visually marked. Spans are highlighted, table and form cells are boxed, and symbolic definitions are underlined.

All conditions ask the model to return a compact JSON answer and evidence IDs. The prompt text tells image-condition models to read unit labels visible in the rendered image.

Task	Required operation	Image mark
Reading	choose a gold sentence among distractors	highlight
Table	bind row/column headers to a cell	boxes
Symbolic	substitute definitions into an expression	underlines
Form / FUNSD	retrieve value linked to a field	boxes, highlights
Multi-hop	compose two marked facts	highlights
DocVQA	answer from document text and layout	boxes, highlights

Table 1: Task families are grouped by the operation required to use the annotated evidence.

Tasks. Table 1 separates the evaluated tasks by the operation needed to use the evidence. Reading is a single-span retrieval task with distractor sentences. Table lookup requires binding a row key and column key to a target cell. Symbolic tasks require using marked definitions in an expression. Form tasks ask for values associated with fields. Multi-hop tasks require composing two facts, typically person → project and project → code. Length sweeps vary the number of reading units, table rows, form fields, symbolic-chain depth, and multi-hop distractors. DocVQA and FUNSD provide real-document counterparts with ANLS and token-F1 scoring; synthetic tasks use exact match.

Models and estimands. We report complete paired cells for Qwen2.5-VL-3B/7B/32B/72B (Bai et al., 2025) and InternVL3-9B (Zhu et al., 2025). Pairing is item-level: a row enters the main analysis only when the same item has the four primary conditions T_{plain}, T_{mark}, I_{plain}, and I_{ann}. For each model-task cell, we compute:

$$\Delta_T = P(T_{\text{mark}}) - P(T_{\text{plain}}), \quad (1)$$

$$\Delta_I = P(I_{\text{ann}}) - P(I_{\text{plain}}), \quad (2)$$

$$A = \Delta_I - \Delta_T, \quad (3)$$

$$I_a - T_m = P(I_{\text{ann}}) - P(T_{\text{mark}}). \quad (4)$$

Positive Ameans that image-side annotation improves over its image baseline more than text-side markup improves over plain text. Positive $I_a - T_m$means the marked image directly beats the marked text. Confidence intervals use bootstrap resampling over paired items.

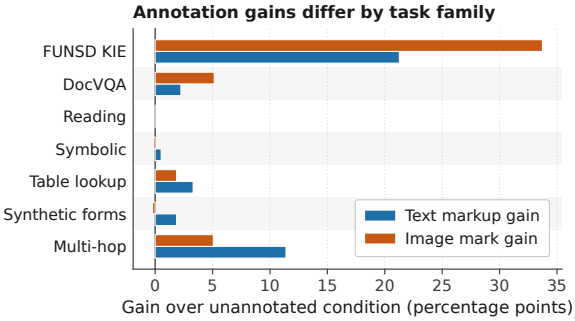

Figure 1: Text markup and image marks help different task families. Bars show paired gains over the corresponding unannotated condition, averaged over clean Qwen paired rows. Visual marks dominate real document tasks; text markup dominates multi-hop retrieval and harder table lookup.

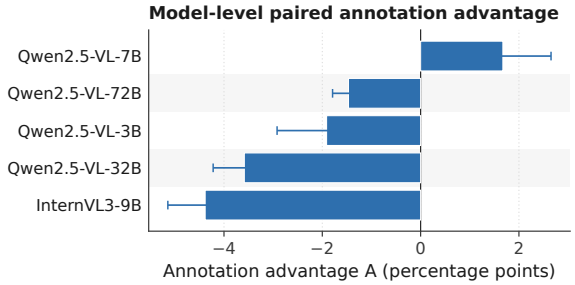

Figure 2: Model-level annotation advantage A averaged across paired tasks. Positive values mean visual marks help more than text markup relative to their own baselines. Only Qwen2.5-VL-7B shows a positive average advantage; the other four models favor text markup overall, reflecting the heavier weight of multi-hop and table-style tasks in the average.

