

# NOT SEARCH, BUT SCAN: BENCHMARKING MLLMS ON SCAN-ORIENTED ACADEMIC PAPER REASONING

Rongjin Li<sup>1</sup>, Zichen Tang<sup>1</sup>, Xianghe Wang<sup>1</sup>, Xinyi Hu<sup>1</sup>, Zhengyu Wang<sup>1</sup>,  
 Zhengyu Lu<sup>1</sup>, Yiling Huang<sup>1</sup>, Jiayuan Chen<sup>1</sup>, Weisheng Tan<sup>1</sup>, Jiacheng Liu<sup>1</sup>,  
 Zhongjun Yang<sup>1</sup>, Haihong E<sup>1,\*</sup>

<sup>1</sup>Beijing University of Posts and Telecommunications

\*Corresponding author.

 [github.com/Staudinger0325/ScholScan](https://github.com/Staudinger0325/ScholScan)

 [huggingface.co/datasets/Staudinger/ScholScan](https://huggingface.co/datasets/Staudinger/ScholScan)

## ABSTRACT

With the rapid progress of multimodal large language models (MLLMs), AI already performs well at literature retrieval and certain reasoning tasks, serving as a capable assistant to human researchers, yet it remains far from autonomous research. The fundamental reason is that current work on scholarly paper reasoning is largely confined to a search-oriented paradigm centered on pre-specified targets, with reasoning grounded in relevance retrieval, which struggles to support researcher-style full-document understanding, reasoning, and verification. To bridge this gap, we propose ScholScan, a new benchmark for scholarly paper reasoning. ScholScan introduces a scan-oriented task setting that asks models to read and cross-check entire papers like human researchers, scanning the document to identify consistency issues. The benchmark comprises 1,800 carefully annotated questions drawn from 9 error families across 13 natural-science domains and 715 papers, and provides detailed annotations for evidence localization and reasoning traces, together with a unified evaluation protocol. We assessed 15 models across 24 input configurations and conduct a fine-grained analysis of MLLM capabilities across error families. Across the board, retrieval-augmented generation (RAG) methods yield no significant improvements, revealing systematic deficiencies of current MLLMs on scan-oriented tasks and underscoring the challenge posed by ScholScan. We expect ScholScan to be the leading and representative work of the scan-oriented task paradigm.

## 1 INTRODUCTION

Enabling multimodal large language models (MLLMs) (OpenAI, 2025; Anthropic, 2025; ByteDance Seed Team, 2025; Meta, 2025; xAI, 2025) to conduct comprehensive understanding and generation based on academic literature is the ultimate goal of Deep Research, and a critical milestone on the path toward artificial general intelligence (AGI) (Ge et al., 2023; Morris et al., 2024; et al., 2025c). With rapid advances, MLLMs are increasingly capable of supporting academic workflows through retrieval, reading, and writing. For example, PaSa (He et al., 2025) can invoke a series of tools to answer complex academic queries with high-quality results, while Google Deep Research (et al., 2025b) is capable of producing human-level research reports based on specific queries.

However, most of the existing work still follows *a search-oriented paradigm*, where models retrieve a few relevant passages and reason over local evidence based on prespecified targets (Gao et al., 2023; Lou et al., 2025). Such methods are effective for tasks with clearly predefined targets, but struggle with researcher-style full-document reasoning and verification (Zhou et al., 2024). ***To function as researchers, models must move beyond reactive question answering and toward proactive discovery of implicit problems.***

To fill this gap, as shown in Figure 1, we introduce *a scan-oriented paradigm*, where models address queries with targets absent and are required to actively **construct a document-level evidence view**,

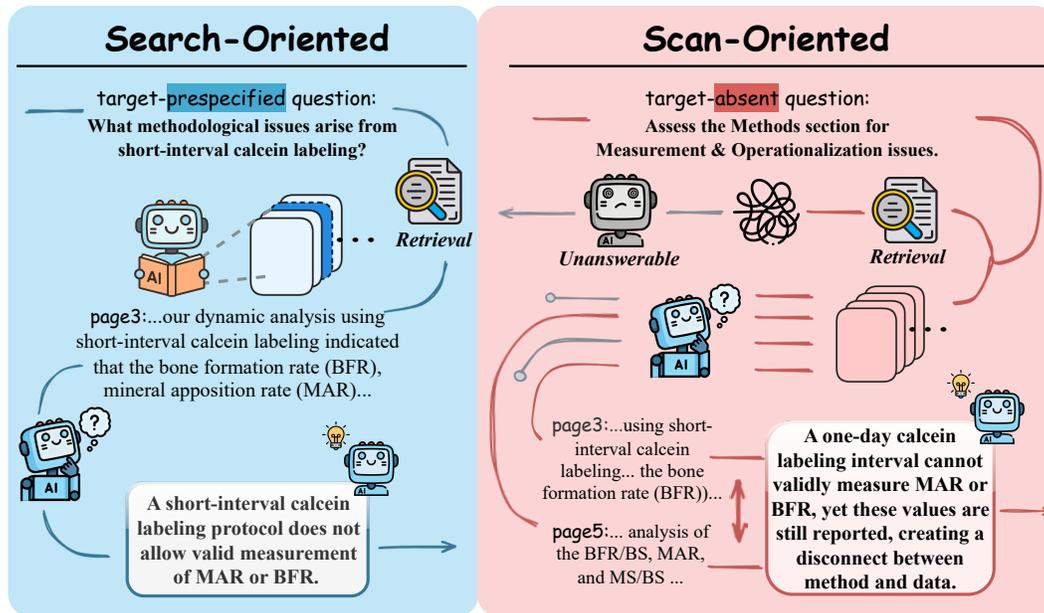


Figure 1: A comparison between search-oriented and scan-oriented task paradigms. Unlike the former, the scan-oriented paradigm provides no prespecified targets, requiring the model to actively scan the entire paper, construct a document-level evidence view.

**perform exhaustive scanning over the full paper, and conduct evidence-based reasoning.** In contrast to search-oriented tasks that assess a model’s ability to identify and reason over *relevant* fragments, scan-oriented tasks emphasize *consistency*. *Instead of relying on prespecified targets or hints, models must derive all necessary concepts and inferences solely from given documents.*

We instantiate this setting via scientific error detection, as it naturally demands discovering non-obvious flaws without target cues, and present ScholScan, a new multimodal benchmark for scholarly reasoning. ScholScan features the following key highlights:

- **Scan-Oriented Task Paradigm.** ScholScan receive one or more complete academic papers together with target-absent queries, presenting a rigorous challenge to their evidence-based reasoning capabilities. The benchmark comprises 715 papers spanning 13 natural science disciplines.
- **Comprehensive Error Types.** ScholScan covers 9 categories of scientific errors across the entire research workflow. It also includes citation and referencing errors, providing a rigorous test of a model’s cross-source reasoning ability.
- **Process-Aware Evaluation Framework.** ScholScan provides fine-grained annotations for both evidence location and reasoning steps, enabling a comprehensive evaluation framework that assesses model performance in terms of both process and outcome.

We evaluate 15 models across 24 input configurations and 8 retrieval-augmented generation (RAG) frameworks. All models exhibit limited performance, and none of the RAG methods deliver significant improvements. These results highlight the inadequacy of search-oriented frameworks when applied to scan-oriented tasks, and underscore both the challenges and the potential of enabling MLLMs to perform reliable, document-level reasoning over full academic papers.

## 2 RELATED WORK

### 2.1 MULTIMODAL LARGE LANGUAGE MODELS

With the rapid progress of MLLMs, models have evolved beyond perception tasks (e.g., image recognition and explanation) (Liu et al., 2024) toward deep understanding of structured, multimodal long documents. Their strengths lie in the ability to integrate cross-modal information and perform multi-hop reasoning over extended contexts. These capabilities are not only valuable for

specific question answering or instruction-following tasks (Yue et al., 2024) but are particularly well suited for simulating human thought processes and generating explainable reasoning trajectories (Zheng et al., 2023). Consequently, achieving comprehensive understanding of entire documents has emerged as a core challenge that MLLMs are inherently equipped to address.

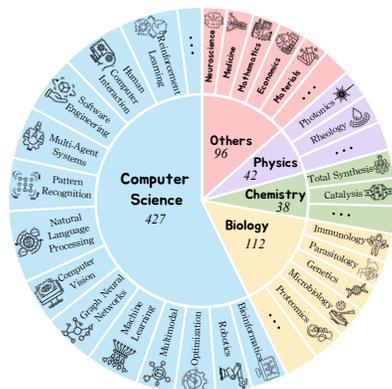
## 2.2 DOCUMENT UNDERSTANDING BENCHMARK

Document understanding tasks challenge models to identify relevant context and perform accurate reasoning grounded in that information. Progress in document understanding benchmarks has followed two main axes. Along the input dimension, it has evolved from short to long contents, from everyday to specialized domains, and from plain text to multimodal format (Chen et al., 2021; Yang et al., 2018; Tito et al., 2021; Deng et al., 2025). Along the scenario dimension, it has shifted from limited-output formats to more open-ended responses (Pramanick et al., 2024). DocMath-Eval (Zhao et al., 2024) evaluates numerical reasoning on long, specialized documents, revealing large performance gaps even for strong models in expert domains, while MMLongBench-Doc (Ma et al., 2024) builds a multimodal benchmark with layout-rich documents. However, a comprehensive benchmark that integrates all challenges above has yet to be introduced.

## 2.3 ACADEMIC PAPER UNDERSTANDING BENCHMARK

Compared with general documents, academic papers are distinguished by their rich domain knowledge and logical rigor. Reasoning over papers has emerged as a major challenge in recent research. Some studies ask for local elements like charts or snippets, leveraging their internal complexity, but neglect the need for cross-source integration and domain-specific interpretation within the full document (Wang et al., 2024; Li et al., 2024). Recent studies extend inputs to the document level and adopt image-based formats to better simulate real-world reading scenarios. (Auer et al., 2023; Yan et al., 2025) However, benchmarks based on the QA paradigm face inherent limitations, as they typically presuppose answer existence and embed explicit cues in the question itself, reducing the need for comprehensive understanding and information organization. Moreover, mainstream evaluation protocols focus on the final outcome, with limited assessment of whether intermediate reasoning is evidentially grounded and logically valid. More examples and analysis are shown in Appendix B.

## 3 THE SCHOLEVAL BENCHMARK



Benchmark	Mod.	Para.	Eval.	# Dom.
<i>Document Understanding</i>				
DocMath-Eval <sub>CompLong</sub>	T+TD	Search	A	N/A
MMLongbench-Doc	T+MD	Search	A	N/A
LongDocURL	T+MD	Search	A	N/A
SlideVQA	T+MD	Search	A	N/A
<i>Academic Paper Understanding</i>				
CharXiv	I	Search	A	8
ArXivQA	I	Search	A	10
MMCR	T+MD	Search	A	CS
AAAR-1.0	T+MD	Search	A	CS
<b>ScholScan (ours)</b>	T+MD	Scan	A+P	13

Figure 2: Left: Overview of ScholScan. Right: Comparison to related benchmarks. **Mod.:** Modalities; **Para.:** Task Paradigm; **Eval.:** Evaluation; **T:** Text; **I:** Image; **TD:** Text-Form Document; **MD:** Multimodal Document; **A:** Answer; **P:** Process; **Dom.:** Number of academic domains in the dataset.

### 3.1 OVERVIEW OF SCHOLSCAN

We introduce ScholScan, a benchmark designed to comprehensively evaluate MLLMs’ ability to detect scientific flaws in academic papers under scan-oriented task settings. As illustrated in Figure 2, ScholScan spans 13 disciplines across the natural sciences, including physics, chemistry, and

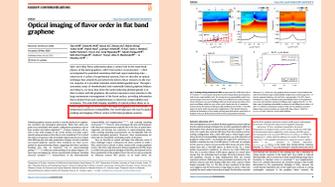
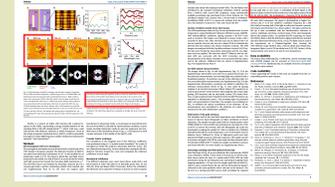
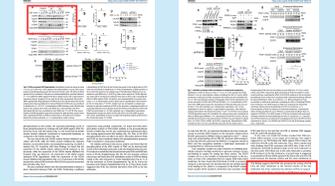
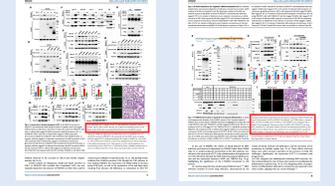
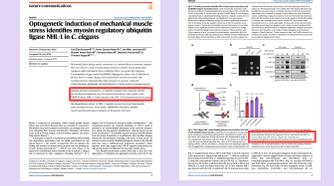
 <b>Research Question &amp; Definitions</b>	 <b>Design &amp; Identifiability</b>	 <b>Sampling &amp; Generalizability</b>
 <p><b>Explanation:</b> The definition of "actionable variants" shifts across sections (LOE 1–5 in Abstract, LOE 1–3 in Results), causing ambiguity.</p>	 <p><b>Explanation:</b> The design is described as probing both short- and long-range interactions, yet the paper still claims unique large-<math>q</math> selectivity, creating a disconnect.</p>	 <p><b>Explanation:</b> The experiments use a narrow diabetic mouse substrain, yet the paper generalizes findings to all patients, creating an invalid sample-to-population inference.</p>
 <b>Measurement &amp; Operationalization</b>	 <b>Data Handling &amp; Preprocessing</b>	 <b>Computation &amp; Formulae</b>
 <p><b>Explanation:</b> First-harmonic demodulation is dominated by far-field background and cannot produce the reported high-quality near-field images.</p>	 <p><b>Explanation:</b> Feature selection for NSCLC and HCC models was done on the full dataset before splitting, causing data leakage, while the Discussion falsely claims unbiased validation.</p>	 <p><b>Explanation:</b> The Methods claim a 200-fold concentration, but the 200 <math>\mu</math>L subsample is incorrectly said to represent <math>\sim</math>20 mL instead of 40 mL, creating a twofold calculation error.</p>
 <b>Inference &amp; Conclusions</b>	 <b>Referential &amp; Citation Alignment</b>	 <b>Language &amp; Expression</b>
 <p><b>Explanation:</b> The data show PGK1 promotes EGFR degradation, yet the Discussion claims inhibiting PGK1 as therapy, directly contradicting the results.</p>	 <p><b>Explanation:</b> Figure 1 report an LPS dose of 1.5 mg/kg, but Figure 5 reports 15 mg/kg, creating a tenfold discrepancy that makes the actual experimental dose unclear.</p>	 <p><b>Explanation:</b> The paper swaps <i>C. elegans</i> gene and protein nomenclature (e.g., 'unc-45' vs. 'UNC-45'), creating technically misleading references.</p>

Figure 3: Sampled ScholScan examples with 9 error types, covering the whole process of scientific research, each requiring the model to perform thorough cross-source evidence-based reasoning.

computer science, and spans over 100 subfields such as immunology, total synthesis, and machine learning. The benchmark comprises 1,800 questions derived from 715 real academic papers, and covers 9 major error categories (Figure 3) that commonly observed in real-world research scenarios. These include issues in numerical and formulaic computation, experimental design, inference and conclusion, and citation misuse, among others. Figure 2 also provides a comparison ScholScan with existing benchmarks for multimodal paper understanding and long-document reasoning.

### 3.2 DATA COLLECTION & QUESTION GENERATION

We curated papers from ICLR 2024/2025 and Nature Communications, and collected public reviews for the former. Questions were constructed based on two dimensions, where the source is either generated or sampled, and the context is either within-paper or cross-paper.

**Generation.** On high-quality accepted papers, we prompt Gemini 2.5 Pro to perform coordinated sentence-level edits spanning multiple sections or pages. It then synthesizes composite errors and generates the corresponding question along with an explanation grounded in the edited context.

**Sampling.** From rejected ICLR submissions and their public reviews, we prompt Gemini 2.5 Pro to extract explicit, falsifiable scientific errors and convert them into questions with initial explanations. Subjective remarks about novelty or writing quality are excluded.

**Within-paper.** This setting focuses on verifiable facts and internal consistency within a single paper, and supports both Generation and Sampling.

**Cross-paper.** This setting examines citation consistency across papers. For each instance, Gemini 2.5 Pro receives an accepted paper and one of its cited sources, then edits the accepted paper to introduce paraphrases or reasoning errors about the citation. As public reviews mainly address nonfalsifiable aspects such as appropriateness, all cross-paper instances are constructed exclusively using the generation method.

### 3.3 QUALITY CONTROL & ANNOTATION

Despite explicit instructions, initial outputs exhibited substantial hallucinations, logical inconsistencies, and low-quality questions. To ensure the quality, 10 domain experts conducted a rigorous annotation process. Each instance underwent independent dual review, and disagreements were resolved by a third expert. Among the 3,500 initially generated candidates, 1,700 were discarded, and 1,541 of the remaining were revised, including 535 question rewrites, 1,207 explanation edits, and 1,141 corrections to error categories or metadata. Further details are provided in Appendix C.

## 4 EXPERIMENTS

### 4.1 EXPERIMENTS SETTING

**Models.** We benchmark a total of 24 input configurations by feeding academic papers as either images or OCR text using the Tesseract (Smith, 2007) engine, covering 15 mainstream models (Yang et al., 2025; Bai et al., 2025; et al., 2025a; Guo et al., 2025; et al., 2025d).

**Evaluation Protocol.** Inspired by MMLongBench-Doc (Ma et al., 2024), we prompt models to generate necessary reasoning chains from evidence to detected anomalies without constraining the output format, which aims to assess the ability for evidence-grounded reasoning rather than mere instruction-following. For open-ended responses, we use GPT-4.1 (OpenAI, 2025) to extract cited evidence and reasoning steps, and quantify alignment with annotated explanations. Human evaluation confirms high agreement between our pipeline and expert annotations. Further implementation details are provided in Appendix E.

**Metrics.** We define a structured evaluation framework by parsing the model response  $a$  into a tuple:

$$\Psi(a) \Rightarrow (\mathbf{1}_{\text{exist}}, \mathbf{1}_{\text{contain}}, \widehat{\mathcal{E}}, \widehat{\mathcal{R}}, n). \quad (1)$$

Here,  $\mathbf{1}_{\text{exist}}$  and  $\mathbf{1}_{\text{contain}}$  are binary indicators for whether output contains any error and includes the annotated target error;  $\widehat{\mathcal{E}}, \widehat{\mathcal{R}}$  and  $\mathcal{E}^*, \mathcal{R}^*$  are the predicted and gold evidence sets and reasoning chains;  $\hat{g} = \text{prefix\_match}(\widehat{\mathcal{R}}, \mathcal{R}^*)$  counts matched reasoning steps;  $n \in \mathbb{N}$  is the number of unrelated errors.  $\text{HasError}(a)$  is 1 if the output contains any predicted error, and 0 otherwise. Based on  $\Psi(a)$ , we define an end-to-end score  $S(m) \in [0, 1]$  that combines all aspects of prediction quality:

(i) *Existence.*  $S_{\text{exist}}(a) = 1$  if and only if the response includes the annotated target error.

$$S_{\text{exist}}(a) = \mathbf{1}\{\text{HasError}(a)\} \cdot \mathbf{1}\{\widehat{\mathcal{E}} \cap \mathcal{E}^* \neq \emptyset\} \quad (2)$$

(ii) *Evidence location score.* Even when the target error is identified, the cited evidence may be incomplete or noisy. We compute a Dice score with a squared penalty for over-reporting:

$$S_{\text{location}} = \max \left\{ 0, \frac{2|\widehat{\mathcal{E}} \cap \mathcal{E}^*| + \mathbf{1}\{|\widehat{\mathcal{E}}| + |\mathcal{E}^*| = 0\}}{\max(|\widehat{\mathcal{E}}| + |\mathcal{E}^*|, 1)} - 0.8 \left( \frac{|\widehat{\mathcal{E}} \setminus \mathcal{E}^*|}{\max(|\widehat{\mathcal{E}}|, 1)} \right)^2 \right\}. \quad (3)$$

(iii) *Reasoning process score.* Even if the target error is detected, the reasoning may diverge from the gold chain. We use prefix match to assess reasoning completeness:

$$S_{\text{reasoning}} = \mathbf{1}\{g_r = 0\} + \mathbf{1}\{g_r > 0\} \left( \frac{\hat{g}}{g_r} \right)^2. \quad (4)$$

(iv) *Unrelated-error penalty*. Models may list unrelated items to inflate recall at the cost of precision. We penalize this with a rapidly increasing function of unrelated error count:

$$P_{\text{unrelated\_err}}(n) = 0.9^{\min(n,2)} \exp\left(-0.6 [\max(n-2,0)]^{1.5}\right). \quad (5)$$

(v) *Overall outcome score*. The final score for  $a$  is defined as:

$$S(m) = S_{\text{exist}}(a) \sqrt{S_{\text{location}} \cdot S_{\text{reasoning}}} \cdot P_{\text{unrelated\_err}}(n). \quad (6)$$

Table 1: Model performance (scaled by 100) across input configurations. **RQD**: Research Question & Definitions; **DI**: Design & Identifiability; **SG**: Sampling & Generalizability; **MO**: Measurement & Operationalization; **DHP**: Data Handling & Preprocessing; **CF**: Computation & Formulae; **IC**: Inference & Conclusions; **RCA**: Referential and Citation Alignment; **LE**: Language & Expression.