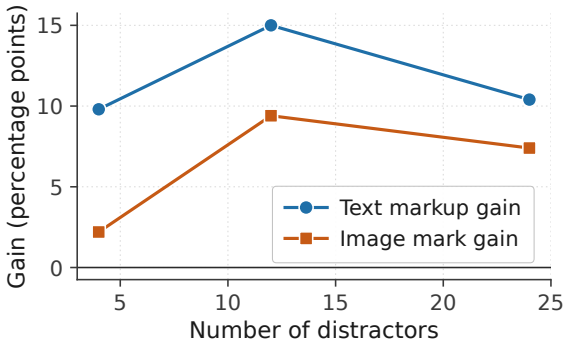

Figure 3: Qwen2.5-VL-72B multi-hop length sweep. Text markup gives a larger gain than visual marks at each distractor count, and the gap persists as the chain becomes harder.

4 Analysis

4.1 The annotation channel matters most when the task is spatial

Figure 1 compares text-markup and image-mark gains across Qwen paired rows, excluding control-only conditions. The positive cases are real document tasks. On FUNSD-style form understanding, image marks improve score by 33.7 points over plain images, while text markup improves by 21.3 points over plain text. On DocVQA, the corresponding gains are 5.1 and 2.2 points. These tasks ask the model to use document layout or field/value structure, so the mark is aligned with the operation.

The negative cases are text-like composition tasks. Multi-hop retrieval gains 11.4 points from text markup but 5.1 points from image marks, giving $A = -6.3$ points. Table lookup averages $A = -1.4$ points, and the difference appears when the table is hard enough: for Qwen2.5-VL-72B on 16-row tables, text markup adds 19.0 points while image marks add 6.2 points. Reading is saturated and produces no measurable gain in either channel. Figures 10 and 11 show that this gap persists across length sweeps: the text-markup advantage widens with the number of distractors and table rows, while image marks plateau.

4.2 Structured text is a strong diagnostic baseline

The structured text condition helps interpret the gap between seeing a mark and using it. On Qwen2.5-VL-72B, T_{struct}scores 99.3% on expanded multi-hop, compared with 88.9% for $T_{\text{mark}}$and 85.1% for I_{ann}. On the 24-distractor multi-hop sweep, T_{struct}scores 96.8%, T_{mark}scores 73.8%, and I_{ann}scores 72.4%. On 16-row table lookup, T_{struct}reaches 100.0%, T_{mark}99.6%, and I_{ann}86.8%. These comparisons suggest that explicitly naming evidence IDs is often easier for the model than interpreting human-like highlighting, especially when the answer requires chaining or indexing.

This does not mean visual marks are ignored. The document tasks show clear image-side gains. Rather, the bottleneck differs by task: for layout questions, a visual region can directly identify where to read; for compositional questions, the model must convert the marked region into symbols and maintain those symbols across steps.

4.3 Controls separate annotation use from task saturation

Control annotations help rule out a weak interpretation of high scores. In easy settings, random or mis-

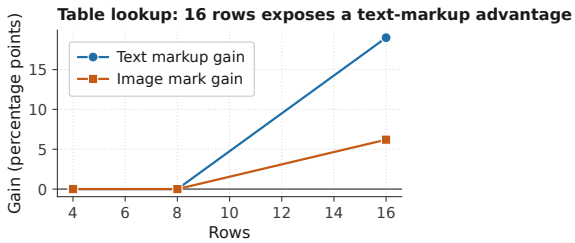

Figure 4: Qwen2.5-VL-72B table lookup sweep. Small tables are saturated for both channels; the 16-row setting exposes a sizable text-markup advantage that image marks do not match.

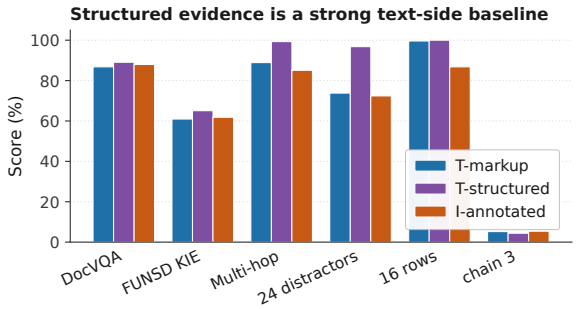

Figure 5: Structured text evidence (T_{struct}) is a strong baseline on selected Qwen2.5-VL-72B cells. On expanded multi-hop and 16-row table lookup, the JSON evidence sidecar matches or beats both the marked-text and marked-image channels, indicating that the bottleneck on compositional tasks is symbolic rather than perceptual.

leading marks can score near gold marks because the base task is already solved. For Qwen2.5-VL-72B symbolic controls, gold visual marks score 89.2%, while random and misleading visual marks both score 89.8%; all-unit marks score 88.2%. The result is not evidence that misleading marks help. It shows that accuracy on saturated tasks is not enough to establish semantic use of the annotation. We therefore treat non-saturated tasks and paired gains as the main evidence.