Models	Avg.	RQD	DI	SG	MO	DHP	CF	IC	RCA	LE
<b>MLLM (Image Input)</b>										
<i>Proprietary MLLMs</i>										
Gemini 2.5 Pro	15.6	11.9	12.6	35.7	12.3	27.0	4.6	14.7	15.2	7.4
GPT-5	19.2	10.1	9.7	28.2	14.6	26.6	13.8	25.3	25.3	6.9
Grok 4	4.0	0.0	1.9	16.7	3.2	7.4	0.7	1.9	3.6	0.0
Doubao-Seed-1.6-thinking	10.2	3.4	3.5	22.3	7.5	15.1	10.2	12.2	10.9	3.3
Doubao-Seed-1.6	9.9	3.0	4.4	29.2	4.9	15.0	6.3	17.9	8.0	3.9
<i>Open-source LLMs</i>										
Llama 4 Maverick	7.0	7.0	7.3	9.4	4.5	4.0	6.5	6.7	8.8	3.0
Gemma 3 27B	1.7	0.5	2.7	2.3	1.7	1.0	1.0	1.3	2.6	0.0
Mistral Small 3.1	3.3	0.1	2.0	2.0	1.5	0.1	1.0	2.2	8.6	1.0
Qwen2.5 VL 72B	0.1	0.0	0.7	0.0	0.0	0.0	0.0	0.0	0.2	0.0
<b>OCR + LLM (Text Input)</b>										
<i>Proprietary LLMs</i>										
Gemini 2.5 Pro	30.3	21.5	34.2	44.3	27.6	56.6	10.3	28.8	35.6	8.1
GPT-5	22.5	16.1	21.4	26.0	20.3	36.7	4.7	29.8	30.0	2.6
Claude Sonnet 4	5.7	3.7	2.5	10.8	4.3	10.3	1.4	8.4	6.6	3.5
Grok 4	20.8	9.3	7.7	37.4	12.3	34.4	9.0	20.0	31.2	7.2
Doubao-Seed-1.6-thinking	15.3	8.2	10.1	24.3	10.1	24.2	6.4	19.2	21.0	4.2
Doubao-Seed-1.6	13.9	5.4	6.9	26.4	10.3	23.6	6.3	20.1	17.5	2.3
<i>Open-source LLMs</i>										
Qwen3 A22B (Thinking)	17.4	8.9	16.2	31.9	15.1	23.7	5.6	22.3	21.1	2.3
Qwen3 A22B	1.7	1.2	0.0	2.7	0.4	1.0	0.1	4.3	2.5	1.1
gpt-oss-120b	7.3	6.3	5.7	18.3	4.9	14.5	1.6	12.5	5.5	0.0
DeepSeek-R1	11.4	5.1	11.9	25.4	8.7	22.5	4.7	16.3	9.8	3.5
DeepSeek-V3.1	1.7	1.2	2.0	1.7	1.0	5.8	0.5	2.2	2.1	0.0
Llama 4 Maverick	2.3	1.5	2.0	4.8	3.0	3.6	0.0	5.8	1.6	0.2
Gemma 3 27B	2.0	2.1	1.6	3.0	2.7	0.2	0.7	7.7	1.0	0.0
Mistral Small 3.1	6.9	3.0	2.7	5.5	7.0	2.0	8.5	4.0	12.2	3.0
Qwen2.5 VL 72B	0.2	0.0	0.7	0.0	0.0	0.0	0.0	0.0	0.6	0.0

## 4.2 MAIN RESULT

Table 1 presents our evaluation results. Our main findings are summarized as follows:

**Overall performance remains unsatisfactory.** GPT-5 achieves the highest average score in the image input group (19.2), while Gemini 2.5 Pro, the best-performing model in the text input setting, still fails to surpass the 60-point threshold on any subtask. Even in the SG category, which yields the best performance overall, nearly half of the models receive single-digit scores. Most models perform poorly under the scan-oriented task formulation and fail to detect any issues in many papers. This challenge is particularly pronounced for open-source models.

**Reasoning-enhanced models demonstrate clear advantages.** Across both input configurations, reasoning-enhanced variants consistently achieve higher scores. Almost all top-performing models, measured by both subtask-specific and overall metrics, fall into this category. Notably, Qwen3-Thinking and Deepseek-R1 outperform their base versions by more than 10% in average scores,

with substantial gains observed across all error types. These results indicate that reasoning-enhanced models are better able to simulate the iterative process of extraction followed by reasoning, which is essential for effectively handling scan-oriented tasks and producing higher-quality responses.

**MLLMs face significant bottlenecks in handling long multimodal inputs.** Across most evaluation metrics, text inputs outperform image inputs. Among the nine MLLMs tested, the average performance gap between text and image inputs reaches 4.81 points, highlighting visual processing as a key limitation in current MLLM capabilities.

**In most evaluation metrics, text inputs consistently outperform image inputs.** Among the nine MLLMs evaluated, the average performance gap between text and image inputs is 4.81 points, underscoring visual processing as a key limitation in current MLLM capabilities.

**Although overall performance is generally weaker, multimodal input remains indispensable.** In certain categories such as CF, where OCR-based text extraction leads to substantial loss of formulaic or tabular content, image inputs outperform their text counterparts. This highlights the essential role of multimodal reasoning and the irreplaceable value of visual information in addressing specific types of errors.

### 4.3 FINE-GRAINED ANALYSIS

**Capability Dimensions.** We compute pairwise Spearman correlations between error types across two input configurations (text and image) for the eight evaluated MLLMs excluding Qwen2.5-VL-72B, as shown in Figure 4. We derive the following insights:

(i) *With image input, CF exhibits consistently low correlations with other error categories, suggesting that the skills required for mathematical reasoning are relatively distinct.* In contrast, with text input, CF shows moderate correlation with LE, indicating that OCR-flattened formulas lose their structural specificity and are interpreted by models in a manner more akin to natural language. Combined with the overall poor performance on CF tasks, this underscores the unique challenges of this category and the need for targeted improvements.

(ii) *Although DI is also related to experimental settings, it does not exhibit strong correlations with SG, MO, or DHP.* This indicates that DI primarily emphasizes causal framing and variable identifiability, rather than the procedural understanding of experimental operations.

(iii) *OCR severely degrades structured content such as figures and formulas, making questions that depend on multimodal information unanswerable.* This diminishes the expression of multimodal reasoning capabilities and artificially inflates inter-category correlations under text input.

Based on the above analysis, we consolidate the original 9 error categories, each defined by its objective target, into 5 core latent skill dimensions evaluated by ScholScan under the image input setting. While each dimension highlights the primary competence emphasized by its corresponding error types, they are not mutually exclusive, as many questions involve overlapping reasoning abilities.

RQD and DI correspond to research concept comprehension, which requires models to *identify the scope and definition* of research objectives by integrating contextual cues and prior knowledge. SG, MO, and DHP fall under *experimental process modeling*, which tests a model’s ability to reconstruct procedural workflows such as sampling, measurement, and data handling. CF captures *formal reasoning and symbolic computation*, focusing on syntactic parsing and numerical logic. IC evaluates causal inference, where models must *synthesize dispersed causal evidence* to reach sound conclusions. RCA and LE reflect referential alignment and linguistic consistency, which assess the ability to *verify citations and maintain coherent expression* throughout the document.

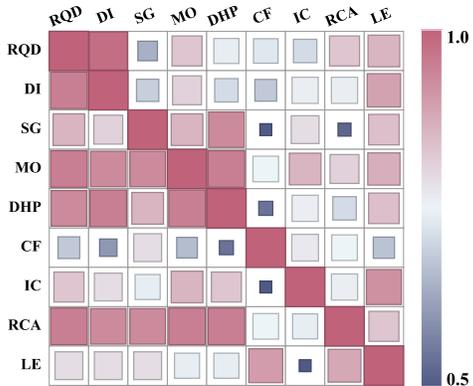


Figure 4: Spearman correlation matrix among the 9 error types.

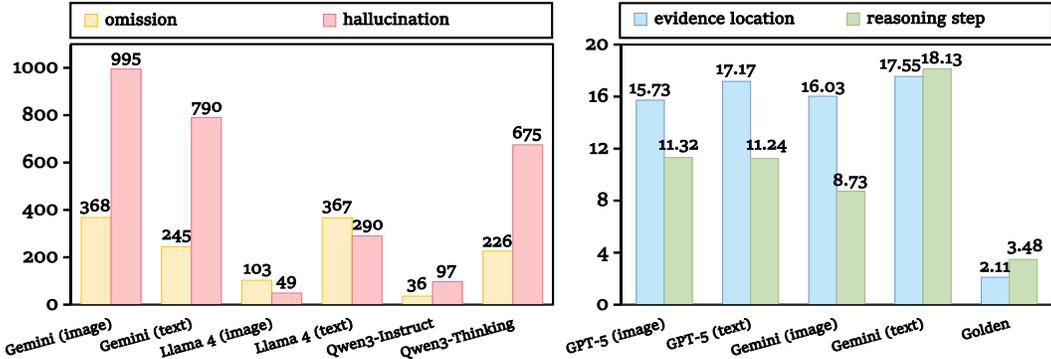


Figure 5: Left: Distribution of omission and hallucination errors. Right: Average reasoning steps and evidence locations involved in the answer generation, compared against the golden reference.

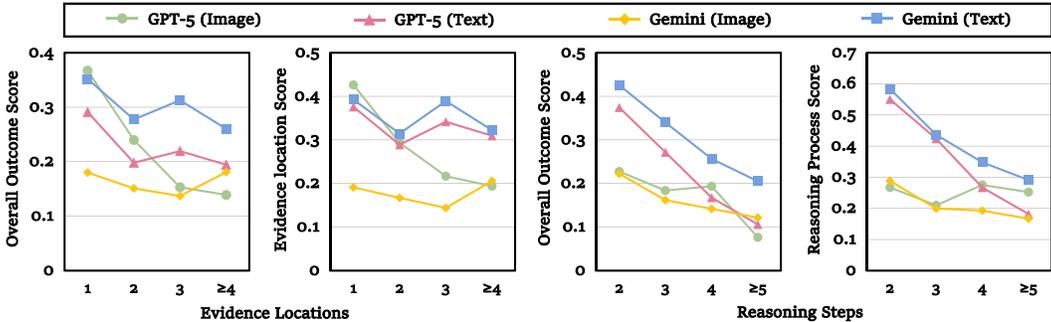


Figure 6: Performance trends across varying reasoning depths and evidence counts.

**Hidden Complexity in Scan-Oriented Tasks.** We analyze the reasoning traces of GPT-5 and Gemini 2.5 Pro under both input configurations, focusing on the number of evidence pieces scanned and the reasoning steps performed. As illustrated in Figure 5, even the most advanced models often scan up to 8 times more evidence and execute 3.5 times more reasoning steps than the reference answers, merely to approximate a correct response, yet they still frequently fail. This highlights the substantial hidden complexity inherent in scan-oriented tasks, which significantly amplifies the challenge of successful task completion.

#### 4.4 ERROR ANALYSIS

**Omission and Hallucination.** Most zero-score cases fall into two categories: either the model fails to detect any errors in the paper, or it becomes overwhelmed by hallucinations and entirely overlooks the actual errors present in the reference answer. We analyze the number of zero-score questions and the proportion of these two failure modes across models, as shown in Figure 5. Stronger models tend to have fewer zero-score cases overall, but are more prone to overconfident hallucinations.

**Fragile Reasoning under Complex Evidence.** Figure 6 shows how top-performing models behave under different numbers of reasoning steps and evidence locations. As reasoning steps increase, both reasoning and overall scores steadily decline, revealing a clear bottleneck in MLLMs’ ability to construct long causal chains. In contrast, variation in evidence count has a weaker and less consistent impact. However, this does not imply that multi-evidence questions pose only marginal difficulty. Since the evaluation metric allows partial evidence omissions, more evidence items do not necessarily incur large score penalties. Still, heavier evidence loads often require longer reasoning chains, which substantially affect the coherence and completeness of inferred logic. These results highlight the persistent challenge for MLLMs in integrating evidence and maintaining logical structure as task complexity grows.

Table 2: Scores of RAG methods across the 9 error types (scaled by 100).

Models	Avg	RQD	DI	SG	MO	DHP	CF	IC	RCA	LE
<i>Text Input (Base Model: Qwen3 Thinking)</i>										
Baseline	17.4	8.9	16.2	31.9	15.1	23.7	5.6	22.3	21.1	2.3
Oracle	24.5	20.6	27.9	43.6	21.3	40.8	7.4	26.9	26.0	1.9
bm25	16.7	9.7	13.7	33.0	17.3	23.8	6.8	25.4	16.5	3.0
BGE-M3	11.3	8.6	7.5	24.8	9.1	15.4	5.3	15.6	11.4	1.0
Contriever-msmacro	16.6	9.7	18.2	33.7	10.7	20.8	6.4	18.5	19.8	1.8
nv-embed-v2	6.8	4.0	4.0	9.4	6.1	4.9	5.5	5.7	10.0	2.0
<i>Image Input (Base Model: Llama4 Maverick)</i>										
Baseline	7.0	7.0	7.3	9.4	4.5	4.0	6.5	6.7	8.8	3.0
Oracle	6.5	3.0	4.5	15.6	8.2	9.4	4.9	10.0	4.4	1.4
ColPali-v1.3	0.8	1.5	0.0	0.5	0.0	0.9	0.5	1.3	1.4	0.0
ColQwen2.5	1.2	2.1	0.7	0.5	0.0	1.2	0.2	2.7	2.0	0.0
VisRAG	1.0	2.0	0.0	1.0	0.0	1.0	1.6	1.3	1.2	0.0
VRAG-RL	10.9	9.8	11.6	17.8	8.2	11.0	6.8	13.1	10.8	8.1

#### 4.5 RAG ANALYSIS

We evaluated 8 RAG methods under both input configurations (Robertson et al., 1994; Chen et al., 2024; Lee et al., 2025; Faysse et al., 2025; Yu et al., 2025; Wang et al., 2025; Izacard et al., 2022). Key findings are presented below, with detailed results shown in Tables 2 and 3.

**Oracle Condition Yields Significant Accuracy Gains.** Providing gold-standard images alleviates the scanning burden in long-context inputs, increasing the chances of generating correct answers. While overall performance improves, gains are limited for CF errors and minimal for LE errors. For CF, sparse formulaic content means gold images offer slight help. For LE, dense text distribution makes even direct access to target regions insufficient to reduce complexity for current models.

**In consistency-centric scan-oriented tasks, most retrieval-based enhancement methods show minimal effectiveness.** All embedding models exhibit poor retrieval accuracy. None achieves recall of 50% within the top-5 retrieved items. More critically, performance deteriorates after retrieval, especially for multimodal embedding models, where post-retrieval responses are almost entirely incorrect and scores approach 0.

**Complex embedding model architectures do not yield better performance.** Providing gold-standard images alleviates the scanning burden in long-context inputs, increasing the chances of retrieving correct answers. While overall performance improves, gains are limited for CF and minimal for LE errors. For CF, sparse formulaic content means gold images offer only slight localization help. For LE, dense error distribution makes even direct access to target regions insufficient to reduce task complexity for current models.

**Reinforcement learning frameworks with a visual-centric focus have distinguished themselves as leading approaches.** Despite being built on a compact 7B model, VRAG-RL consistently delivers improved performance and is the only method that achieves gains in the image-input setting following RL optimization. Its enhanced retrieval sharpens evidence selection, while strong reasoning provides effective guidance during document scanning. The retrieval and reasoning components are interleaved in design, with each stage informing the other in an iterative loop. This tightly coupled interaction contributes to the method’s superior performance potential.

## 5 CONCLUSION

In this paper, we introduce ScholScan, a benchmark designed to evaluate the performance of MLLMs on scan-oriented tasks that require detecting scientific errors across entire academic pa-

Table 3: Summary of retrieval performance for RAG methods.

Models	MRR@5	Recall@5
<i>Text Input (Base Model: Qwen3 Thinking)</i>		
bm25	0.41	0.48
BGE-M3	0.16	0.21
Contriever-msmacro	0.31	0.39
nv-embed-v2	0.30	0.38
<i>Image Input (Base Model: Llama4 Maverick)</i>		
ColPali-v1.3	0.26	0.31
ColQwen2.5	0.30	0.35
VisRAG	0.41	0.46

pers. We conduct a comprehensive evaluation and in-depth analysis of mainstream MLLMs and RAG methods. The results demonstrate that current MLLMs remain far from capable of reliably addressing such tasks, and that existing RAG approaches provide little to no improvement. This highlights the complexity, integrative demands, and originality of the ScholScan benchmark. Looking ahead, we aim to develop scan-oriented task paradigms suited to diverse academic scenarios and explore new techniques for enhancing model performance on target-suppressed inputs. These directions support the broader goal of advancing MLLMs from passive assistants to active participants in scientific research.

## 6 ETHICS STATEMENT

All data used in this paper were constructed by the authors and do not include any external public or proprietary datasets. The included academic papers and author names are publicly available through arXiv and OpenReview and can be freely accessed.

A team of 10 domain experts was assembled to comprehensively review all task instances initially generated by Gemini 2.5 Pro. All annotators gave informed consent to participate. To ensure the accuracy and neutrality of both model-generated and human-verified content, we employed a rigorous multi-stage validation process involving cross-review and third-party adjudication.

Evaluation across 15 mainstream models and 24 input configurations was conducted via legally authorized API access through the VolcEngine, Alibaba Cloud’s LLM services, and OpenRouter.

ScholScan is fully open-sourced and freely available for academic and non-commercial research purposes. We provide the complete download link and documentation through an anonymous GitHub repository. All personally identifiable information has been removed from the dataset, and its collection and release comply with the ethical and legal requirements in place at the time of data acquisition.

## REFERENCES

- Anthropic. System card: Claude opus 4 & claude sonnet 4. <https://www.anthropic.com/claude-4-system-card>, May 2025. Updated Sep 2, 2025.
- S. Auer, Dante Augusto Couto Barone, Cassiano Bartz, E. Cortes, Mohamad Yaser Jaradeh, Oliver Karras, Manolis Koubarakis, Dmitry I. Mouromtsev, Dmitrii Pliukhin, Daniil Radyush, Ivan Shilin, Markus Stocker, and Eleni Tsalapati. The sciq scientific question answering benchmark for scholarly knowledge. *Scientific Reports*, 13, 2023. URL <https://api.semanticscholar.org/CorpusID:258507546>.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao 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, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL <https://arxiv.org/abs/2502.13923>.
- ByteDance Seed Team. Introduction to techniques used in seed1.6. <https://seed.bytedance.com/en/blog/introduction-to-techniques-used-in-seed1-6>, June 2025. Official blog post describing Seed1.6 techniques.
- Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024. URL <https://arxiv.org/abs/2402.03216>.
- Zhiyu Chen, Wenhui Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. FinQA: A dataset of numerical reasoning over financial data. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3697–3711, Online and Punta Cana, Dominican

- Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.300. URL <https://aclanthology.org/2021.emnlp-main.300/>.
- Chao Deng, Jiale Yuan, Pi Bu, Peijie Wang, Zhong-Zhi Li, Jian Xu, Xiao-Hui Li, Yuan Gao, Jun Song, Bo Zheng, and Cheng-Lin Liu. LongDocURL: a comprehensive multimodal long document benchmark integrating understanding, reasoning, and locating. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1135–1159, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.57. URL <https://aclanthology.org/2025.acl-long.57/>.
- DeepSeek-AI et al. Deepseek-v3 technical report, 2025a. URL <https://arxiv.org/abs/2412.19437>.
- Gheorghe Comanici et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025b. URL <https://arxiv.org/abs/2507.06261>.
- Long Phan et al. Humanity’s last exam, 2025c. URL <https://arxiv.org/abs/2501.14249>.
- OpenAI et al. gpt-oss-120b & gpt-oss-20b model card, 2025d. URL <https://arxiv.org/abs/2508.10925>.
- Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, CELINE HUDELLOT, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=ogjBpZ8uSi>.
- Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *ArXiv*, abs/2312.10997, 2023. URL <https://api.semanticscholar.org/CorpusID:266359151>.
- Yingqiang Ge, Wenyue Hua, Kai Mei, Jianchao Ji, Juntao Tan, Shuyuan Xu, Zelong Li, and Yongfeng Zhang. Openagi: When llm meets domain experts. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 5539–5568. Curran Associates, Inc., 2023. URL [https://proceedings.neurips.cc/paper\\_files/paper/2023/file/1190733f217404edc8a7f4e15a57f301-Paper-Datasets\\_and\\_Benchmarks.pdf](https://proceedings.neurips.cc/paper_files/paper/2023/file/1190733f217404edc8a7f4e15a57f301-Paper-Datasets_and_Benchmarks.pdf).
- Daming Guo, Dongdong Yang, Hongyi Zhang, et al. Deepseek-rl incentivizes reasoning in llms through reinforcement learning. *Nature*, 645:633–638, 2025. doi: 10.1038/s41586-025-09422-z.
- Yichen He, Guanhua Huang, Peiyuan Feng, Yuan Lin, Yuchen Zhang, Hang Li, and Weinan E. Pasa: An llm agent for comprehensive academic paper search, 2025. URL <https://arxiv.org/abs/2501.10120>.
- Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning, 2022. URL <https://arxiv.org/abs/2112.09118>.
- Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. NV-embed: Improved techniques for training LLMs as generalist embedding models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=lgsyLSsDRe>.
- Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal ArXiv: A dataset for improving scientific comprehension of large vision-language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14369–14387,

- Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.775. URL <https://aclanthology.org/2024.acl-long.775/>.
- Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL <https://llava-vl.github.io/blog/2024-01-30-llava-next/>.
- Renze Lou, Hanzi Xu, Sijia Wang, Jiangshu Du, Ryo Kamoi, Xiaoxin Lu, Jian Xie, Yuxuan Sun, Yusen Zhang, Jihyun Janice Ahn, Hongchao Fang, Zhuoyang Zou, Wenchao Ma, Xi Li, Kai Zhang, Congying Xia, Lifu Huang, and Wenpeng Yin. AAAR-1.0: Assessing AI’s potential to assist research. In *Forty-second International Conference on Machine Learning*, 2025. URL <https://openreview.net/forum?id=RHAWcjIy12>.
- 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. MMLONGBENCH-DOC: Benchmarking long-context document understanding with visualizations. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL <https://openreview.net/forum?id=loJMLacwzf>.
- Meta. Llama 4 | model cards and prompt formats. <https://www.llama.com/docs/model-cards-and-prompt-formats/llama4/>, 2025. Official model card and prompt format documentation.
- Meredith Ringel Morris, Jascha Sohl-Dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet, and Shane Legg. Position: Levels of AGI for operationalizing progress on the path to AGI. In *Forty-first International Conference on Machine Learning*, 2024. URL <https://openreview.net/forum?id=0ofzEysK2D>.
- OpenAI. Gpt-5 system card. Technical report, OpenAI, August 2025. URL <https://cdn.openai.com/gpt-5-system-card.pdf>. Version: August 13, 2025.
- OpenAI. gpt-4.1 — openai api documentation, 2025. URL <https://platform.openai.com/docs/models/gpt-4.1>. Accessed: 2025-09-25.
- Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. SPIQA: A dataset for multimodal question answering on scientific papers. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL <https://openreview.net/forum?id=h3lddsY5nf>.
- Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. Okapi at trec-3. In *Text Retrieval Conference*, 1994. URL <https://api.semanticscholar.org/CorpusID:41563977>.
- R. Smith. An overview of the tesseract ocr engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pp. 629–633, 2007. doi: 10.1109/ICDAR.2007.4376991.
- Rubèn Tito, Minesh Mathew, C. V. Jawahar, Ernest Valveny, and Dimosthenis Karatzas. ICDAR 2021 competition on document visualquestion answering. *CoRR*, abs/2111.05547, 2021. URL <https://arxiv.org/abs/2111.05547>.
- Qiuchen Wang, Ruixue Ding, Yu Zeng, Zehui Chen, Lin Chen, Shihang Wang, Pengjun Xie, Fei Huang, and Feng Zhao. Vrag-rl: Empower vision-perception-based rag for visually rich information understanding via iterative reasoning with reinforcement learning, 2025. URL <https://arxiv.org/abs/2505.22019>.
- Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. Charxiv: Charting gaps in realistic chart understanding in multimodal llms, 2024. URL <https://arxiv.org/abs/2406.18521>.
- xAI. Grok 4 fast model card. Technical report, xAI, September 2025. URL <https://data.x.ai/2025-09-19-grok-4-fast-model-card.pdf>. Last updated: September 19, 2025.

- Dawei Yan, Yang Li, Qing-Guo Chen, Weihua Luo, Peng Wang, Haokui Zhang, and Chunhua Shen. Mmcr: Advancing visual language model in multimodal multi-turn contextual reasoning, 2025. URL <https://arxiv.org/abs/2503.18533>.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chuji Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL <https://arxiv.org/abs/2505.09388>.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *CoRR*, abs/1809.09600, 2018. URL <http://arxiv.org/abs/1809.09600>.
- Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and Maosong Sun. VisRAG: Vision-based retrieval-augmented generation on multi-modality documents. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=zG459X3Xge>.
- Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhui Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9556–9567, 2024. doi: 10.1109/CVPR52733.2024.00913.
- Yilun Zhao, Yitao Long, Hongjun Liu, Ryo Kamoi, Linyong Nan, Lyuhao Chen, Yixin Liu, Xiangru Tang, Rui Zhang, and Arman Cohan. DocMath-eval: Evaluating math reasoning capabilities of LLMs in understanding long and specialized documents. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16103–16120, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.852. URL <https://aclanthology.org/2024.acl-long.852/>.
- Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibe Yang. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 5168–5191. Curran Associates, Inc., 2023. URL [https://proceedings.neurips.cc/paper\\_files/paper/2023/file/108030643e640ac050e0ed5e6aace48f-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2023/file/108030643e640ac050e0ed5e6aace48f-Paper-Conference.pdf).
- Ruiyang Zhou, Lu Chen, and Kai Yu. Is LLM a reliable reviewer? a comprehensive evaluation of LLM on automatic paper reviewing tasks. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 9340–9351, Torino, Italia, May 2024. ELRA and ICCL. URL <https://aclanthology.org/2024.lrec-main.816/>.

## Table of Contents in Appendix

<b>A Prompts</b>	<b>16</b>
A.1 Within-Generate Prompt . . . . .	16
A.2 Within-Sample Prompt . . . . .	19
A.3 Extractor Prompt . . . . .	21
A.4 System Prompt . . . . .	23
<b>B Examples from Existing Datasets</b>	<b>24</b>
B.1 Example from DocMath-Eval . . . . .	24
B.2 Example from SlideVQA . . . . .	25
B.3 Example from MMLongBench-Doc . . . . .	26
B.4 Example from LongDocURL . . . . .	27
B.5 Example from ArXivQA . . . . .	28
B.6 Example from Charxiv . . . . .	29
B.7 Example from AAAR . . . . .	30
B.8 Example from MMCR . . . . .	31
B.9 Example from DocVQA . . . . .	32
B.10 Example from SPIQA . . . . .	33
<b>C Dataset Annotation and Construction</b>	<b>34</b>
C.1 Human Annotator Guidelines . . . . .	34
C.2 Annotation Statistics . . . . .	34
C.3 Examples of Annotation . . . . .	35
C.3.1 Case 1: Discard Directly . . . . .	35
C.3.2 Case 2: Modify Question . . . . .	36
C.3.3 Case 3: Modify Explanation . . . . .	37
<b>D Common Failure Cases of MLLMs</b>	<b>38</b>
D.1 RQD (Research Question & Definitions) . . . . .	38
D.2 DI (Design & Identifiability) . . . . .	39
D.3 SG (Sampling & Generalizability) . . . . .	40
D.4 RCA (Referential and Citation Alignment) . . . . .	42
D.5 MO (Measurement & Operationalization) . . . . .	43
D.6 DHP (Data Handling & Preprocessing) . . . . .	45
D.7 CF (Computation & Formulae) . . . . .	47
D.8 IC (Inference & Conclusions) . . . . .	48
D.9 LE (Language & Expression) . . . . .	49
<b>E Human-Machine Consistency Evaluation</b>	<b>50</b>

**F Hyperparameter Sensitivity Analysis**

**51**

## A PROMPTS

### A.1 WITHIN-GENERATE PROMPT

#### Within-Generate Prompt

You will receive a high-quality, already accepted scientific paper as a PDF. Working only with the PDF itself (and any appendix embedded in the same PDF), edit specific textual spans to inject one or more errors chosen only from the taxonomy below, such that the errors are hard yet clearly identifiable by a professional reviewer reading the PDF alone.

Error Type (fixed):

Research Question & Definitions

Definition: The core construct/hypothesis/variable is insufficiently or inconsistently defined (conceptual vs operational), leaving the estimand ambiguous.

Design & Identifiability

Definition: Given a clear estimand, the design violates structural identification conditions so the effect is not identifiable even with infinite data and perfect measurement.

Sampling & Generalizability

Definition: The sampling frame/process/composition or cluster/power setup does not support valid or stable sample→population claims.

Measurement & Operationalization

Definition: Measures/manipulations lack feasibility/reliability/validity/timing, so observed variables systematically diverge from the intended construct/treatment.

Data Handling & Preprocessing

Definition: Pipeline choices in missing handling, joins/keys, temporal splitting, feature construction, or partitioning introduce bias (incl. leakage or unit/scale conflicts).

Computation & Formulae

Definition: Arithmetic/algebra/notation errors (totals/ratios, unit conversion, CI vs point estimate, p-value vs label, symbol reuse, undefined variables, dimension mismatch).

Inference & Conclusions

Definition: Interpretations or causal statements exceed what methods/data support, or contradict the shown statistics/tables/captions.

Referential and Citation Alignment

Definition: Contradictions about the same quantity/term across text, tables, captions, or appendix within the paper.

Language & Expression

Definition: Terminology/capitalization/grammar ambiguities that affect meaning or domain-critical term consistency (not cosmetic typos).

## Within-Generate Prompt (Continued)

Global constraints (must comply)

1. Each error must map to exactly one primary category in the taxonomy. Do not mix causes.
2. Each error must involve more than 2 micro-edits (each edit  $\leq$  20 English words) spread across distinct pages or paragraphs.
3. If an edit would create an immediate contradiction in the same sentence/paragraph/caption, you may add shadow patch(es) for the same error to keep the text natural (still counted as edit locations).
4. Independence across errors (per-copy generation)
 

Generate each error on a separate copy of the original PDF . Different errors must be logically and operationally independent:

No progression or variant relations: an error must not be a stricter/looser version, superset/subset, or minor wording variant of another error.

No anchor reuse: do not target the same sentence/caption/table cell or reuse the same old\_str (or a near-duplicate paraphrase) across different errors.

Applying any single error in isolation to the original PDF must still yield a detectable, clearly categorizable error according to the taxonomy.
5. Every error must be supportable using text inside the PDF. Do not rely on external supplementary files or prior knowledge.
6. Design as difficult as possible but clean errors. Prefer edits that force cross-checking between two spots (e.g., Methods vs Results). Avoid trivialities. Edits must remain locally plausible and not advertise themselves via obviously artificial phrases (e.g., avoid contrived tokens purely added to be detectable).
7. ``No cosmetic issues`` applies except for I (Language & Expression). For I, edits must affect meaning or domain-critical terminology (e.g., ambiguous phrasing, inconsistent technical terms). Pure typos, punctuation tweaks, or layout nits are not allowed.
8. Do not edit titles, author lists, bibliography entries, equation numbering, figure images, or add new figures/tables/references.
9. Frame each question as a neutral imperative that asks for a decision about a specific condition, using (but not limited to) Decide/Determine/Judge/Evaluate/Assess whether.... Do not presuppose an outcome or use suggestive intensifiers (e.g., clearly/obviously/likely/suspicious as examples).

## Within-Generate Prompt (Continued)

```
10. Output English-only and strictly follow the JSON schema
    below. Do not include any additional text outside the
    JSON:
[
  {
    "id": "1-based integer as string",
    "modify": [
      {
        "location": "Page number + short unique nearby quote (
          ≤15 tokens).",
        "old_str": "Exact original text from the PDF (verbatim)
          .",
        "new_str": "Edited text after your change."
      }
      /* Add 1-2 more locations; each location ≤ 20 words
        changed.
        Shadow patches for local coherence count as locations.
        */
    ],
    "question": "One neutral audit-style task (1-25 words).",
    "explanation": "Explain in 2-4 sentences why a reviewer can
      detect this error from the edited PDF alone.",
    "Type": "Name the primary category (e.g., Inference &
      Conclusions).",
  }
  /* More Errors */
]
```

## A.2 WITHIN-SAMPLE PROMPT

### Within-Sample Prompt

You will receive a paper PDF and the weaknesses mentioned in its peer-review comments. Your task is, based only on the content of that PDF, to sample from the review comments and verify possible errors related to the categories below, and for each confirmed or highly plausible error, generate one question and one explanation.

Error Type (fixed):

Research Question & Definitions

Definition: The core construct/hypothesis/variable is insufficiently or inconsistently defined (conceptual vs operational), leaving the estimand ambiguous.

Design & Identifiability

Definition: Given a clear estimand, the design violates structural identification conditions so the effect is not identifiable even with infinite data and perfect measurement.

Sampling & Generalizability

Definition: The sampling frame/process/composition or cluster/power setup does not support valid or stable sample→population claims.

Measurement & Operationalization

Definition: Measures/manipulations lack feasibility/reliability/validity/timing, so observed variables systematically diverge from the intended construct/treatment.

Data Handling & Preprocessing

Definition: Pipeline choices in missing handling, joins/keys, temporal splitting, feature construction, or partitioning introduce bias (incl. leakage or unit/scale conflicts).

Computation & Formulae

Definition: Arithmetic/algebra/notation errors (totals/ratios, unit conversion, CI vs point estimate, p-value vs label, symbol reuse, undefined variables, dimension mismatch).

Inference & Conclusions

Definition: Interpretations or causal statements exceed what methods/data support, or contradict the shown statistics/tables/captions.

Referential and Citation Alignment;

Definition: Contradictions about the same quantity/term across text, tables, captions, or appendix within the paper.

Language & Expression

Definition: Terminology/capitalization/grammar ambiguities that affect meaning or domain-critical term consistency (not cosmetic typos).

## Within-Sample Prompt (Continued)

Global constraints (must comply)  
 Output only the specified categories; even if other error types appear in the reviews, do not output them.  
 Sample first, then verify: extract candidates from the review comments, then confirm them in the PDF. If you cannot locate supporting anchors in the PDF (page number plus phrase/label), do not output that candidate.  
 Questions must be neutral and non-leading: use an "audit task + decision" style, avoiding yes/no bias.  
 Independence: each question must target a different figure or different textual anchor; no minor variants of the same issue.  
 Evidence first: the explanation must cite locatable anchors in the PDF (page number + original phrase/caption). You may mention a key short phrase from the review as a clue, but write the question and explanation in your own words  
 Language & format: both question and explanation must be in English; output JSON only, with no extra text.  
 Quantity: sort by evidence strength and output up to 5 items; if none qualify, output an empty array [].

Example output

```
[
  {
    "id": "1",
    "question": "Audit y-axis baselines and possible axis breaks in Figure 2; decide presence/absence and cite evidence.",
    "explanation": "The review flags possible exaggeration in Fig.2. In the PDF (p.6, caption 'Performance vs baseline'), the y-axis starts at 0.85 with a break, magnifying small differences; panels use different ranges."
    "Type": "Visualization & Presentation Bias"
  }
]
```

## A.3 EXTRACTOR PROMPT

## Extractor Prompt

You will receive three inputs:

Q: the open-ended question;

E: the gold explanation (describes exactly one error; extra details still belong to the same single error);

A: the model's answer to be evaluated.

Your job is to extract counts only and output a single JSON object with the exact schema below. Do not compute any scores. Do not add fields.

Core selection rule (multiple errors in A)

1. Parse E into a single gold error (the "target error").
2. From A, identify how many distinct error claims are made. Cluster together mentions that support the same error (multiple locations for one error are still one error).
3. Existence decision (binary correctness only):

Let the gold existence be 1 if E asserts an error exists, else 0.

Let the predicted existence be 1 if A asserts any error, else 0 (e.g., states no error).

Set existence = 1 if predicted existence equals gold existence; otherwise set existence = 0.

4. If existence = 0: set contains\_target\_error = 0; set all location and reasoning counts to 0; and set unrelated\_errors to the total number of distinct error claims in A. Then output the JSON.

5. If existence = 1:

If the gold existence is 1: determine whether A contains the target error (match by the main error idea in E: category/intent/scope; treat E's subpoints as the same error).

If yes, set contains\_target\_error = 1 and compute location and reasoning only for the target error. Count all other error claims in A as unrelated\_errors.

If no, set contains\_target\_error = 0; set all location and reasoning counts to 0; set unrelated\_errors to the total number of distinct error claims in A.

If the gold existence is 0: set contains\_target\_error = 0; set all location and reasoning counts to 0; set unrelated\_errors to the total number of distinct error claims in A. (These negative items are for binary accuracy only; they are not used for detailed scoring.)

Matching guidance (A error ↔ target error): match by the main error idea in E (category/intent/scope), not by wording. Treat E's subpoints as part of the same single error. Prefer the best-matching cluster in A; if ties, choose the one with stronger alignment to E's core claim.

## Extractor Prompt (Continued)

```

Counting rules
Location (for the target error only when existence=1 and
  contains_target_error=1):
gold_steps: number of unique error locations described in E (
  after normalization and deduplication).
hit_steps: number of predicted locations in A that match any
  gold location for the target error.
extra_steps: number of predicted locations in A for the
  target error that do not match any gold location.

Reasoning (for the target error only when existence=1 and
  contains_target_error=1):
Convert E into a canonical set or ordered chain of reasoning
  steps for the target error.
gold_steps: total number of such steps.
reached_steps:
  single-chain tasks: length of the longest valid prefix of
    A along the gold chain;
  multi-path/parallel tasks: size of the intersection
    between A's steps and the gold step set (or the
    maximum across gold paths if multiple are defined).
missing_steps: gold_steps - reached_steps (non-negative
  integer).
Unrelated errors:
unrelated_errors: number of distinct error claims in A that
  are not the target error (0 if none).
Output schema (return exactly this JSON; integers only)
{
  "existence": 0,
  "contains_target_error": 0,
  "location": {
    "gold_steps": 0,
    "hit_steps": 0,
    "extra_steps": 0
  },
  "reasoning": {
    "gold_steps": 0,
    "reached_steps": 0,
    "missing_steps": 0
  },
  "unrelated_errors": 0
}

```

#### A.4 SYSTEM PROMPT

##### System Prompt

You are a neutral, careful academic reviewer. You will receive an open-ended question and the paper content. The paper may or may not have issues related to the question. Do not assume there are errors. If the question is about citations, you will be given a citing paper and a cited paper; evaluate only the citing paper for possible issues and use the cited paper only as the reference for comparison. Write in natural prose with no fixed template

##### Rules:

- Speak only when sure. State an error only if you are confident it is a real error (not a mere weakness).
- Stay on scope. Discuss only what the question asks about.
- Evidence completeness. For every error you state, list all distinct evidence cues you are confident about from the PDF. Include plain identifiers (figure/table/section/equation/citation) or quotes. Avoid redundant repeats of the exact same instance; include all distinct locations needed to support the error.
- Be clear and brief. Use short, direct sentences.
- No metaphors. No fancy wording. No guesses or outside sources. Do not invent figures, tables, equations, citations, or results.
- Report as many distinct, well-supported errors as you can within scope. If none are clear, write exactly: "No clear issue relevant to the question." and nothing else.

## B EXAMPLES FROM EXISTING DATASETS

### B.1 EXAMPLE FROM DOCMATH-EVAL

#### One Example from DocMath-Eval

**Question\_ID:** complong-testmini-30

**Question:** What is the percentage of total offering cost on the total amount raised in the IPO if the total offering cost is \$14,528,328 and each unit sold is \$10?