Item-level outcomes show the same pattern. Most items tie because both annotation channels either solve or fail the item. Among non-ties, multi-hop favors text: 11.4% of items have a larger text gain, compared with 5.4% with a larger image gain. FUNSD reverses this direction: 23.6% favor image marks and 9.6% favor text markup. Figure 8 summarizes this. The appendix gives the full numerical tables.

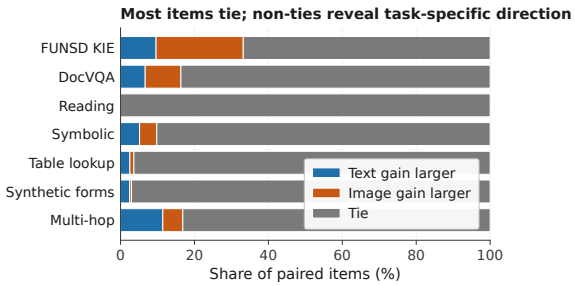

Figure 6: Most paired items are ties; the non-tie direction depends on task family. FUNSD and DocVQA tilt toward image marks, while multi-hop tilts toward text markup. Reading is saturated and has essentially no non-ties.

5 Discussion

The analysis supports a constrained interpretation of visual annotation. Visual marks are effective when the evidence is spatially grounded: a box around a form field or document region matches the way the answer is found. Text markup and structured sidecars are more reliable when the evidence must be composed as discrete symbols. This explains why document tasks favor the image channel while multi-hop and harder table tasks favor text-side evidence.

6 Conclusion

Annotating text as an image is not generally stronger than annotating text as text. Visual marks help most on real document and form tasks, where they point to a spatial region. Text markup and structured evidence help more when the model must chain, index, or calculate over discrete facts. A good annotation channel should match the operation required by the task.

Limitations

The study measures outputs under controlled prompt variants, not internal attention, OCR traces, or causal circuits. Rendering choices are also fixed rather than swept; font size, resolution, and color could change image-condition scores. Finally, several tasks are saturated, which makes null effects hard to interpret. The safest claim is therefore conditional: in this experiment set, visual annotation is a layout aid, while text-side evidence remains better for symbolic composition.

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, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Giacomo Frisoni, Lorenzo Molfetta, Mattia Buzzoni, and Gianluca Moro. 2026. Graph-of-mark: Promote spatial reasoning in multimodal language models with graph-based visual prompting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(36):30726–30734.

Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. Funsd: A dataset for form understanding in noisy scanned documents. *Preprint*, arXiv:1905.13538.

Qing'an Liu, Juntong Feng, Yuhao Wang, Xinzhe Han, Yujie Cheng, Yue Zhu, Haiwen Diao, Yunzhi Zhuge, and Huchuan Lu. 2026. Vista-bench: Do vision-language models really understand visualized text as well as pure text? *Preprint*, arXiv:2602.04802.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021. Docvqa: A dataset for vqa on document images. *Preprint*, arXiv:2007.00398.

Kaiser Sun, Xiaochuang Yuan, Hongjun Liu, Chen Zhao, Cheng Zhang, Mark Dredze, and Fan Bai. 2026. Reading, not thinking: Understanding and bridging the modality gap when text becomes pixels in multimodal llms. *Preprint*, arXiv:2603.09095.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. 2023. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *Preprint*, arXiv:2310.11441.

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? *Preprint*, arXiv:2512.15649.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *Preprint*, arXiv:2504.10479.

A Additional Numerical Results

All appendix values are percentage-point differences unless the table is reporting raw scores. $\Delta_T = T_{\text{mark}} - T_{\text{plain}}$, $\Delta_I = I_{\text{ann}} - I_{\text{plain}}$, $A = \Delta_I - \Delta_T$, and $I_a - T_m = I_{\text{ann}} - T_{\text{mark}}$.