**Context Modalities: Text Documents**

1. Offering costs consist of legal, accounting and other costs incurred through the balance sheet date that are directly related to the Initial Public Offering. Offering costs amounting to \$14,528,328 were charged to shareholders' equity upon the completion of the Initial Public Offering.
2. Pursuant to the Initial Public Offering on July 20, 2020, the Company sold 25,300,000 Units, which includes the full exercise by the underwriter of its option to purchase an additional 3,300,000 Units, at a purchase price of \$10.00 per Unit. Each Unit consists of one Class A ordinary share and one-half of one redeemable warrant ("Public Warrant"). Each whole Public Warrant entitles the holder to purchase one Class A ordinary share at an exercise price of \$11.50 per whole share (see Note 7).

**Covered areas:**

**Mathematics Only**

**Cross-evidence Reasoning:**

**Limited**

**Task Paradigm:**

**Search-oriented**

B.2 EXAMPLE FROM SLIDEVQA

**One Example from SlideVQA**

**Question ID:** 1  
**Question:** How much difference in INR is there between the average order value of CY2013 and that of CY2012?

**Context Modalities: Multi-Modal Documents and Texts**

**Executive summary**  
Key findings

**Increasing average order value**

Year	Value (INR)	Growth Rate
CY2012	1,080	-
CY2013	1,860	67%
CY2016P	3,600	25% CAGR

**Fashion + Footwear + Accessories GMV**

Year	Value (M)	Growth Rate
CY2012	278	-
CY2013	559	100%
CY2016P	2,811	71% CAGR

*Accel estimates and Industry sources*

KEY FINDING

ACCEL

Average order values climbing up rapidly

1. Last year there was a significant jump in average order value as there was a penetration of new categories like jewellery, home décor etc.
2. Also, users are becoming more comfortable buying higher priced items online.

Fashion category doubled last year

1. Last year was the rise of the fashion category – fashion e-commerce GMV doubled since 2012.
2. Given the young demographic which is shopping for latest looks online and increasing choice online – we estimate that this category will see 400% growth in the next 3 years and rival electronics and mobile category in GMV.

**Covered areas:** Limited

**Cross-evidence Reasoning:** None

**Task Paradigm:** Search-oriented

B.3 EXAMPLE FROM MMLONGBENCH-DOC

One Example from MMLongBench-Doc

**Doc ID:** afe620b9beac86c1027b96d31d396407.pdf  
**Question:** How much higher was the proposed dividend paid (Rupees in lacs) in 2002 compared to 2001?

**Context Modalities: Multi-Modal Documents and Texts**



**SHAREHOLDER REFERENCER**

**Unclaimed Dividend**

Unclaimed dividend for the years prior to and including the financial year 1998-99 has been transferred to the General Revenue Account of the Central Government / the Investor Education and Protection Fund established by the Central Government (IEPF), as applicable.

Shareholders who have not encashed their dividend warrants relating to financial year(s) up to and including 1993-94 may claim such dividend (transferred to the General Revenue Account) from the Registrar of Companies, West Bengal, Government of India, Neam Palace, # 1850 Building, 2nd Floor, 23/44 A.L.I.C. Bose Road, Kolkata 700 020, in the prescribed form. This form can be furnished by the Investor Service Centre of the Company (ISC) on request or can be downloaded from the Company's corporate website www.itcportal.com under the section 'Investor Relations'.

The dividend for the undemanded years, if unclaimed for 7 years, will be transferred by the Company to IEPF in accordance with the schedule given below. Attention is drawn that the unclaimed dividend for the financial year 1999-2000 will be due for transfer to IEPF later this year. Communication has been sent by the Company to the concerned Shareholders advising them to lodge their claims with respect to unclaimed dividend.

Once unclaimed dividend is transferred to IEPF, no claim shall lie in respect thereof.

**ITC Limited**

Financial Year	Dividend Identification No.	Date of Declaration of Dividend	Total Dividend (Rs.)	Unclaimed Dividend as on 31/03/2007		Due for transfer to IEPF on
				(Rs.)	%	
1999-00	70th	28th July, 2000	1,84,06,11,780.00	1,26,32,297.00	0.69	18th September, 2007*
2000-01	71st	30th August, 2001	2,45,41,49,040.00	2,04,42,133.91	0.84	18th September, 2008
2001-02	72nd	28th July, 2002	3,36,14,27,743.00	2,56,63,749.00	0.77	31st August, 2009
2002-03	73rd	25th July, 2003	3,71,26,78,200.00	2,38,48,718.00	0.64	30th August, 2010
2003-04	74th	20th July, 2004	4,05,25,77,000.00	3,35,88,520.00	0.83	4th September, 2011
2004-05	75th	28th July, 2005	7,73,24,48,386.00	6,07,53,301.00	0.78	3rd September, 2012
2005-06	76th	21st July, 2006	9,95,12,91,267.00	7,38,87,332.00	0.74	28th August, 2013

\* It will not be possible to entertain claims received by ISC after 14th September, 2007.

**Erstwhile ITC Hotels Limited**

Financial Year	Date of Declaration of Dividend	Total Dividend (Rs.)	Unclaimed Dividend as on 31/03/2007		Due for transfer to IEPF on
			(Rs.)	%	
1999-00	25th August, 2000	3,02,16,492.00	3,13,648.00	1.04	10th October, 2007*
2000-01	17th August, 2001	3,02,16,492.00	3,04,562.00	1.01	21st September, 2008
2003-04	14th July, 2004	6,04,32,984.00	6,99,704.00	1.16	18th August, 2011

\* It will not be possible to entertain claims received by ISC after 9th October, 2007.

**Bank Details**

Shareholders holding Shares in the physical form are requested to notify / send the following to ISC to facilitate better servicing:-

- i) any change in their address / mandate / bank details, and
- ii) particulars of the bank account in which they wish their dividend to be credited, in case the same have not been furnished earlier.

Shareholders are advised that respective bank details and addresses as furnished by them or by NSDL / CDSL to the Company, for Shares held in the physical form and in the dematerialized form respectively, will be printed on dividend warrants as a measure of protection against fraudulent encashment.

30

**Covered areas:**

Limited (7 Areas)

**Cross-evidence Reasoning:**

Limited

**Task Paradigm:**

Search-oriented

B.4 EXAMPLE FROM LONGDOCURL

One Example from LongDocURL

**Question\_ID:** free\_gemini15\_pro\_4061601\_47\_71\_8  
**Question:** What was the total fair value of options that vested in 2016, 2015, and 2014, in millions of Canadian dollars?

**Context Modalities: Multi-Modal Documents and Texts**

The following table summarizes additional stock option information:

Year ended December 31 (millions of Canadian \$, unless otherwise noted)	2016	2015	2014
Total intrinsic value of options exercised	31	10	21
Fair value of options that have vested	126	91	95
<b>Total options vested</b>	<b>2.1 million</b>	<b>2.0 million</b>	<b>1.7 million</b>

As at December 31, 2016, the aggregate intrinsic value of the total options exercisable was \$86 million and the total intrinsic value of options outstanding was \$130 million.

**21. PREFERRED SHARES**  
 In March 2014, TCPL redeemed all of the 4 million outstanding Series Y preferred shares at a redemption price of \$50 per share for a gross payment of \$200 million.

**22. OTHER COMPREHENSIVE (LOSS)/INCOME AND ACCUMULATED OTHER COMPREHENSIVE LOSS**  
 Components of Other comprehensive (loss)/income, including the portion attributable to non-controlling interests and related tax effects, are as follows:

Year ended December 31, 2016 (millions of Canadian \$)	Before Tax Amount	Income Tax Recovery/(Expense)	Net of Tax Amount
Foreign currency translation gains on net investment in foreign operations	3	—	3
Change in fair value of net investment hedges	(14)	4	(10)
Change in fair value of cash flow hedges	44	(14)	30
Reclassification to net income of gains and losses on cash flow hedges	71	(29)	42
Unrealized actuarial gains and losses on pension and other post-retirement benefit plans	(30)	12	(18)
Reclassification to net income of actuarial loss on pension and other post-retirement benefit plans	22	(9)	13
Other comprehensive loss on equity investments	(117)	20	(97)
<b>Other Comprehensive Loss</b>	<b>(29)</b>	<b>(9)</b>	<b>(38)</b>

Year ended December 31, 2015 (millions of Canadian \$)	Before Tax Amount	Income Tax Recovery/(Expense)	Net of Tax Amount
Foreign currency translation gains on net investment in foreign operations	798	15	813
Change in fair value of net investment hedges	(505)	133	(372)
Change in fair value of cash flow hedges	(82)	35	(47)
Reclassification to net income of gains and losses on cash flow hedges	144	(56)	88
Unrealized actuarial gains and losses on pension and other post-retirement benefit plans	74	(23)	51
Reclassification to net income of actuarial loss and prior service costs on pension and other post-retirement benefit plans	41	(9)	32
Other comprehensive income on equity investments	62	(15)	47
<b>Other Comprehensive Income</b>	<b>522</b>	<b>90</b>	<b>612</b>

155 TCPL, Consolidated financial statements 2016

Covered areas:

Limited

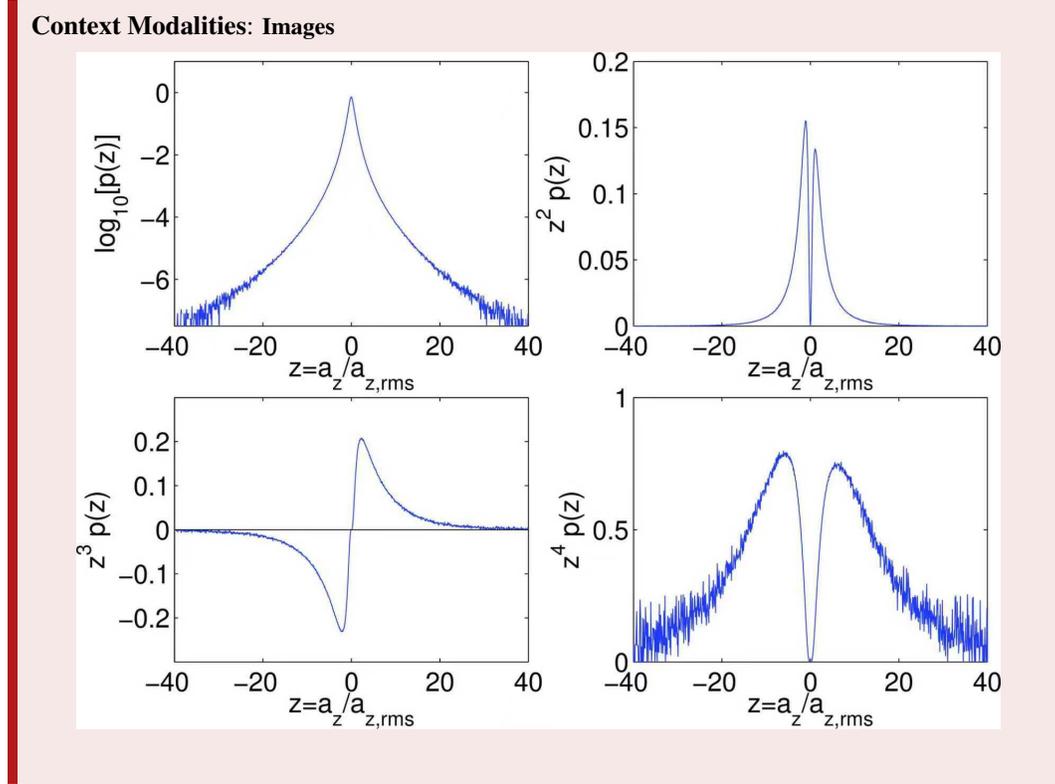
Task Paradigm: search

Search-oriented

B.5 EXAMPLE FROM ARXIVQA

**One Example from ArXivQA**

**Question ID:** physics-8049  
**Question:** Based on the top-right graph, how would you describe the behavior of  $P(z)$  as  $z$  approaches zero?



**Covered areas:** Limited

**Cross-evidence Reasoning:** None

**Task Paradigm:** Search-oriented

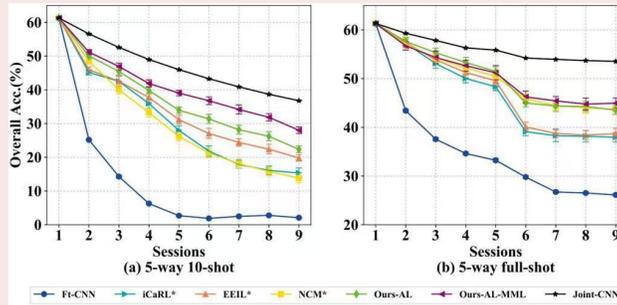
B.6 EXAMPLE FROM CHARXIV

One Example from Charxiv

**Question\_ID:** 2004.10956

**Question:** Which model shows a greater decline in accuracy from Session 1 to Session 9 in the 5-way full-shot scenario?

**Context Modalities:** Images



**Covered areas:**

**Limited (8 Areas)**

**Cross-evidence Reasoning:**

**None**

**Task Paradigm:**

**Search-oriented**

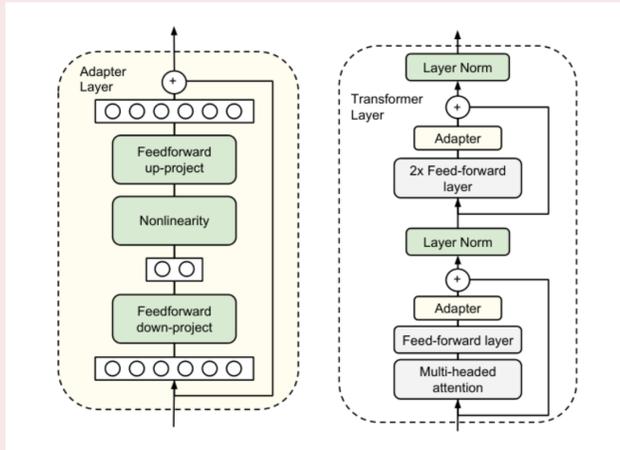
B.7 EXAMPLE FROM AAAR

One Example from AAAR

**Question ID:** 1902.00751

**Question:** What experiments do you suggest doing? Why do you suggest these experiments?

**Context Modalities:** Multi-Modal Documents



**Covered areas:**

**Limited**

**Task Paradigm:**

**Search-oriented**

B.8 EXAMPLE FROM MMCR

**One Example from MMCR**

**Question\_ID:** 1  
**Question:** Which module’s weights are frozen?

**Context Modalities: Multi-Modal Documents and Texts**

**Re-mine, Learn and Reason: Exploring the Cross-modal Semantic Correlations for Language-guided HOI detection**

Yichao Cao Southeast University caoyichao@seu.edu.cn	Qingfei Tang Nanjing Erbo Tech. qingfeitang@gmail.com	Feng Yang Southeast University yangfeng@seu.edu.cn
Xiu Su* University of Sydney xiu55992@uni.sydney.edu.au	Shan You SenseTime youshan@sensetime.com	Xiaobo Lu Southeast University xblu@seu.edu.cn
Chang Xu University of Sydney c.xu@sydney.edu.au		

**Abstract**

*Human-Object Interaction (HOI) detection is a challenging computer vision task that requires visual models to address the complex interactive relationship between humans and objects and predict classes, action, object's triplets. Despite the challenges posed by the numerous interaction combinations, they also offer opportunities for multi-modal learning of visual texts. In this paper, we present a systematic and unified framework (RMRL) that enhances HOI detection by incorporating structural text knowledge. Firstly, we qualitatively and quantitatively analyze the loss of interaction information in the two-stage HOI detector and propose a re-mining strategy to generate more comprehensive visual representation. Secondly, we design more fine-grained sentence- and word-level alignment and knowledge transfer strategies to effectively address the many-to-many matching problem between multiple interactions and multiple texts. These strategies alleviate the matching confusion problem that arises when multiple interactions occur simultaneously, thereby improving the effectiveness of the alignment process. Finally, HOI reasoning by visual features augmented with textual knowledge substantially improves the understanding of interactions. Experimental results illustrate the effectiveness of our approach, where state-of-the-art performance is achieved on public benchmarks.*

**1. Introduction**

Human-object interaction (HOI) detection [16, 6] is an emerging field of research that builds upon object detection and requires more advanced high-level visual understanding. A high-performing HOI detector should not only accurately localize all interacting Human-Object pairs but also recognize their specific interactions, typically represented as an HOI triplet in the format of *class, action, objects* [6]. Previous approaches for achieving HOI detection can be divided into two pipelines: those that treat object detection and interaction recognition as separate stages [3, 4, 13, 14, 24, 20], and those that aim to handle both simultaneously [15, 26, 67, 36, 7]. Although both paradigms have made significant progress, the task remains challenging due to the vast variety of human-object interaction combinations in the real world [39, 60]. For example, the HICO-DET dataset [6] contains 600 human-object interaction combinations. A common approach is to optimize the model by mapping these various triplet labels into a discrete one-hot labels. However, this method oversimplifies the intricacy of the HOI task and can be cumbersome for model optimization.

In recent years, multi-modal learning has gained significant attention in the vision-and-language learning domains, where it has achieved state-of-the-art performance on various tasks [25, 3, 4, 31, 1, 23]. By integrating information from multiple modalities, such as images [50, 48, 31, 49] and text [65], multi-modal learning can provide a more comprehensive understanding of entities or events. In the field of HOI, several recent studies [66, 21, 57, 59, 60] have applied image-and-text models to improve interaction detection performance. For example, HOI-VP [66] used a set of

arXiv:2307.13529v2 [cs.CV] 18 Sep 2023

**Covered areas:** Limited

**Cross-evidence Reasoning:** Limited

**Task Paradigm:** Search-oriented

B.9 EXAMPLE FROM DOCVQA

One Example from DocVQA

Question ID: 24581

Question: What is name of university?

Context Modalities: Multi-Modal Documents

UNIVERSITY OF CALIFORNIA, SAN DIEGO

To Paul

Date 11/30/82 Time 2:04 <sup>A.M.</sup> <sub>P.M.</sub>

**WHILE YOU WERE OUT**

Dr. Wilson 455-8056

Ms. Sanjiv Clinic

From Sanjiv Clinic

Telephoned     Will phone again     Please phone  
 Came to see you     Will come again     Rush

**MESSAGE**

Re Program Committee  
Kidney Fdn. It will  
probably be 1st or 2nd  
week in March (1983)  
rather than latter half.  
Phone party at (None to call for)

Taken by Mary

74375-136 Source: https://www.industrydocuments.ucsf.edu/docs/nkbl0226

Covered areas:

Limited

Cross-evidence Reasoning:

None

Task Paradigm:

Search-oriented

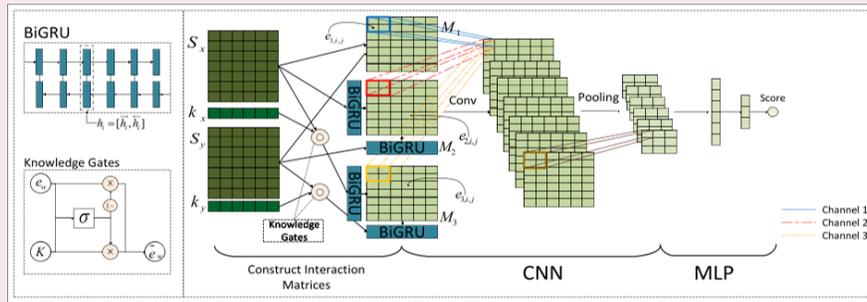
B.10 EXAMPLE FROM SPIQA

One Example from SPIQA

**Question ID:** 1611.04684v1

**Question:** What is the role of the knowledge gates in the KEHNN architecture?

**Context Modalities: Multi-Modal Figures and Charts**



**Covered areas:**

**Limited**

**Cross-evidence Reasoning:**

**None**

**Task Paradigm:**

**Search-oriented**

## C DATASET ANNOTATION AND CONSTRUCTION

### C.1 HUMAN ANNOTATOR GUIDELINES

The defective academic papers in our dataset are curated from three primary sources:

1. We synthetically inject nine types of errors into papers accepted at ICLR and *Nature Communications*.
2. For papers rejected by ICLR, we identify the shortcomings based on reviewers' comments and categorize them into the same nine error types.
3. For accepted ICLR papers, we generate consistency-related errors by cross-referencing their content against the cited literature.

To ensure the quality of each error, all entries undergo a rigorous, multistage validation protocol executed by human annotators. For synthetically generated errors, annotators manually embed them into the source papers following this protocol:

- **Credibility Validation:** Each error must be logically sound and verifiable. For generated errors, annotators first confirm their logical coherence and unambiguity. Flawed error descriptions are revised whenever possible; only irreparable cases are discarded.
- **Evidence Verification:** All evidence substantiating an error must be either directly traceable to the source document or grounded in established domain-specific knowledge. Annotators are required to meticulously verify the origin and accuracy of all supporting data and background information.
- **Category Classification:** Each error must be accurately classified into one of the 9 predefined categories according to their formal definitions. Annotators verify the correctness of the assigned category and reclassify it if necessary.
- **Paper Revision:** Upon successful validation, annotators embed the generated error into the original manuscript by adding, deleting, or modifying relevant text segments as dictated by the error's specification.

This unified and standardized annotation protocol enables the creation of a high-quality dataset of academic papers with curated errors, providing a robust benchmark for evaluating the document scanning and error detection capabilities of Large Multimodal Models.

### C.2 ANNOTATION STATISTICS

Initially, we generated or sampled a pool of 3,500 academic paper instances containing potential errors. During the manual annotation phase, following the protocol described above, we discarded 1,700 instances to ensure the logical rigor of the errors, the accuracy of the evidence, and a balanced distribution of categories.

Of the remaining 1,800 instances, 1,541(85.6%) underwent manual revision. The distribution of these modifications is as follows:

- **535 questions** were rewritten to eliminate ambiguity or to increase their retrieval and reasoning difficulty.
- **1,207 explanations** were revised to correct erroneous evidence references and resolve logical flaws.
- **1,141 instances** underwent category reclassification or manual paper editing. This process served to fix classifications that were inconsistent with our definitions and, for errors generated, to manually inject them into the source papers to create the flawed documents.





### C.3.3 CASE 3: MODIFY EXPLANATION

#### Example

##### Article

<https://doi.org/10.26434/chemrxiv-2025-5553-8>

(Supplementary Fig. 4G; Supplementary Movie 2). In the PC, the nearby flexible BP6 winged helix domains 1 and 2 (WH1 and WH2) become stabilized, forming a channel which stabilizes the DNA prior to entry into the active site (Supplementary Fig. 4H). Consistent with the role of the BP2 residue C56 in helix regulation<sup>16</sup>, assembly is found solvent exposed on the periphery of the complex with a calculated solvent accessible surface area of 241 Å<sup>2</sup>, together with an ordered structure of the clip domain binding interface in BP2 with significant distal also observed between the N-terminus of BP1 and BP2 and SNAPC4 in the human L6 PC structure (Fig. 3). Removal of the stabilizing interactions likely leads to the observed loss of BP1 structure in the human system. Given the established role of this region in transcription initiation<sup>16</sup>, this suggests a potential difference in the mechanism of transcription initiation at human and yeast. Despite PC assembly with the full-length protein, the density for the SNAPC<sup>mini</sup> complex corresponded to a core assembly (termed miniSNAPC) consisting of SNAPC2 and SNAPC3 subunits not observed (residues 142-303). SNAPC2 and SNAPC3 subunits not observed (residues 141), with the SNAPC2 and SNAPC3 subunits not observed (Supplementary Fig. 4C), consistent with their dispensable role in DNA binding.

For a direct comparison, we imaged the class III PC assembled with SNAPC<sup>mini</sup> consisting of SNAPC1 (S50), SNAPC2 and SNAPC3 (L268) using Cryo-EM. Three-dimensional classification was carried out as previously, yielding exclusively closed clamp and open-clamp OC (Supplementary Fig. 3). Only reconstructions for the closed OC produced TIRBSNAPC density sufficient quality for TIRBSNAPC<sup>mini</sup> classification and refinement, producing two OC structures (OC1<sup>mini</sup> resolved at 3.76 Å resolution and OC2<sup>mini</sup> resolved at 4.1 Å resolution which corresponded to SNAPC<sup>mini</sup> fully and partially engaged TIRBSNAPC OC structures, respectively (Supplementary Figs. 4, 6). In both instances, both SNAPC<sup>mini</sup> structures were very similar (Supplementary Table 1) and each displayed a similar fold to the miniSNAPC-DNA complex resolved in isolation (Supplementary Fig. 4, Supplementary Fig. 4H-K). Together, these results suggest that the SNAPC core is unaffected by the presence of the mobile SNAPC2 and SNAPC3 subunits and is well defined, presenting PC assembled component interfaces without any requirement for complex remodeling. As all OC structures are highly similar and represent equivalent functional states, the OC1<sup>mini</sup> structure, which displayed the highest quality reconstruction, was selected and is hereafter referred to as OC<sup>mini</sup>. All the following presented structural analysis was carried out with this model unless otherwise specified.

**BP2 uses a conserved interaction for transcription initiation**  
Structural comparison of the human H2A class III PC to the Saccharomyces cerevisiae H2A PC, formed on the SMO promoter (PFB047) revealed that BP2 shared a common mechanism of BP1 interaction with the S.c. BP1. Sequence and structural alignment of the N-terminal 29 amino acid linker region for both S.c. BP1<sup>16</sup> and H2A BP2 displayed a high level of conservation and structural similarity (Fig. 5A). Alignment of both the human and S.c. PC structures identified an equivalent docking position for BP1 and BP2 on BP1 between the human and yeast structures, with the BP1 BP2 region involved presenting highly conserved interactions (Fig. 5C; Supplementary Fig. 7A), demonstrating an equivalent mode of interaction irrespective of species and protein.

In contrast, the BP1 and BP2-associated factor BP3 displays significant structural differences between the human and yeast PCs. Both H.c. and S.c. BP3 form a conserved SNAPC domain which facilitates interactions with TIRBSNAPC (Fig. 3D). However, in the yeast PC, BP3 undergoes a significant conformational change from an unfolded to a largely ordered structure. The N-terminal 'yester' domain folds to engage the C1' initiation termination loop and stabilizes the WH1 and WH2 domains of the S.c. BP3 homologue C34<sup>16</sup>. C-terminal to this,

the clip domain (also previously identified as the BP3 region, encompassing residues 200-312 in yeast<sup>16</sup>) folds to form contacts with the Cyclics domain of BP1. No such transition is observed for the human system, with both regions N- and C-terminal to the SANT domain remaining disordered upon PC formation (Fig. 3D). Alignment of human and yeast structures revealed a loss of the ordered initiation-termination loop in human BP3 (Fig. 3D) together with an altered structure of the clip domain binding interface in BP2 with significant distal also observed between the N-terminus of BP1 and BP2 and SNAPC4 in the human L6 PC structure (Fig. 3). Removal of the stabilizing interactions likely leads to the observed loss of BP1 structure in the human system. Given the established role of this region in transcription initiation<sup>16</sup>, this suggests a potential difference in the mechanism of transcription initiation at human and yeast. Despite the loss of BP3 structure, the human BP3 and yeast BP3 adopt a highly similar conformation in the yeast, which was unexpected in the absence of an ordered BP3 stem and other, nevertheless, a comparison between both structures revealed a basic surface patch presented by BP3C in both human and yeast structures (Fig. 3G), suggesting that the interaction with the upstream DNA is sufficient to stabilize BP3C WH1 and 2 upon formation of the PC.

The structural comparison also discerned differences within the polymerase active site. Whilst most functional regions including fork loops 1 and II, which II, the bridge helix, and the trigger loop were identical, as suggested by their sequence conservation (Supplementary Fig. 7), C1' differences were observed in the ruler and switch loops. Comparison of H.c. and S.c. ruler sequences revealed a divergence between both regions which was also observed in the structure, with the human ruler adopting a highly similar folded structure to yeast, elongating and PC engaged states in contrast to the yeast polymerase, where this region is disordered in the PC (Fig. 3H) and folds upon transition to the elongating complex (EC)<sup>17</sup>. The ruler forms stabilizing interactions with the RNAP-DNA hybrid in the active site, suggesting a difference in transcription initiation<sup>18</sup>. Differences were also observed in the switch III loop despite high sequence conservation. Importantly, S.c. BP1 N-terminal domain is critical for full expansion of the initial transcription bubble around position +9 via an allosteric mechanism through direct contact with the Switch III loop<sup>19</sup>. Indeed, S.c. BP1 Zr1808 reconfigures the switch III loop, preventing the occlusion of the template strand from the active site and allowing for the stabilization of the transcription bubble (PFB047, ec1, ec2)<sup>19</sup>. However, the human switch III loop is not conservedly open in all states (PFB268, 7ae1)<sup>19</sup> (Fig. 3I) suggesting that human and yeast BP1 BP2-mediated DNA opening, and therefore transcription initiation, may be differentially regulated.

**SNAPC structure within the Class III PC**  
Resolution of the SNAPC<sup>mini</sup> core structure in the OC context observed a similar structure to the BP2-bound miniSNAPC complex (PFB268) (Supplementary Table S1). SNAPC<sup>mini</sup> presents a central structural unit together with the entire SNAPC complex is assembled (Fig. 3A). The first 5 helices in the N-terminus of SNAPC2 the 'base' domain together with an anti-parallel β-sheet of the ubiquitin like domain (ULD) completely envelop the N-terminus of SNAPC2 (Fig. 3B). SNAPC<sup>mini</sup> assembles on the opposing face of SNAPC3, forming multiple contacts

##### Article

<https://doi.org/10.26434/chemrxiv-2025-5553-8>

transcribed by Pol II despite their defined BP III PSE-TATA box spacing<sup>17</sup>. As a result, the small spacing difference observed between L2 and L6 promoters is unlikely to drive the flipping of the SNAPC. We propose that the flipped orientation to a specific adaptation of SNAPC to interact with TIRBSNAPC complex at Pol II and BP III promoters and is governed by a mutually exclusive set of interactions and steric clashes that by the SNAPC conformation in each complex. Thus, the only determinant of polymerase selectivity is the presence or absence of a TATA box, as previously demonstrated<sup>17</sup>. The presence of a canonical TATA box favours the recruitment of BP2 TBP complex and subsequently RNA Pol II. In the absence of a TATA box, recruitment of TBP-TBA prevents, resulting in the subsequent recruitment of RNA Pol II. In both scenarios, SNAPC can interact in a similar manner with the cognate PSE box while adopting two distinct and mutually exclusive conformations to engage with either BP2 TBP or TBA/TBP, resulting in the recruitment of RNA Pol II or Pol II, respectively.

Despite the increasing amount of structural data available for SNAPC and the class III PC<sup>17</sup>, the arrangement of SNAPC2 and SNAPC3 subunits in the SNAPC complex remained elusive. Employing sulfur-SDA crosslinking mass spectrometry analysis, the location and arrangement of these subunits was investigated. Combining this with AlphaFold predictions, the SNAPC2 binding site was localized to a structural module in the C-terminus of SNAPC4, consistent with previous mutational analysis which identified the region spanning 1281-1393 of SNAPC4 as the interaction site<sup>17</sup>. We therefore confirm this as the binding site of SNAPC2 while defining contact sites in the N- and C-termini (Fig. 4E). SNAPC2 was also identified as part of the wing 2 helical bundle despite the absence of a defined wing 2 structure in the Pol II BP1 and miniSNAPC-DNA structures, consistent with the requirement for this region to produce a stable SNAPC complex<sup>17</sup>. Multiple crosslinks were identified between both wing 2 and SNAPC2, suggesting that these subunits together form a large structural module which is highly mobile. MAP modelling placed this structure in proximity to the bound DNA, consistent with the role of these subunits in regulating DNA binding affinity<sup>17</sup>. The observed localization and flexibility of this module allows us to speculate that SNAPC2-SNAPC4 module may regulate DNA access to the binding site here modulating the SNAPC-DNA binding affinity. Indeed, the identified mobile module encompassed the OC1 intersecting region (OR) in BP1<sup>16</sup> which was close to several SNAPC2 crosslinking sites (Supplementary Fig. 8F) and has been implicated in allowing RNA-DNA binding repression and stabilizing SNAPC at the promoter<sup>17</sup>, a function which has also been suggested for SNAPC2<sup>17</sup> leading to the possibility that OC1 may interact with a SNAPC2 gene to regulate SNAPC DNA binding. In the class III PC context, this SNAPC2-containing mobile module faces the upstream region, pointing towards the nucleosome and bound OC1, potentially facilitating the interaction. As a result, future work will focus on characterizing the interaction between SNAPC-containing PC1 and upstream regulatory factors to give greater mechanistic insight into class III promoter transcription and a more complete structural description of the complex class III PC.

##### Methods and Protein Expression

All TIRBSNAPC components were expressed recombinantly in Rosetta (DE3) pLys Escherichia coli cells (Sigma Aldrich, 74031M)<sup>20</sup>. Briefly, N-terminally His-tagged TIRBSNAPC (10x30x10x100)<sub>1-200</sub> was expressed in a pET43b(+) vector. Following transformation, cells were grown in terrific broth (TB) at 37 °C to an optical density (A<sub>600</sub>) of 0.6 and subsequently induced with 0.5 mM isopropyl β-D-1-thiogalactopyranoside (IPTG; Thermo Scientific, R0393) at 20 °C for 18 hours. Cells were harvested by centrifugation at 2000 × g for

30 mins at 4 °C and stored at -80 °C prior to purification. Full length, N-terminally tagged BP2 was expressed in a pORFME construct. Transformed cells were grown in TB at 37 °C to a final A<sub>600</sub> of 0.6 and induced for 4 hours at 30 °C with 1 mM IPTG. Cells were collected by centrifugation at 2000 × g for 20 mins at 4 °C and stored at -80 °C. N-terminally His-tagged BP1 (1-484) was expressed in a pORFME vector. Cells were grown in TB at 37 °C until A<sub>600</sub> of 0.6. The temperature was reduced to 16 °C and cultures subsequently induced with 0.5 mM IPTG once A<sub>600</sub> reached 0.6. Induction was carried out for 4 hours and cells collected by centrifugation at 2000 × g for 20 mins at 4 °C. Cell culture supernatant at -80 °C.

Full length SNAPC (SNAPC<sup>full</sup>) and mini SNAPC (SNAPC<sup>mini</sup>) were expressed using the BP2lac baculovirus expression system<sup>17</sup>. Briefly, for SNAPC<sup>full</sup> expression, a double-tagged SNAPC2 subunit, carrying a T1-cleavable N-terminal stop and C-terminal His tag, SNAPC1, SNAPC2, SNAPC3 and SNAPC4 were each sub-cloned into a pUB vector and subsequently combined into a pR262 expression construct by NdeI/BstI Hifi DNA Assembly Cloning Kit (NEB, M2290A) using the bicluc expression system. For SNAPC<sup>mini</sup> construction, an N-terminal T1-cleavable Stop and His tag were added to SNAPC1 (S50) and SNAPC3 respectively. These constructs were sub-cloned into pUB vectors alongside SNAPC1 (268aa), SNAPC2 and SNAPC3 and subsequently combined into a pR262 construct for expression. Following cloning, both SNAPC<sup>full</sup> and SNAPC<sup>mini</sup> pR262 constructs were used for baculovirus production via transfection into Drosophila S2 cells (Genescreen Bioassay) for baculovirus generation, baculovirus was isolated and transfected into 2 ml of adherent Sf9 cells (Thermo Fisher Scientific, 14496031) at 0.5 × 10<sup>6</sup> cell/ml density in Insect-XPRESS media (Lonza, L2-7000) in a 96-well plate. Cells were incubated at 27 °C for 72 hours and supernatant (total volume of 2 ml), containing the baculovirus, was collected post-confluent. The P1 virus, viral amplification was achieved by P1 infection of a 25 culture of Sf9 cells growing in suspension at a cell density of 0.5 × 10<sup>6</sup> cells/ml in Insect-XPRESS media (Lonza, L2-7000). Cells were incubated at 27 °C for 130 h for 5 days and supernatant collected following centrifugation of the cultures at 1000 × g for 15 mins at 4 °C. This produced the P2 virus which was stored at 4 °C. Large-scale protein expression was carried out through the addition of 2 ml of P2 viral stock to 500 ml of Five Cells (Thermo Fisher Scientific, B65502) grown to 0.5 × 10<sup>6</sup> cells/ml in Insect-XPRESS media (Lonza, L2-7000) in a roller bottle flask (Genescreen). Cells were grown at 27 °C for 3 days at 100 rpm, and harvested through centrifugation at 3000 × g for 20 mins at 4 °C. Cell pellets were resuspended in phosphate buffered saline (PBS), centrifuged as previously and the resulting washed pellet stored at -80 °C prior to purification.

Flagged endogenous human RNA polymerase II was purified from HEK293T cells (CRISPR-Cas9 modified to include a C-terminal mCherry-Ter199 stop tag on the BP1C1 (POLR2C) subunit)<sup>21</sup>. To produce sufficient material for purification, large-scale cell cultivation was carried out in a 96 L Anticlonal Biotech Reactor (ezControl Bioreactor, Applikon) using a batch cultivation in Dynamic Flex Medium (Gibco, A36520) supplemented with 1% penicillin-streptomycin, 1 mM stable glutamine (Gibco, C6304043), 0.3% Pluronic F-68 non-ionic surfactant (Gibco, 2404032), and 90 ppm silicone-based Antifoam C Emulsion (Sigma-Aldrich, AR01). Culture conditions included a temperature of 37 °C, pH 7.0 (controlled by CO<sub>2</sub>), and dissolved oxygen at 40%. The latter was controlled by sugar gating with air flow ranging from 300 to 500 l/min, along with a second cascade of CO<sub>2</sub> gating from 1 to 0.100 l/min. Agitation was maintained at 100 rpm using a three-pitched blade impeller. Initially, cells underwent sub-culture and expansion and reaching a density of 4 × 10<sup>6</sup> viable cells/ml. Subsequently, cells were used to inoculate the bioreactor at a concentration of 0.5 × 10<sup>6</sup> viable cells/ml, depending on the desired cell count. Harvesting took place at 10 × 10<sup>6</sup> viable cells/ml.

Nature Communications | (2025)16:141

4

Nature Communications | (2025)16:141

12

**Question:** Evaluate if the composition of the SNAPCmini construct is consistently defined throughout the paper.

**Explanation:** The results on page 4 state that the assembled SNAPCmini includes the SNAPC2 subunit. However, the methods on page 12 describe the construction of SNAPCmini using only SNAPC4, SNAPC3, and SNAP1, with SNAPC2 explicitly removed from the cloning description. A third conflicting statement on page 6 implies SNAPC2 was expected to be part of the minimal core, creating conceptual and operational inconsistency regarding this key experimental complex.

**Error Type:** RQD (Research Question & Definitions)

#### Before:

**Explanation:** ...with SNAPC2 explicitly removed from the cloning description. A third conflicting statement on page 6...

#### Decision: Modify

#### After:

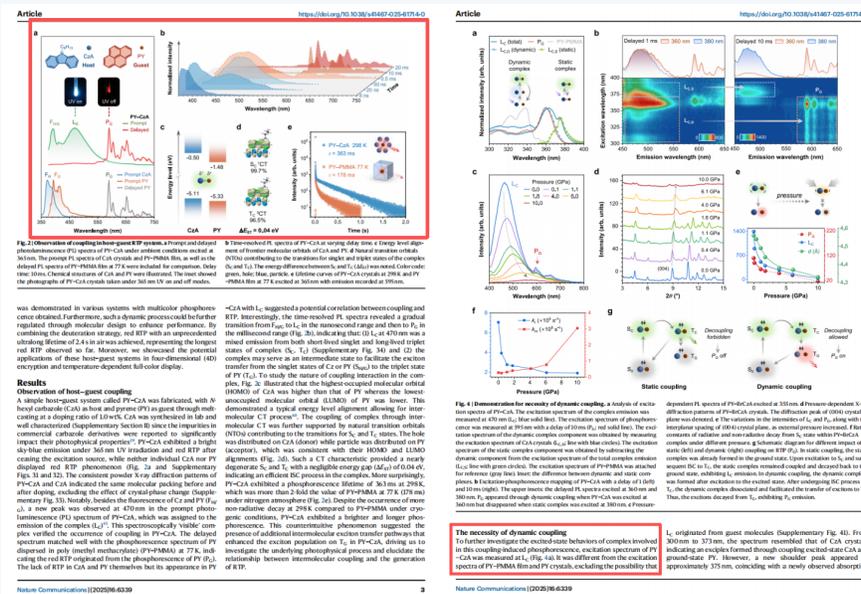
**Explanation:** ...with SNAPC2 explicitly removed from the cloning description.

**Analysis:** This instance targets an inconsistency in the operational definition of the SNAPCmini construct. Specifically, the results section states that the assembled SNAPCmini includes the SNAPC2 subunit, while the methods section explicitly describes the construction of SNAPCmini using only SNAPC4, SNAPC3, and SNAP1, with SNAPC2 removed from the cloning procedure. The original explanation additionally referenced a speculative statement regarding SNAPC2's expected presence in the minimal core, which introduced unnecessary ambiguity and reduced the clarity of the definition-level inconsistency. By removing this auxiliary statement, the modified instance focuses on a clear and realistic mismatch between experimental description and implementation, which is representative of RQD-type errors commonly encountered in academic writing.