Family	n	Δ_T	Δ_I	A	$I_a - T_m$
Multi-hop	7442	11.4	5.1	-6.3	-8.7
Synthetic forms	5500	1.9	-0.2	-2.1	1.1
Table lookup	7954	3.3	1.9	-1.4	-1.2
Symbolic	7988	0.5	-0.2	-0.7	1.0
Reading	6500	0.0	0.0	0.0	0.0
DocVQA	600	2.2	5.1	2.9	-0.3
FUNSD KIE	1187	21.3	33.7	12.5	12.5

Table 2: Task-family summary over Qwen paired rows.

Model	n	Δ_T	Δ_I	A	$I_a - T_m$
InternVL3-9B	6000	2.9	-1.5	-4.4	-5.9
Qwen2.5-VL-32B	7500	6.1	2.5	-3.6	-3.1
Qwen2.5-VL-3B	6084	5.1	3.2	-1.9	-3.8
Qwen2.5-VL-72B	19040	3.1	1.6	-1.5	-0.4
Qwen2.5-VL-7B	4547	3.7	5.4	1.7	1.9

Table 3: Model-level paired summary.

Task	T_m	T_s	I_a
DocVQA	86.8	89.1	88.0
FUNSD KIE	61.0	65.1	61.8
Multi-hop	88.9	99.3	85.1
24 distractors	73.8	96.8	72.4
16 rows	99.6	100.0	86.8
chain 3	5.2	4.4	5.4

Table 4: Selected Qwen2.5-VL-72B scores comparing marked text, structured text, and marked images.

B Additional Figures

Family	n	Text > img	Img > text	Tie	A
Multi-hop	7442	11.4	5.4	83.1	-6.3
Synthetic forms	5500	2.5	0.5	97.0	-2.1
Table lookup	7954	2.5	1.1	96.4	-1.4
Symbolic	7988	5.1	4.7	90.2	-0.7
Reading	6500	0.0	0.0	100.0	0.0
DocVQA	600	6.7	9.7	83.7	2.9
FUNSD KIE	1187	9.6	23.6	66.8	12.5

Table 5: Item-level direction of annotation gains over Qwen paired rows.

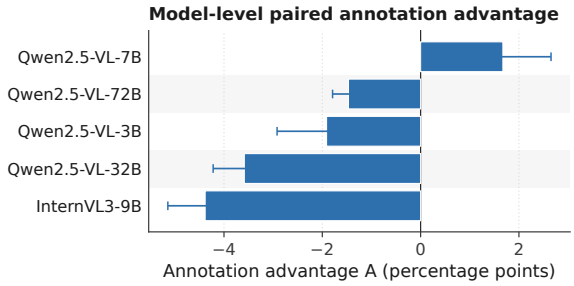

Figure 7: Model-level annotation advantage. Positive values mean visual marks help more than text markup relative to their own baselines.

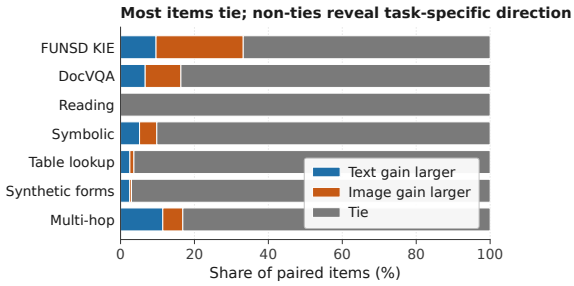

Figure 8: Most paired items are ties; the non-tie direction depends on task family.

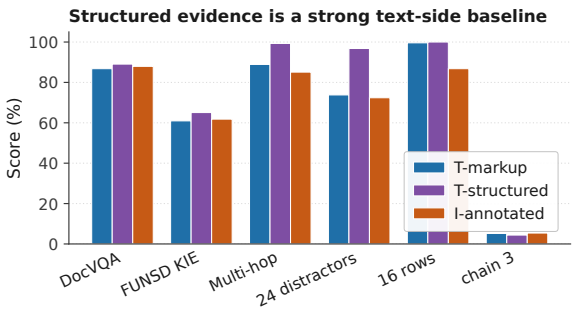

Figure 9: Structured text evidence is a strong baseline on selected Qwen2.5-VL-72B cells.