D.2 DI (DESIGN & IDENTIFIABILITY)

Example: 1006



**Question:** Assess the Experiments section for Design & Identifiability issues.

**Explanation:** The paper’s core argument is that it identifies a specific ‘dynamic coupling’ pathway as essential for RTP, distinct from a ‘static coupling’ pathway. The edits state that the key experiment (excitation-phosphorescence mapping) cannot distinguish between these two pathways, as the final phosphorescence shows spectral signatures of originating from both. This introduces a structural identification problem: with two potential causal pathways leading to the same outcome and no way to isolate their effects, the claim that the dynamic pathway is the definitive mechanism is not identifiable from the data presented.

**Error Type:** DI (Design & Identifiability)

**Type:** Within-Generate

### D.3 SG (SAMPLING & GENERALIZABILITY)

Example1: 1014

**Article** <https://doi.org/10.1038/s41467-025-59630-4>

**Methods**  
**The simulation of photon propagation in biological tissues via Monte Carlo method**  
 The fluorescence signal source is set to a line of 2 mm long and 0.1 mm wide, and emits a total of 1 × 10<sup>6</sup> photons in random directions. The refractive index of tissues is set as 1.37 and the scattering anisotropy factor is set as 0.9. The dimensions of the tissue in the simulation are set as 100 mm in both length and width. After the photon has propagated through the tissue, it is caught through the microscopic imaging system and reaches the detector. The detector is a 2D array plane with a resolution of 512 × 512 and the pixel size is 20 μm. In simulations of vascular imaging in mouse tissues, the absorption coefficient of water is assumed to be the absorption coefficient of the tissue, considering that water is the most predominant component of biological tissues. The scattering coefficient is assumed to be equal to that in the vascular imaging simulation, the line source is set as a depth of 1 mm and extended background from different depths are added to simulate background such as tissue autofluorescence. For the simulation of a vascular imaging scene under strong background interference at depth, in addition to a vertical line light source (depth = 1 mm), a horizontal line source (depth = 1 mm) with a length of 2 mm and a width of 0.5 mm which emits a total of 1 × 10<sup>6</sup> photons in random directions is set to act as the on-axis background caused by the liver. In the simulation of imaging in adipose tissue, the previously measured absorption and reduced scattering spectra of fresh porcine adipose by the integrating sphere method are used for the simulation. The specific measurement steps are described in the supporting information. On this basis, separate simulations are performed for vertical line source located at tissue depths of 1.2 and 3 mm.

**Optical system for fluorescence imaging**  
 The self-developed NIR-II fluorescence macro imaging system includes an excitation light source and an imaging module. Two CW lasers of 793 and 656 nm are used as excitation sources, in which 793 nm excitation offers the advantage of higher excitation efficiency (1.6 SEI) while their emission provides the benefit of a higher safe power density limit. After being collimated, the laser beam is expanded through a lens and then further homogenized by a ground glass diffuser, ensuring that a uniform illumination is produced on the observed sample. The fluorescence emitted from the sample passes through the filter and objective lens, and then is imaged on the detector of a customized sCMOS camera (spectral range of 900–2500 nm). According to the NIR of the camera's standard interface communication protocol, an independent software has been customized with the function of loading, collecting and saving images. The objective lens used is OCELOS OPTICS' F45mm, focal length = 30.7 mm, transmission spectral range of 600–2500 nm. In all our experiments, we realize the detection and imaging of different wavelengths to achieve correct focusing, ensuring that image clarity is maintained across the entire field of view (Fig. S23).  
 In the in vivo fluorescence imaging experiments, the highest excitation intensity used is 0.6 W/cm<sup>2</sup> for the 793 nm CW laser and 0.4 W/cm<sup>2</sup> for the 656 nm CW laser, which are both below the laser safety limits for NIR lasers (U.S. National Standard for Safe Use of Lasers, ANSI Z136.1-2014). The highest excitation intensity used is 0.29 W/cm<sup>2</sup> for the 793 nm CW laser in the in vivo fluorescence imaging experiments. The integration time for imaging are all 70 ms.

**Chemicals**  
 PEGs, GDs, and sulfur powder were purchased from Alfa. Oleic acid, oleic acid, 1-oleoyl-3-glycerol, poly(acrylic acid) (PAA), DMSO, 2,4,6-triphenylsulfoniumhexafluorophosphate (TSP), N,N-dimethylformamide (DMF), N,N-dimethylacetamide (DMAc), N,N-dimethylpropyleneamine dihydrochloride (DMAP), and sodium acetate were purchased from Sigma-Aldrich. sMPC-amine (DMV-5) was purchased from Laysan Bio. 8-AM PEG-amine (DMV-40K) was purchased from Advanced Biochemicals.

**Synthesis of PEG-CAS GDs**  
 The sulfur precursor solution was prepared by mixing 0.08 g of sulfur powder with 7.5 mL of oleic acid in a microwave flask under argon at 120°C for 30 min. The lead precursor solution was prepared by combining 0.53 g of PEG<sub>400</sub> and 7.2 mL of oleic acid in a three-neck flask, degassing under argon for 30 min, and then heating to 45–50°C (depending on the desired particle size). Under continuous stirring, 2.25 mL of the sulfur precursor solution was injected into the lead precursor solution. Once the desired growth (typically 3–60 min) was reached, the reaction was quenched by adding 30 mL of cold hexane. The products were collected via centrifugation and redispersed in a mixture of 30 mL hexane and 20 mL oleic acid. The mixture was agitated for 10 min to remove excess sulfur from the products. The GDs were then precipitated by centrifugation. This precipitation procedure was repeated three times until the supernatant became colorless. The GDs were redispersed in a mixture of toluene and ODE. The PEG-CAS GDs were synthesized via a calcium-exchange procedure. GDS (0.5 g), oleic acid (4 mL), and ODE (15 mL) were washed to 200°C purged with argon, and then cooled to 250°C. 5 mL of PEG-CAS in toluene/ODE was bubbled with argon for 1 min and then injected into the GDS precursor. The reaction flask was quenched by stopping stirring and being immediately after the growth reaction was carried out at 100°C for 20–120 min. The PEG-CAS GDs were precipitated with ethanol and then redispersed in hexane.

**Modification and PEGylation of GDs**  
 0.1 g of polystyrene acid powder (average MW = 8000) and 1.5 g of GDS were transferred into a round-bottom flask. The mixture was dissolved in 10 mL of DMF, and approximately 1.2 mL of oleic acid was added separately. The solution was stirred overnight. After the reaction, 50 mL of 0.3 M HCl was added to the mixture. The resulting precipitate was separated by centrifugation and redispersed in a mix of methanol. Subsequently, 20 mL of 1 M HCl was added to the solution, and the procedure was again repeated by centrifugation. This purification process was repeated at least five times. The final precipitate was dissolved in 5 mL of chloroform and washed with 30 mL of HCl. The organic phase was collected and dried over anhydrous Na<sub>2</sub>SO<sub>4</sub>. The chloroform was removed under vacuum, yielding olefin-terminated polystyrene acid (OPA). Asymmetrical PEG-CAS GDS (5.0 mg) were dissolved in 2.0 mL of chloroform containing 12 mg of OPA. The mixture was stirred at room temperature for 30 min, and the solvent was removed under vacuum using a rotary evaporator. The residue was then dissolved in 50 mM Na<sub>2</sub>CO<sub>3</sub> solution under vacuum. The modified GDs were precipitated by ultra-centrifugation at 75,000 rpm for 1 h. The purified product was dissolved in pH 5.5 MBS buffer (0.1 M NaCl and 0.05 M Tris) and stored at 4°C. 0.1 mg of sMPC-amine (DMV-5) and 1 mg of 8-AM PEG-amine (DMV-40K) were dissolved in 1 mL of MBS buffer at a molar ratio of 2:1 and added to the GDS dispersion under stirring. Additionally, 10 mg of DMSO was dissolved in 500 μL of MBS buffer and added to the GDS dispersion. The mixture was stirred at room temperature overnight. The PEG-CAS GDs were purified using a 100 kDa filter and washed the times with PBS to remove excess reactants. The final product was dissolved in PBS and stored at 4°C.

**Animal handling**  
 All experimental procedures were approved by the Animal Use and Care Committee of Zhejiang University (1920202053). The experiment animals involved in this research included a **specific substrain of diabetic mice** (National Standard for Genetic Germplasm Resources of Diabetic Mice (C57BL/6J) mice) (C-20) and rabbits (C14), which were obtained from S.M.C. Laboratory Animal Co., Ltd. (Ningbo, China). All the

**nature communications**

**Article** <https://doi.org/10.1038/s41467-025-59630-4>

**High-contrast in vivo fluorescence imaging exploiting wavelengths beyond 1880 nm**

Received: 25 September 2024  
 Accepted: 25 April 2025  
 Published online: 13 May 2025

Check for updates

Jiayi Liu<sup>1,2</sup>, Qingxi Xia<sup>1,2</sup>, Tianxiang Wu<sup>1,2</sup>, Yuhuang Zhang<sup>1</sup>, Shiyi Peng<sup>1</sup>, Yifei Li<sup>1</sup>, Yuzean Li<sup>1,2</sup>, Hua Lei<sup>1,2</sup>, Mingqi Zhang<sup>1,2</sup> & Jun Qian<sup>1,2</sup>

The second near-infrared (NIR-II) window is widely acknowledged for its excellent potential in in vivo fluorescence imaging. Currently, NIR-II fluorescence imaging predominantly operates within the 900–1850 nm spectral range, while the region beyond 1880 nm has been disregarded due to the large light absorption of water. Based on a refined understanding of the effect of light absorption on imaging, we propose an approach that utilizes the previously neglected region surrounding the water absorption peak at 1910 nm for imaging. Both simulations and experiments confirm that the water absorption contributes positively to imaging, enabling high-contrast in vivo fluorescence imaging in the 1880–2080 nm window. To further assess the applicability of this approach in different biological media, we extend our focus to fluorescence imaging in adipose tissue. This leads to the expansion of the imaging window to 1700–2080 nm, owing to the unique light absorption characteristics of adipose tissue. Our results demonstrate that the 1700–2080 nm region provides optimal imaging quality in adipose tissue, attributing to its moderate absorption and low scattering. This work, using a specialized disease model, advances our understanding of the interplay between light absorption and photon scattering in bioimaging, providing a definitive insight for selecting optimal imaging windows to achieve high-contrast fluorescence imaging for all patients.

Fluorescence imaging, combining the advantages of high-power excitation and optical resolution and radiation-free properties, has a wide range of applications in the biomedical field. In recent years, the second near-infrared window (NIR-II, 900–1800 nm) fluorescence imaging, has gained significant attention due to its low absorption scattering and reduced biological autofluorescence<sup>1–3</sup>. In particular, based on the varying properties of scattering and absorption, NIR-II is divided into several specific imaging sub-windows, including NIR-IIa (1300–1400 nm)<sup>4</sup>, NIR-IIb (1500–1700 nm)<sup>5</sup>, and NIR-IIc (1700–1880 nm)<sup>6,7</sup>. Indeed, light propagation through biological media is an essential process in fluorescence bioimaging. This scattering perturbations, which have conventionally been regarded as detrimental to the performance of bioimaging. As a result, a common approach improving imaging quality is to minimize scattering by increasing wavelength and avoid spectral regions with strong absorption peaks<sup>8–10</sup>. However, recent studies have challenged this conventional opinion by highlighting the beneficial role of absorption in enhancing the imaging contrast<sup>11–13</sup>. The improvement in contrast is attributed to the greater attenuation of background signals with longer optical paths in scattering media, thereby increasing the relative proportion of useful signals. This

**Author contributions**  
 J.L. and Q.X. conceived the idea and designed the experiments. J.L. performed the experiments and analyzed the data. T.W. and Y.Z. provided technical support. Y.H.Z., S.Y.P., Y.F.L., H.L., and M.Q.Z. contributed to the data analysis and manuscript preparation. J.Q. supervised the project and provided funding. All authors discussed the results and contributed to the final manuscript.

**Additional Information**  
 Correspondence: Jun Qian (qianjun@zhujiu.com)

**Supplementary Information**  
 Supplementary Information is available for this article.

**Reprints and permissions**  
 Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Check for updates**  
 Check for updates information is available at [www.nature.com/check-for-updates](http://www.nature.com/check-for-updates).

**© The Author(s) 2025**

**Question:** Assess the Methods section for Sampling & Generalizability issues.

**Explanation:** The Methods section is edited to state that the experiments used a “specific substrain of diabetic” mice, a highly specialized sample. However, the Abstract and Discussion make broad, unsupported claims of generalizability to “all patients” and the “general patient population.” This constitutes an invalid sample-to-population inference.

**Error Type:** SG (Sampling & Generalizability)

**Type:** Within-Generate



## D.4 RCA (REFERENTIAL AND CITATION ALIGNMENT)

Example: 0

Published as a conference paper at ICLR 2025

Second, they assume the sensors form a fixed spatial grid and record data simultaneously, which is not the case in our setup where the data could come from as little as a single sensor at each time point (Fig. 1). For example, some earlier methods assume the observations are made on a fixed and regular spatiotemporal grid (Gong et al., 2018; Geyer & Johnson, 2005), other methods work with irregular but still fixed observation locations (Jakovlev et al., 2021; Lioy & Glimm, 2022), and other works go further and allow the observation locations to change over time (Fitz et al., 2021; Yu et al., 2022) but fit the observation times and assume dense observation. Whereas, another line of research has proposed methods that model only the spatiotemporal observation process without modeling the system dynamics (Chen et al., 2021; Zhu et al., 2021; Zhou et al., 2022; Zhou & Yu, 2023; Du et al., 2024).

Our work fills this gap and proposes a model for randomly observed spatiotemporal dynamical systems. Our model incorporates techniques from amortized variational inference (Kingma & Welling, 2013), neural differential equations (Chen et al., 2018; Rackačius et al., 2020), neural point processes (Mei & Flierler, 2017; Chen et al., 2021), and implicit neural representations (Chen et al., 2022; Yin et al., 2022) to efficiently learn both the underlying system dynamics and the random observation process. Our model uses initial observations to obtain the variational estimate of the latent initial state via a transformer encoder (Vaswani et al., 2017), simulates the latent trajectory with neural ODEs (Chen et al., 2018), and uses implicit neural representations to parameterize the point process and observation distribution. Furthermore, we identify a computational bottleneck in the latent state evaluation and propose a technique to alleviate it, resulting in up to 4x faster training. Our model shows strong empirical results outperforming other models from the literature on challenging spatiotemporal datasets.

## 2 BACKGROUND

### 2.1 SPATIOTEMPORAL POINT PROCESSES

Spatiotemporal point processes (STPP) model sequences of events occurring in space and time. Each event has an associated event time  $t_i \in \mathbb{R}_{>0}$  and event location  $\mathbf{x}_i \in \mathbb{R}^d$ . Given an event history  $H_t = \{(t_i, \mathbf{x}_i)\}_{i=1}^t$  with all events up to time  $t$ , we can characterize an STPP by its conditional intensity function

$$\lambda^*(t, \mathbf{x}) \triangleq \lim_{\delta t \rightarrow 0} \frac{P(t_i \in [t, t + \delta t], \mathbf{x}_i \in B_{\delta t}(\mathbf{x}) | H_t)}{\delta t | B_{\delta t}(\mathbf{x})} \quad (1)$$

where  $\delta t$  denotes an infinitesimal time interval, and  $B_{\delta t}(\mathbf{x})$  denotes a  $d$ -ball centered at  $\mathbf{x}$ . Given a history  $H_t$  with  $t-1$  events,  $\lambda^*(t, \mathbf{x})$  describes the instantaneous probability of the next,  $t$ th, event occurring at time  $t$  and location  $\mathbf{x}$ . Given a sequence of  $N$  events  $\{(t_i, \mathbf{x}_i)\}_{i=1}^N$  on a bounded domain  $\mathcal{A} \subset [0, T] \times \mathbb{R}^d$ , the log-likelihood for the STPP is evaluated as (Oshley et al., 2007)

$$\log p(\{(t_i, \mathbf{x}_i)\}_{i=1}^N) = \sum_{i=1}^N \log \lambda^*(t_i, \mathbf{x}_i) - \int_{\mathcal{A}} \lambda^*(t, \mathbf{x}) d\mathbf{x} dt. \quad (2)$$

Marked STPP extends the above simple STPP by a mark  $\mathbf{y}_i \in \mathbb{R}^k$  that is associated to each event  $(t_i, \mathbf{x}_i)$ .

### 2.2 ORDINARY AND PARTIAL DIFFERENTIAL EQUATIONS

Given a deterministic continuous-time dynamic system with state  $\mathbf{x}(t) \in \mathbb{R}^k$ , we can describe the evolution of its state in terms of an ordinary differential equation (ODE)

$$\frac{d\mathbf{x}(t)}{dt} = \mathbf{f}(\mathbf{x}, t(t)). \quad (3)$$

For an initial state  $\mathbf{x}_1$  at time  $t_1$ , we can solve the ODE to obtain the system state  $\mathbf{x}(t)$  at later times  $t > t_1$ . The solution exists and is unique if  $\mathbf{f}$  is continuous in time and Lipschitz continuous in state (Coddington et al., 1956), and can be obtained either analytically or using numerical ODE solvers (Hairer et al., 1997). In this work we solve ODEs numerically using ODE solvers from

Published as a conference paper at ICLR 2021

Model	Preheat		Earthquake JP		COVID-19 NJ		BOLD5000	
	Temporal	Spatial	Temporal	Spatial	Temporal	Spatial	Temporal	Spatial
Hawkes Process	-4731.00 ± 0.11	-4111.00 ± 0.09	-4111.00 ± 0.09	-4111.00 ± 0.09	-4111.00 ± 0.09	-4111.00 ± 0.09	-4111.00 ± 0.09	-4111.00 ± 0.09
Self-exciting Process	-2111.00 ± 0.11	-1811.00 ± 0.09	-1811.00 ± 0.09	-1811.00 ± 0.09	-1811.00 ± 0.09	-1811.00 ± 0.09	-1811.00 ± 0.09	-1811.00 ± 0.09
Neural Hawkes Process	-4731.00 ± 0.11	-4111.00 ± 0.09	-4111.00 ± 0.09	-4111.00 ± 0.09	-4111.00 ± 0.09	-4111.00 ± 0.09	-4111.00 ± 0.09	-4111.00 ± 0.09
Neural Hawkes Process	-4731.00 ± 0.11	-4111.00 ± 0.09	-4111.00 ± 0.09	-4111.00 ± 0.09	-4111.00 ± 0.09	-4111.00 ± 0.09	-4111.00 ± 0.09	-4111.00 ± 0.09
Conditional KDE	-2111.00 ± 0.11	-1811.00 ± 0.09	-1811.00 ± 0.09	-1811.00 ± 0.09	-1811.00 ± 0.09	-1811.00 ± 0.09	-1811.00 ± 0.09	-1811.00 ± 0.09
Temporal CNF	-2111.00 ± 0.11	-1811.00 ± 0.09	-1811.00 ± 0.09	-1811.00 ± 0.09	-1811.00 ± 0.09	-1811.00 ± 0.09	-1811.00 ± 0.09	-1811.00 ± 0.09
Neural Jump SDE (ODE)	4086.00 ± 0.07	4186.00 ± 0.05	4186.00 ± 0.05	4186.00 ± 0.05	4186.00 ± 0.05	4186.00 ± 0.05	4186.00 ± 0.05	4186.00 ± 0.05
Jump CNF	4022.00 ± 0.06	4166.00 ± 0.04	4166.00 ± 0.04	4166.00 ± 0.04	4166.00 ± 0.04	4166.00 ± 0.04	4166.00 ± 0.04	4166.00 ± 0.04
Attentive CNF	4086.00 ± 0.07	4186.00 ± 0.05	4186.00 ± 0.05	4186.00 ± 0.05	4186.00 ± 0.05	4186.00 ± 0.05	4186.00 ± 0.05	4186.00 ± 0.05

Table 1: Log-likelihood per event on held-out test data (higher is better). Standard devs. estimated over 5 runs.

**Results & Analyses** The results of our evaluation are shown in table 1. We highlight all results where the intervals containing one standard deviation away from the mean overlap.

Across all data sets, the Time-varying CNF outperforms the conditional KDE baseline despite not being conditional on history. This suggests that the overall spatial distribution is rather complex. We also see from Figure 7 that Gaussian clusters tend to compensate for false events by learning a larger band-width whereas a flexible CNF can only model multi-modal event propagation.

The Jump and Attentive CNF models achieve better log-likelihoods than the Time-varying CNF, suggesting prediction in these data sets benefit from modeling dependence on event history.

For Covid-19, the self-exciting Hawkes process is a strong baseline which aligns with similar results for other infectious diseases (Fitz et al., 2021), but Neural STPPs can achieve substantially better spatial likelihoods. Overall, STPP is competitive with the Neural Jump SDEs; however, it tends to fall short of the Attentive CNF which jointly models spatial and temporal variables.

In a closer comparison to the temporal likelihood of Neural Jump SDEs (Yu & Brown, 2019), we find that overly-restrictive spatial models can negatively affect the temporal model since both domains are highly coupled. Since our realization of Neural Jump SDEs and our STPP use the same underlying architecture to model the temporal domain, the temporal likelihood values are often close. However, there is still a statistically significant difference between our Neural STPP models and Neural Jump SDEs even for the temporal log-likelihood on all data sets.

Finally, we note that the results of the Jump and Attentive CNFs are typically close. The Attentive model generally achieves better temporal log-likelihoods while maintaining competitive spatial log-likelihoods. This difference is likely due to the Attentive CNF’s ability to attend to all previous events, while the Jump CNF has to compress all history information inside the hidden state at the time of event. The Attentive CNF also enjoys substantially faster computation (see Appendix A).

## 6 CONCLUSION

To learn high-fidelity models of stochastic events occurring in continuous space and time, we have proposed a new class of parametrizations for spatio-temporal point processes. Our approach combines ideas of Neural Jump SDEs with Continuous Normalizing Flows and allows us to retain the flexibility of neural temporal point processes while enabling highly expressive models of continuous marks. We leverage Neural ODEs as a computational method that allows computing, up to negligible numerical error, the likelihood of the joint model, and we show that our approach achieves state-of-the-art performance on spatio-temporal datasets collected from a wide range of domains.

A promising area for future work are applications of our method in earth and climate science which often are concerned with modeling highly complex spatio-temporal data. In this context, the use of Riemannian CNFs (Matsen & Nobile, 2020; Lee et al., 2020; Fawcett & Froyen, 2020) is especially interesting as it allows us to model Neural STPPs on manifolds (e.g. the earth’s surface) by simply replacing the CNF in our models with a Riemannian counterpart.

**Question:** Scan the errors in cited reference Chen et al. (2021)

**Explanation:** The edited P contains a Type H error by misrepresenting the performance of the cited model. P (p. 8) claims that the NSTPP model from Chen et al. (2021) ‘reported performance comparable to a standard Hawkes process baseline’. This contradicts the results in S, where the proposed models (i.e., NSTPP) consistently outperform the Hawkes process baseline, often by a large margin. For example, S (p. 9, Table 1) shows on the BOLD5000 dataset that the ‘Attentive CNF’ model achieves a temporal log-likelihood of  $5.842 \pm 0.005$ , which is substantially better than the Hawkes Process at  $2.860 \pm 0.050$ .

**Error Type:** RCA (Referential and Citation Alignment)

**Type:** Cross-Generate

42

D.5 MO (MEASUREMENT & OPERATIONALIZATION)

**Example1: 1015**

**Article** <https://doi.org/10.1038/s41467-025-59630-4>

**Fig. 2 | Comparison of in vivo fluorescence imaging of rats within various NIR bands.** a Schematic design of the PEGylated PEG-Cy5.5 core-shell QDs. b Normalized emission spectra of four kinds of PEG-Cy5.5 in aqueous dispersion. c Normalized absorption spectra of four kinds of PEG-Cy5.5 QDs mixed in a certain ratio in water. The blue line fluorescence images of the window to 1880–2080 nm could significantly suppress the background (Fig. 3), providing imaging performance even better than NIR-II window. Many fluorescent dyes are excreted from the body through hepatic and renal metabolism, and these organs are also where they accumulate<sup>31</sup>. For example, KC, a near-infrared fluorescent dye approved by the Food and Drug Administration, rapidly accumulates in organs like the liver after entering the bloodstream<sup>32</sup>. The dye accumulating in the liver and spleen would create a bright background, and interfere with the detection of targeted vessels above

**Fig. 2 | Comparison of in vivo fluorescence imaging of rats within various NIR bands.** a Schematic design of the PEGylated PEG-Cy5.5 core-shell QDs. b Normalized emission spectra of four kinds of PEG-Cy5.5 in aqueous dispersion. c Normalized absorption spectra of four kinds of PEG-Cy5.5 QDs mixed in a certain ratio in water. The blue line fluorescence images of the window to 1880–2080 nm could significantly suppress the background (Fig. 3), providing imaging performance even better than NIR-II window. Many fluorescent dyes are excreted from the body through hepatic and renal metabolism, and these organs are also where they accumulate<sup>31</sup>. For example, KC, a near-infrared fluorescent dye approved by the Food and Drug Administration, rapidly accumulates in organs like the liver after entering the bloodstream<sup>32</sup>. The dye accumulating in the liver and spleen would create a bright background, and interfere with the detection of targeted vessels above

**Fig. 2 | Comparison of in vivo fluorescence imaging of rats within various NIR bands.** a Schematic design of the PEGylated PEG-Cy5.5 core-shell QDs. b Normalized emission spectra of four kinds of PEG-Cy5.5 in aqueous dispersion. c Normalized absorption spectra of four kinds of PEG-Cy5.5 QDs mixed in a certain ratio in water. The blue line fluorescence images of the window to 1880–2080 nm could significantly suppress the background (Fig. 3), providing imaging performance even better than NIR-II window. Many fluorescent dyes are excreted from the body through hepatic and renal metabolism, and these organs are also where they accumulate<sup>31</sup>. For example, KC, a near-infrared fluorescent dye approved by the Food and Drug Administration, rapidly accumulates in organs like the liver after entering the bloodstream<sup>32</sup>. The dye accumulating in the liver and spleen would create a bright background, and interfere with the detection of targeted vessels above

**Article** <https://doi.org/10.1038/s41467-025-59630-4>

**Fig. 3 | The schematic diagram of photon propagation in tissue and the simulation results of simulating the Monte Carlo method.** a Schematic of photon propagation in biological tissues with high and low light absorption, and the corresponding imaging effects. b The absorption spectrum of water within 700–2300 nm<sup>33</sup>. c Simulated images within 1200–1300, 1300–1400, 1400–1500, 1500–1700, 1700–1880, and 1880–2080 nm, through realistic biological tissue of 1 mm thickness, where the absorption coefficient of water was considered as the tissue absorption coefficient and the setting of tissue scattering coefficient could be found in the “Methods” section. d SBR analysis of the simulation results for each NIR window. Background was added as a single minimum pixel intensity value > 1. e Simulated images within 1200–1300, 1300–1400, 1400–1500, 1500–1700, 1700–1880, and 1880–2080 nm, through realistic biological tissue of 1 mm thickness, where the absorption coefficient of water was considered as the tissue absorption coefficient and the setting of tissue scattering coefficient could be found in the “Methods” section. d SBR analysis of the simulation results for each NIR window. Background was added as a single minimum pixel intensity value > 1. e The structure similarity index measure (SSIM) analysis of the simulation results

**Fig. 3 | The schematic diagram of photon propagation in tissue and the simulation results of simulating the Monte Carlo method.** a Schematic of photon propagation in biological tissues with high and low light absorption, and the corresponding imaging effects. b The absorption spectrum of water within 700–2300 nm<sup>33</sup>. c Simulated images within 1200–1300, 1300–1400, 1400–1500, 1500–1700, 1700–1880, and 1880–2080 nm, through realistic biological tissue of 1 mm thickness, where the absorption coefficient of water was considered as the tissue absorption coefficient and the setting of tissue scattering coefficient could be found in the “Methods” section. d SBR analysis of the simulation results for each NIR window. Background was added as a single minimum pixel intensity value > 1. e The structure similarity index measure (SSIM) analysis of the simulation results

**Fig. 3 | The schematic diagram of photon propagation in tissue and the simulation results of simulating the Monte Carlo method.** a Schematic of photon propagation in biological tissues with high and low light absorption, and the corresponding imaging effects. b The absorption spectrum of water within 700–2300 nm<sup>33</sup>. c Simulated images within 1200–1300, 1300–1400, 1400–1500, 1500–1700, 1700–1880, and 1880–2080 nm, through realistic biological tissue of 1 mm thickness, where the absorption coefficient of water was considered as the tissue absorption coefficient and the setting of tissue scattering coefficient could be found in the “Methods” section. d SBR analysis of the simulation results for each NIR window. Background was added as a single minimum pixel intensity value > 1. e The structure similarity index measure (SSIM) analysis of the simulation results

**Question:** Assess the Figures/Tables section for Measurement & Operationalization issues.

**Explanation:** The figure captions on pages 3 and 4 have been edited to specify that the background for Signal-to-Background Ratio (SBR) calculations was defined as the single minimum pixel intensity in the image. This is not a valid or reliable operationalization of the “background” construct, as it’s highly susceptible to single-point noise or detector artifacts. This flawed measurement procedure systematically undermines all conclusions based on the SBR metric.

**Error Type:** MO (Measurement & Operationalization)

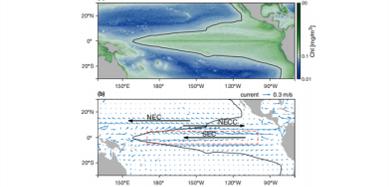
**Type:** Within-Generate

Example2: 1090

Article <https://doi.org/10.1038/s41467-024-5560-8>

(August 2002–July 2022) monthly time series of surface chlorophyll-a concentration (Chl) derived from Moderate-resolution Imaging Spectroradiometer (MODIS) data. The potential forcing underlying CRT changes were discussed based on wind and ocean current data of the same period.

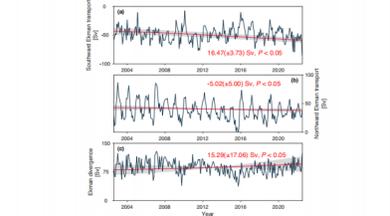
**Results**  
**Westward extension of the chlorophyll-rich tongue in the equatorial Pacific.**  
 Here, we define the boundary of the CRT as the Chl isocline of 0.15 mg/m<sup>3</sup> within the climatology (1982–2002) (Fig. 1a) (see the black curve). Clearly, within the boundary of the CRT are the poleward equatorial upwelling zone and the western IEC (Fig. 1b).  
 The linear trend fitting was first applied to the time series of CRT area, which suggest no statistically significant trend (P > 0.05), see the red line in Fig. 2a. The EMD, which is powerful in detecting trend<sup>38</sup> was then used not to decompose the CRT area time series into trend and interannual components, as well as a residual trend component, as the focus of this analysis is in distinguishing between these variabilities and the residual trend (see “Methods” and Supplementary Fig. 1). Following the approach of Sotillo et al.<sup>38</sup>, we obtained an estimate of the amplitude of the trend, which contributes 9% to the variability of the trend signal, while the interannual and seasonal components explain 39% and 21%, respectively.  
 The seasonal variations of the CRT area highlight peaks of the CRT area in January and a minimum in October–November (see the insert figure in Fig. 3b). This pattern is in general in phase with the known seasonal cycle of the equatorial current system.<sup>39</sup>  
 The interannual variability of the CRT area, the dominant component, is negatively correlated with the Multivariate ENSO Index



**Fig. 1** Distribution of the chlorophyll-rich tongue (CRT) and spatial distribution of ocean current fields in the equatorial Pacific. Chlorophyll-a (Chl) surface concentration (mg/m<sup>3</sup>) and 30-cm current fields (m/s) between August 2002 and July 2022. The CRT (black) boundary is delineated with the black curve, representing the Chl isocline of 0.15 mg/m<sup>3</sup>. The major surface currents are shown in color (black arrows in B), including North Equatorial Current (NEC), North Equatorial Counter Current (NECC) and South Equatorial Current (SEC). The upwelling zone (red shaded area) is defined as the area of positive Ekman divergence (m/s<sup>2</sup>) > 0.05 m/s<sup>2</sup>.

Nature Communications | (2024)15:3033

Article <https://doi.org/10.1038/s41467-024-5560-8>



**Fig. 2** Intensified El Niño divergence in the equatorial Pacific from 2002 to 2022. Monthly variation of (a) the total western El Niño transport calculated at 2% Chl, (b) the total western El Niño transport calculated at 0.1% Chl, and (c) the El Niño divergence. The red line and gray shading represent the long-term trend and the 95% confidence limit, respectively, of the empirical mode decomposition analysis.

In addition, the ENSO-CRT area correlation observed in this effort and prior studies<sup>38,39</sup> suggests that if the occurrences of consecutive La Niña events increase under global warming as projected<sup>40</sup>, the CRT will probably further extend to the west. Of course, the observed long-term trend of the CRT area highly depends on the data span used for trend analysis<sup>38</sup>. Thus, it is necessary to stress that the reported westward extending trend of the CRT is observed within a temporal period of 2002–2022. Nevertheless, our finding of the CRT westward expansion, never reported in previous observational and modeling efforts, highlights the importance of lower frequency variations hidden under major interannual variability of ENSO in the equatorial Pacific. Continued long-term observations spanning over 30 years or more would be essential for a more comprehensive and definitive understanding of the trend of CRT and its impact<sup>38</sup>.

**Methods**  
**Calculation of the area of the high chlorophyll-a tongue.**  
 The standard Level 3 monthly Chl products version R2022.02, at a spatial resolution of 9 km, of Moderate Resolution Imaging Spectroradiometer (MODIS) Ocean Data, August 2002 to July 2022 were acquired from NASA Ocean Color Web (<http://oc2.wr.usgs.gov/oc2v3/>). The computation of the chlorophyll-rich tongue (CRT) area was confined to the open-ocean tropical Pacific region (20°N to 20°S). The eastern boundary was set 1000 km from the shore, and a similar constraint was applied to the western boundary. Pixels with Chl greater than 0.15 mg/m<sup>3</sup> within this region were identified as part of the CRT. The area of each pixel was computed based on its latitude, longitude, Earth radius, and the intervals of latitude and longitude between two adjacent pixels. Subsequently, the CRT area was determined by summing up the area of all pixels within the CRT.

An extended Chl monthly series from September 1997 to August 2023, distributed by the USA Ocean Color project under the Climate Change Initiative (CCI-CCI), was also acquired to supplement our analysis. The CCI-CCI provides merged Chl products at 4 km spatial resolution from several ocean color missions, including MODIS-Aqua, the SeaWiFS Wide Field-of-view Sensor, the Medium Resolution Imaging Spectrometer, and the Ocean and Land Color Instrument<sup>41</sup>. The CCI-CCI monthly Chl products can be directly accessed via <https://www.cci-cci.org/>.

**Definition of the boundaries of the CRT.**  
 The western boundary of the CRT for each month was identified as the longitude of the pixel where the meridional mean Chl between 1°N and 1°S from 20°W to westward first declined to 0.1 mg/m<sup>3</sup>. Note that the western boundary could be absent in some months when Chl in the entire Equatorial Pacific exceeds 0.1 mg/m<sup>3</sup>. In such cases, blank (missing) values for these months. Most blank values were filled using cubic spline interpolation. However, those from May to September 2020, at the end of the time series, remain blank (see Fig. 3a) because it is unreasonable to do extrapolation. For the northern and southern boundaries of CRT in each month, we first computed the zonal mean Chl between 10°W and 10°W and established a latitudinal profile of mean Chl from 20°N to 20°S. We then scanned the latitudinal profile and extracted the two latitudes with mean Chl of 0.1 mg/m<sup>3</sup> as the northern and southern boundaries.

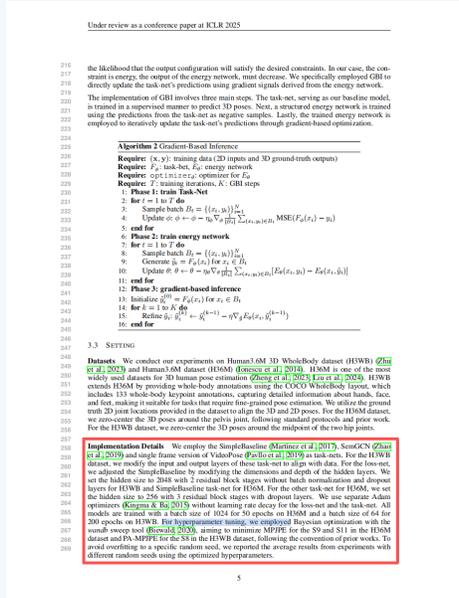
**Empirical mode decomposition for trend analysis.**  
 The Empirical Mode Decomposition (EMD) has demonstrated its efficacy in extracting intrinsic trends and natural variability from diverse data in geographical research<sup>38,39</sup>. Briefly, the original data is decomposed into a finite number of intrinsic Mode Functions (IMFs), which correspond to the scale in the data ranging from the smallest to the largest, and the trend of the original data is defined as the average of the residual components from the decomposed IMFs. There is an upgraded version of the EMD called the Ensemble Empirical Mode Decomposition

Nature Communications | (2024)15:3033

**Question:** Assess the Methods section for Measurement & Operationalization issues.  
**Explanation:** The Results section and the Figure 1 caption define the CRT boundary using a chlorophyll-a (Chl) concentration of 0.15 mg/m<sup>3</sup>. The Methods section also uses this 0.15 mg/m<sup>3</sup> threshold for the western boundary. However, the same Methods section then defines the northern and southern boundaries using a different threshold of 0.1 mg/m<sup>3</sup>, creating an inconsistent operational definition for the paper’s primary construct.  
**Error Type:** MO (Measurement & Operationalization)  
**Type:** Within-Generate

## D.6 DHP (DATA HANDLING & PREPROCESSING)

### Example1: 528



**Question:** Assess the Methods section for Data Handling & Preprocessing issues.

**Explanation:** The reviewer correctly identifies that the authors tuned hyperparameters on the test set. The paper’s “Implementation Details” section on page 5 states: “For hyperparameter tuning, we employed Bayesian optimization with the wandb sweep tool (Biewald, 2020), aiming to minimize MPJPE for the S9 and S11 in the H36M dataset and PA-MPJPE for the S8 in the H3WB dataset, following the convention of prior works.” According to standard protocols for the H36M dataset, subjects S9 and S11 constitute the test set. Tuning hyperparameters directly on the test set introduces data leakage, leading to an optimistic bias in the reported results and invalidating claims of generalization. This is a critical violation of machine learning best practices and fits the Data Handling & Preprocessing (E) category, as a pipeline choice introduces bias.

**Error Type:** DHP (Data Handling & Preprocessing)

**Type:** Within-Sample

Example2: 1566

Article

https://doi.org/10.26434/chemrxiv-2025-40936-v6

hub trusts), many of which opened during lockdown, typically in response to the suspension of planned services during the lockdown. We examined impacts on all non-emergency surgery and, assuming that hubs focus on HVLC elective procedures, we examined HVLC elective surgery separately to better understand the specific contribution of hubs to the overall effect observed across each trust.

**Results**  
A total of 1,077,669 elective surgeries at 71 new-hub trusts, 23 established-hub trusts and 54 trusts without a hub (non-hub trusts) between April 2020 and December 2022 were included in our study, with data for the lockdown months (April 2020 to March 2021) imputed using pre-pandemic average rates to ensure continuity when Supplementary Fig. 1 and 2. An earlier response and effect on hubs as England moved into the second wave across NHS regions in England, mostly in more densely populated areas, with significant variation in the histology, specialties, care provision, and their healthcare provider contracts. Characteristics of trust catchment populations for trusts with and without hubs were broadly similar. However, trusts without hubs (non-hub trusts) tended to have smaller median catchment population size (279,229 vs 400,038 and 226,201) and larger median proportion of individuals with White ethnicity (95% vs 85% and 86%) (Table 1) (supplemental and established-hub trusts respectively (Table 1). The rate of total elective surgery and HVLC elective surgery (number referred to as total-surgery and HVLC-surgery respectively) were consistently higher in non-hub trusts than in new-hub and established-hub trusts throughout extended periods (Fig. 2). In the recovery leading up to the pandemic, the average total-surgery rate at non-hub trusts (447 surgeries per 1000 total catchment population per month (SD = 1.6)), was 8% higher than in new-hub trusts (41.3 (SD = 1.4)) and 9% higher than in established-hub trusts (41.2) (Table 1). During lockdown, the average total-surgery and HVLC-surgery rates decreased across almost all trusts, but larger decreases were observed in non-hub (17.9% vs total and -22.1% for HVLC) than in both new-hub (-13.8% and -20.3%) and established-hub trusts (-13.8% and -15.9%), resulting rates more similar across all types of trust post-pandemic compared with pre-pandemic.