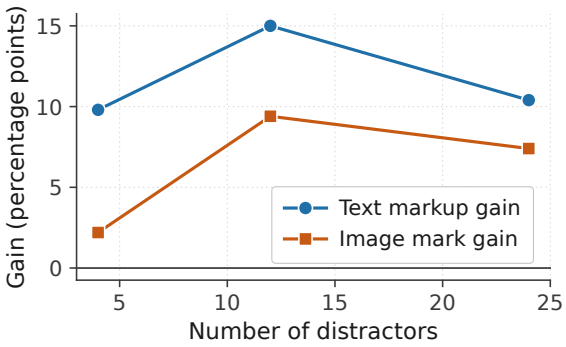

Figure 10: Qwen2.5-VL-72B multi-hop length sweep. Text markup gives a larger gain than visual marks at each distractor count.

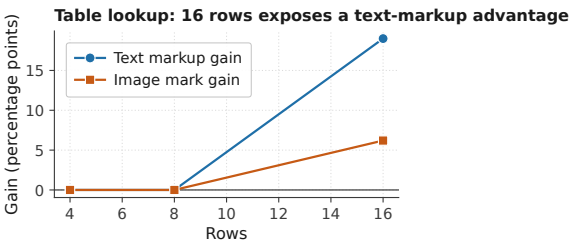

Figure 11: Qwen2.5-VL-72B table sweep. The 16-row setting exposes a larger text-markup gain.

C Full Paired-Cell Table

Table 6 lists every complete-pair model-task cell used for the paired analysis.

Table 6: All complete-pair model-task cells. Scores are percentage-point differences. Control cells are included at the end.

Suite	Model	Task	n	Δ_T	Δ_I	A	$I_a - T_m$
core	Qwen2.5-72B	reading	2000	0.0	0.0	0.0	0.0
core	InternVL3-9B	reading	2000	0.0	0.0	0.0	0.0
core	Qwen2.5-72B	symbolic	2000	-0.1	-2.1	-1.9	-1.1
core	InternVL3-9B	symbolic	2000	8.8	-3.0	-11.8	-16.2
core	Qwen2.5-72B	table	2000	0.0	0.0	0.0	0.0
core	InternVL3-9B	table	2000	0.0	-1.4	-1.4	-1.4
expanded	Qwen2.5-72B	form	1000	0.4	0.0	-0.4	0.0
expanded	Qwen2.5-3B	multi-hop	991	4.5	-2.9	-7.5	-19.7
expanded	Qwen2.5-32B	multi-hop	1000	13.3	7.3	-6.0	-10.3
expanded	Qwen2.5-72B	multi-hop	1000	15.9	8.7	-7.2	-3.8
expanded	Qwen2.5-72B	reading	1000	0.0	0.0	0.0	0.0
expanded	Qwen2.5-3B	symbolic	988	-0.1	0.5	0.6	2.8
expanded	Qwen2.5-7B	symbolic	1000	-0.4	0.9	1.3	3.3
expanded	Qwen2.5-72B	symbolic	1000	1.9	1.2	-0.7	7.0
expanded	Qwen2.5-72B	table	1000	0.0	0.0	0.0	0.0
length	Qwen2.5-3B	fields-12	500	5.8	0.6	-5.2	3.2
length	Qwen2.5-32B	fields-12	500	2.2	0.0	-2.2	1.8
length	Qwen2.5-72B	fields-12	500	0.8	0.0	-0.8	0.0
length	Qwen2.5-3B	fields-20	500	5.4	-3.2	-8.6	2.0
length	Qwen2.5-7B	fields-20	500	-2.2	0.2	2.4	2.0
length	Qwen2.5-32B	fields-20	500	3.2	0.0	-3.2	0.8
length	Qwen2.5-72B	fields-20	500	0.4	0.0	-0.4	0.6
length	Qwen2.5-32B	fields-6	500	4.2	0.0	-4.2	1.8
length	Qwen2.5-72B	fields-6	500	0.0	0.0	0.0	0.0
length	Qwen2.5-3B	hop-d12	451	6.9	2.2	-4.7	-22.2
length	Qwen2.5-7B	hop-d12	500	11.0	4.6	-6.4	-4.8
length	Qwen2.5-32B	hop-d12	500	15.0	4.6	-10.4	-9.2
length	Qwen2.5-72B	hop-d12	500	15.0	9.4	-5.6	-2.0
length	Qwen2.5-7B	hop-d24	500	10.0	6.2	-3.8	-4.2
length	Qwen2.5-32B	hop-d24	500	15.2	9.0	-6.2	-6.6
length	Qwen2.5-72B	hop-d24	500	10.4	7.4	-3.0	-1.4
length	Qwen2.5-32B	hop-d4	500	9.6	3.8	-5.8	-7.8
length	Qwen2.5-72B	hop-d4	500	9.8	2.2	-7.6	-6.0
length	Qwen2.5-72B	read-16	500	0.0	0.0	0.0	0.0
length	Qwen2.5-7B	read-32	500	0.0	0.0	0.0	0.0
length	Qwen2.5-32B	read-32	500	0.0	0.0	0.0	0.0
length	Qwen2.5-72B	read-32	500	0.0	0.0	0.0	0.0
length	Qwen2.5-32B	read-4	500	0.0	0.0	0.0	0.0
length	Qwen2.5-72B	read-4	500	0.0	0.0	0.0	0.0
length	Qwen2.5-72B	read-8	500	0.0	0.0	0.0	0.0
length	Qwen2.5-3B	chain-1	500	3.8	1.6	-2.2	-12.8
length	Qwen2.5-32B	chain-1	500	-2.0	-1.0	1.0	2.6
length	Qwen2.5-72B	chain-1	500	2.2	-1.6	-3.8	1.6
length	Qwen2.5-3B	chain-3	500	0.2	1.0	0.8	1.4
length	Qwen2.5-32B	chain-3	500	-0.2	0.4	0.6	0.2
length	Qwen2.5-72B	chain-3	500	2.2	0.2	-2.0	0.2
length	Qwen2.5-3B	rows-16	500	13.2	4.6	-8.6	-5.2
length	Qwen2.5-7B	rows-16	500	2.4	12.4	10.0	8.4
length	Qwen2.5-32B	rows-16	500	18.0	6.6	-11.4	-9.0
length	Qwen2.5-72B	rows-16	500	19.0	6.2	-12.8	-12.8
length	Qwen2.5-3B	rows-4	454	0.0	0.0	0.0	0.0
length	Qwen2.5-7B	rows-4	500	0.0	0.0	0.0	0.0
length	Qwen2.5-32B	rows-4	500	0.0	0.0	0.0	0.0
length	Qwen2.5-72B	rows-4	500	0.0	0.0	0.0	0.0
length	Qwen2.5-32B	rows-8	500	0.0	0.0	0.0	0.0
length	Qwen2.5-72B	rows-8	500	0.0	0.0	0.0	0.0
real docs	Qwen2.5-3B	DocVQA	200	6.3	10.8	4.5	-3.4
real docs	Qwen2.5-7B	DocVQA	200	-0.2	3.3	3.5	1.4
real docs	Qwen2.5-72B	DocVQA	200	0.6	1.3	0.7	1.2
real docs	Qwen2.5-3B	FUNSD	500	15.8	32.3	16.5	20.1
real docs	Qwen2.5-7B	FUNSD	347	19.3	32.3	13.0	12.3
real docs	Qwen2.5-72B	FUNSD	340	31.3	37.2	5.9	1.4
controls	Qwen2.5-32B	reading	500	0.0	0.0	0.0	0.0
controls	Qwen2.5-72B	reading	500	0.0	0.0	0.0	0.0
controls	Qwen2.5-3B	symbolic	500	1.2	1.0	-0.2	-11.8
controls	Qwen2.5-32B	symbolic	500	-4.0	0.4	4.4	5.4
controls	Qwen2.5-72B	symbolic	500	0.2	-1.8	-2.0	0.0
controls	Qwen2.5-7B	table	500	0.0	0.0	0.0	0.0
controls	Qwen2.5-32B	table	475	0.0	0.0	0.0	0.0
controls	Qwen2.5-72B	table	500	0.0	0.0	0.0	0.0