Excluding the COVID-19 lockdown period, approximately 65-75% of all surgeries were treated as day-case with established-hub trusts consistently having the low-coverage versus alternative but significant differences between trust types for HVLC-surgery (Supplementary Fig. 3). Post-pandemic, non-hub trusts had the shortest average length of hospital stay thereafter reduced to an length of stay for all surgeries, including HVLC (Supplementary Fig. 4). Post-pandemic, non-hub established-hub trusts also had the greatest reduction in length of stay.

**New-hub trusts**

Applying the generalised synthetic control model, estimated effects for the impact of opening a new hub on total-surgery and HVLC surgery rates during the first 12 months were similar (each estimated an additional 0.17-0.18 surgeries per 1000 total catchment population per month), although these were only significant for HVLC-surgery (Table 2). The HVLC-surgery rate was 0.097 surgeries per 1000 trust

catchment population per month (95% confidence interval = 0.090 to 0.240) higher in new-hub trusts during the first 12 months compared with the synthetic control. This is equivalent to a 2.9% (95% CI = 1.7% to 3.2%) higher than expected increase in HVLC-surgery in new-hub trusts when compared with a synthetic control with similar pre-opening trends.

Figure 3 displays observed and counterfactual trends for total-surgery and HVLC-surgery rates (left panels) as well as corresponding impact estimates (right panels) at new-hub trusts. Throughout the pre-intervention period, rates of both total-surgery and HVLC-surgery matched the synthetic control, indicating a good model fit. However, these rates steadily declined, reflecting that the pre-intervention months for many new-hub trusts were just before lockdown, when surgeries were closed (Fig. 3). Positive effects on HVLC-surgery rates were only significant after three months and were driven by a continued reduction in the synthetic control, compared to a marked increase at new-hub trusts.

For all surgeries, the day-case proportion was an average of 0.020 (95% CI = 0.002 to 0.038) higher in new-hub trusts compared to the synthetic control during the first 12 months of opening. There was no significant effect on the day-case proportion for HVLC-surgery or on length of stay for either total-surgery or HVLC-surgery in new-hub trusts during this period (Supplementary Figs. 5, 6).

**Established-hub trusts**

For established-hub trusts, the total-surgery rate was 0.374 (95% CI = 0.483 to 0.270) surgeries per 1000 total catchment population per month higher between April 2020 and December 2022 compared with the synthetic control (Table 2). This corresponds to an average increase of 12.2% (95% CI = 1.3% to 21.2%) in all surgery post-lockdown, due to having an established hub. Similar trends were also seen for the HVLC-surgery rate with an average increase of 12.2% (95% CI = 1.7% to 20.7%) during the post-lockdown period compared with the synthetic control.

As for new-hub trusts, both total-surgery and HVLC-surgery rates matched the synthetic control during the pre-intervention period indicating a good model fit (Fig. 4). Post-intervention, the positive effects on total-surgery and HVLC-surgery rates steadily increased, driven by a linear initial increase in rates in the established-hub trusts not matched over the study period in the estimated synthetic control.

The average length of stay for HVLC-surgery was 0.076 day (95% CI = -0.280 to -0.040) shorter in established-hub trusts than in the synthetic control. There were no significant effects on length of stay across all surgery types at established-hub trusts, nor on the proportion of day cases for either total-surgery or HVLC-surgery (Supplementary Figs. 7, 8).

**Specialties and individual trusts**

When examining each HVLC specialty in turn, we only found a significant effect on the rate of general surgery at new-hub trusts and no significant effect on any individual specialties at established-hub trusts. Overall surgery comprises 17% of all HVLC elective surgery in new-hub trusts and increased by 0.023 surgeries per 1000 total catchment population per month (95% CI = 0.007 to 0.039) during the

**Table 1 | Comparison of population characteristics in new-hub, established-hub and non-hub trusts in England in 2019**

	New-hub trusts (N=71) Mean (SD)	Established-hub trusts (N=23) Mean (SD)	Non-hub trusts (N=54) Mean (SD)
Population from a White ethnic background (%)	85 (77, 92)	86 (77, 92)	95 (84, 96)
Population where male (%)	48 (48, 50)	50 (48, 50)	48 (48, 50)
Population who are aged 65 or over (%)	19 (19, 20)	19 (19, 20)	20 (19, 20)
Catchment population	400,038 (206, 310,448)	408,358 (204,916, 402,270)	279,420 (99,363, 468,008)

Nature Communications | (2025) 16:4182 | 3

**Question:** Assess the Methods section for Data Handling & Preprocessing issues.

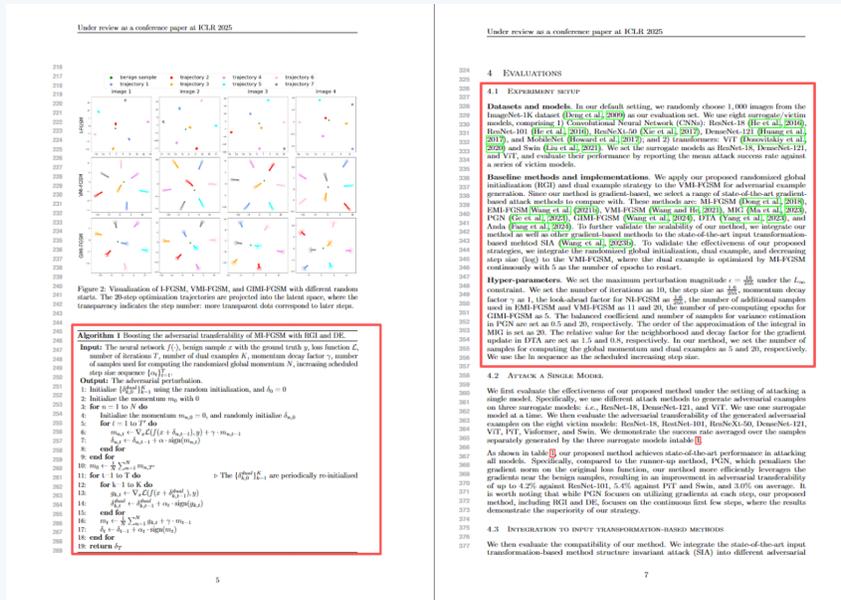
**Explanation:** The modified text on page 3 states that data for the COVID-19 lockdown period were imputed using pre-pandemic averages. This data handling choice is highly problematic, as it smooths over a major, non-random structural break in the time series rather than modeling or excluding it. The imputation method introduces significant bias and data leakage, as a simple average does not accurately reflect the known, drastic reduction in elective surgeries during that specific period, compromising the validity of the causal model.

**Error Type:** DHP (Data Handling & Preprocessing)

**Type:** Within-Generate

## D.7 CF (COMPUTATION &amp; FORMULAE)

Example: 1909



**Question:** Scan the Methods section for Computation & Formulae errors.

**Explanation:** Algorithm 1 on page 5 uses the parameter  $T'$  in the loop definition on line 5: for  $t = 1$  to  $T'$  do. This parameter determines the number of iterations for the Randomized Global Initialization phase. However, the value of  $T'$  is never specified anywhere in the paper, including the ‘‘Hyper-parameters’’ section (Section 4.1 on page 7). An algorithm cannot be implemented or reproduced with an undefined critical parameter. This fits the Computation & Formulae category as an ‘‘undefined variable’’.

**Error Type:** CF (Computation & Formulae)

**Type:** Within-Sample

D.8 IC (INFERENCE & CONCLUSIONS)

Example: 875

Under review as a conference paper at ICLR 2025

FEDGRAPH: A NEW PARADIGM FOR FEDERATED GRAPH LEARNING

Newayam authors  
Paper with specific structure

ABSTRACT

Federated learning is a distributed approach to training global models over multiple clients without sharing their local data. In graph data, the data heterogeneity can complicate graph structure and node features varying extremely different, and the model training of federated graph learning is more challenging in practice. In this paper, we propose FedGraph, a new paradigm for federated graph learning. The key idea is to utilize the graph structure without getting node features as a source knowledge to help of local specific knowledge in clients. The extensive experiments show that FedGraph significantly outperforms the state-of-the-art of federated learning algorithms in anomaly detection tasks. The deep learning models and one existing anomaly detection model are implemented in FedGraph framework.

1 INTRODUCTION

Graph applications are becoming increasingly prevalent in social activities, such as account recommendation, fraud detection, and short video ranking on Twitter, LinkedIn, and TikTok. Most applications developed by these companies follow personalized content to users based on user's personalized interests inferred from their private data. In our social network service (e.g., WeChat), users can obtain similar behaviors, such as browsing on "WeChat Moments" when they read their similar interest distribution. The cross data from WeChat also has the personal interest distribution in a real scenario. In this context, data privacy regulation (GDPR) requires the collection, storage, and processing of personal data, and an effort on most major companies worldwide (Kumar et al., 2020).

However, an important challenge caused from graph data heterogeneity is tackled by recent Federated Graph Learning (FGL) methods (Khan et al., 2020; Wang et al., 2020). In graph data, the number of nodes and edges varies independently on different clients (DCI) properties, and the graph structure also exhibits more heterogeneity. In the federated learning (FL) setting, most Graph Neural Networks (GNNs) consider the structure consistency, and the structure heterogeneity is tackled by the personalized neighbor parameters of the deep models (Khan et al., 2020; Wang et al., 2020). However, on the one hand, the graph structure GNN-based models need more significant communication costs. Additionally, traditional GNN models cannot be directly used for federated learning.

Moreover, a significant challenge, model inference in federated graph task, has been overlooked by the existing FL methods since graph learning models can be only adapted to different graph structures. The evolution of graph task and computation specific algorithms are essential to graph performing federated tasks across multiple graph data, referred to as model inference in federated graph task. In FedGraph, this model is used in the form of a central client (CC) (Khan et al., 2020) to meet federated tasks, and the task of detecting anomaly events on platform such as Twitter and WeChat is implemented for the specific knowledge of the cross application structure of the community and users' behaviors in other related fields which helps to prevent algorithm overfitting process.

Under review as a conference paper at ICLR 2025

Figure 1: An example of a model inference in federated graph task. In large-scale graph learning on WeChat and Weibo, the "CC" (client) is an extremely popular topic on Twitter, Facebook, and Weibo, the top trending topic in the "stock market". Due to data isolation and the privacy of their algorithms, federated federated learning is a better performing federated task.

In traditional federated graph learning, multiple tasks cannot be performed collaboratively if graph data and algorithm are not private. Graph data can be categorized into structure and feature. We present graph feature and algorithm co-optimization process, but the graph structure can be shared. To federated graph learning task and graph data for the graph structure.

Moreover, these federated learning process were proposed for federated graph learning for the graph structure as a help to private with feature in the practice. Knowledge refers to adaptively transfered methods. For instance, a federated graph task can be implemented in multiple devices, allowing the use of exclusive graph learning methods within the organization. We take the graph learning structure as a shared knowledge, which can communicate with other data owners. We present node feature privacy across all data fields. In this new paradigm, we consider node feature privacy and the graph structure of the server can be securely shared. For example, a client may recommend a set of users as an influence on Twitter, however, these users may remain anonymous to other clients for the privacy of the server. The graph data anomaly detection task can be implemented in the federated setting, anomaly detection will be considered as a downstream task in the following text. For anomaly detection on multiple graphs, model optimization has been the federated anomaly detection rather data leakage (Khan et al., 2020; Wang et al., 2020). We measure our contributions as follows:

- Innovative work.** To the best of our knowledge, we are the first to make federated graph learning performance optimization on parameters and task performance exclusively dependent on clients, referred to as model inference in federated graph task. We need employ the graph structure helping federated models (FedGraph).
- Broad application.** A broad spectrum of graph learning models can be easily transferred to FedGraph framework. Specifically, the graph learning models, such as graph neural networks (GNN) models (e.g., DeepWalk, GCN), the models are successfully transferred to FedGraph framework anomaly detection tasks.
- Effectiveness and robustness.** Extensive experiments on the real datasets have verified that anomaly tasks on FedGraph can be effectively applied to different multiple graph settings. The proposed FedGraph achieved a best 95% accuracy under the 30% same level.

Under review as a conference paper at ICLR 2025

Table 1: Accuracy evaluation of our federated graph learning algorithm and baselines on Cora, Citicites, and ACM datasets, the mean average of these datasets on 27, 31, 19, 38, 45, 49. All results are based on the top-1 results identified as anomalies, with a 10% "Anomaly Rate" and anomaly rate (top-1 results).

Dataset	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50																																									
Corat	0.10	0.11	0.12	0.13	0.14	0.15	0.16	0.17	0.18	0.19	0.20	0.21	0.22	0.23	0.24	0.25	0.26	0.27	0.28	0.29	0.30	0.31	0.32	0.33	0.34	0.35	0.36	0.37	0.38	0.39	0.40	0.41	0.42	0.43	0.44	0.45	0.46	0.47	0.48	0.49	0.50	0.51	0.52	0.53	0.54	0.55	0.56	0.57	0.58	0.59	0.60	0.61	0.62	0.63	0.64	0.65	0.66	0.67	0.68	0.69	0.70	0.71	0.72	0.73	0.74	0.75	0.76	0.77	0.78	0.79	0.80	0.81	0.82	0.83	0.84	0.85	0.86	0.87	0.88	0.89	0.90	0.91	0.92	0.93	0.94	0.95	0.96	0.97	0.98	0.99	1.00

Table 2: Evaluation of our federated graph learning method (FedGraph) on baselines on the CoraNet dataset, with the number of clients as 1, 2, and 3.

Method	Accuracy	Privacy	Robustness
FEDGRAPH	95.0%	95.0%	95.0%
GCN	85.0%	85.0%	85.0%
DeepWalk	75.0%	75.0%	75.0%
PageRank	65.0%	65.0%	65.0%
Adaboost	55.0%	55.0%	55.0%
LogitBoost	45.0%	45.0%	45.0%
AdaBoost	35.0%	35.0%	35.0%
AdaBoost.M1	25.0%	25.0%	25.0%
AdaBoost.M2	15.0%	15.0%	15.0%
AdaBoost.M3	5.0%	5.0%	5.0%
AdaBoost.M4	0.0%	0.0%	0.0%
AdaBoost.M5	0.0%	0.0%	0.0%
AdaBoost.M6	0.0%	0.0%	0.0%
AdaBoost.M7	0.0%	0.0%	0.0%
AdaBoost.M8	0.0%	0.0%	0.0%
AdaBoost.M9	0.0%	0.0%	0.0%
AdaBoost.M10	0.0%	0.0%	0.0%
AdaBoost.M11	0.0%	0.0%	0.0%
AdaBoost.M12	0.0%	0.0%	0.0%
AdaBoost.M13	0.0%	0.0%	0.0%
AdaBoost.M14	0.0%	0.0%	0.0%
AdaBoost.M15	0.0%	0.0%	0.0%
AdaBoost.M16	0.0%	0.0%	0.0%
AdaBoost.M17	0.0%	0.0%	0.0%
AdaBoost.M18	0.0%	0.0%	0.0%
AdaBoost.M19	0.0%	0.0%	0.0%
AdaBoost.M20	0.0%	0.0%	0.0%
AdaBoost.M21	0.0%	0.0%	0.0%
AdaBoost.M22	0.0%	0.0%	0.0%
AdaBoost.M23	0.0%	0.0%	0.0%
AdaBoost.M24	0.0%	0.0%	0.0%
AdaBoost.M25	0.0%	0.0%	0.0%
AdaBoost.M26	0.0%	0.0%	0.0%
AdaBoost.M27	0.0%	0.0%	0.0%
AdaBoost.M28	0.0%	0.0%	0.0%
AdaBoost.M29	0.0%	0.0%	0.0%
AdaBoost.M30	0.0%	0.0%	0.0%
AdaBoost.M31	0.0%	0.0%	0.0%
AdaBoost.M32	0.0%	0.0%	0.0%
AdaBoost.M33	0.0%	0.0%	0.0%
AdaBoost.M34	0.0%	0.0%	0.0%
AdaBoost.M35	0.0%	0.0%	0.0%
AdaBoost.M36	0.0%	0.0%	0.0%
AdaBoost.M37	0.0%	0.0%	0.0%
AdaBoost.M38	0.0%	0.0%	0.0%
AdaBoost.M39	0.0%	0.0%	0.0%
AdaBoost.M40	0.0%	0.0%	0.0%
AdaBoost.M41	0.0%	0.0%	0.0%
AdaBoost.M42	0.0%	0.0%	0.0%
AdaBoost.M43	0.0%	0.0%	0.0%
AdaBoost.M44	0.0%	0.0%	0.0%
AdaBoost.M45	0.0%	0.0%	0.0%
AdaBoost.M46	0.0%	0.0%	0.0%
AdaBoost.M47	0.0%	0.0%	0.0%
AdaBoost.M48	0.0%	0.0%	0.0%
AdaBoost.M49	0.0%	0.0%	0.0%
AdaBoost.M50	0.0%	0.0%	0.0%

5.2 EXPERIMENTAL RESULTS

The feasibility of our federated graph learning framework allows us to compare the accuracy performance on our experiment into two sections: same graph structure comparison (Table 1) and different graph structure comparison (Table 2). The same graph structure comparison is conducted on the Cora, Citicites, and ACM datasets, while the different graph structure comparison is performed on the CoraNet dataset. It is apparent to see that the corresponding baseline offers two types of comparison.

5.2.1 EVALUATION ON SAME GRAPH STRUCTURE METRICS

In this evaluation experiment, we use FedGraph and FedGraph (G.D.) to compare against baselines. The accuracy metrics, shown in Table 1, are calculated with anomalies related to the top 10. We divide datasets into 1, 2, 3, and 5 clients to compare accuracy with varying client numbers. The approach consistently achieves the highest accuracy across all clients and datasets. Next, we use the dataset with G.D. FedGraph (G.D.) as a baseline as a source of G.D., implementing the competitive baseline FedGraph (G.D.) by the FedGraph (G.D.). Our approach consistently achieves the highest accuracy.

On the other hand, we compare FedGraph and FedGraph (G.D.) against baselines. It is possible that the best accuracy degree of the two datasets. After dividing the dataset into several parts, we compare FedGraph and FedGraph (G.D.) against baselines. The results show that the FedGraph (G.D.) consistently achieves the highest accuracy across all clients and datasets. Next, we use the dataset with G.D. FedGraph (G.D.) as a baseline as a source of G.D., implementing the competitive baseline FedGraph (G.D.) by the FedGraph (G.D.). Our approach consistently achieves the highest accuracy.

5.2.2 EVALUATION ON DIFFERENT GRAPH STRUCTURE METRICS

To evaluate the effectiveness of FedGraph on various graph structure metrics, we split the CoraNet dataset into many subsets to clients. Besides FedGraph, all baseline methods are one deep learning anomaly detection algorithm. In this experiment, we compare FedGraph (G.D.) against baselines, achieving the highest accuracy over 90% during nearly all anomalies, and surpassing the best baseline by 2%. Additionally, our accuracy results the competitive baseline by 25.75% (4.75%), with an average anomaly rate (AR) higher than the baseline by 1.6 times (1.75%).

**Question:** Evaluate Abstract, Introduction and Experiment section for issues in Inference & Conclusions.

**Explanation:** The paper’s claims of generality are not supported by its evidence. The title and abstract introduce “FedGraph: A New Paradigm for Federated Graph Learning” (page 1), suggesting a broadly applicable framework. However, the methodology is heavily tailored to, and the experiments are exclusively focused on, the single downstream task of anomaly detection. For example, a stated contribution is “Broad application,” but this is immediately qualified with “the models are successfully transferred to FEDGRAPH framework in anomaly detection tasks” (page 2). Furthermore, Section 5, “EXPERIMENTS”, exclusively reports results on anomaly detection tasks. This discrepancy represents an issue of Inference & Conclusions, as the broad conclusion of having created a new “paradigm” for FGL is an overstatement that exceeds what the narrow experimental results can support.

**Error Type:** IC (Inference & Conclusions)  
**Type:** Within-Sample

### D.9 LE (LANGUAGE & EXPRESSION)

#### Example: 1785

Published as a conference paper at ICLR 2025

• We want to support the use of hardware-accelerated instructions (e.g., asynchronous matrix multiply and bulk copy instructions), which also require specific shared memory layouts. In TK, we simplify to 3 layouts – swizzled on 32, 64, and 128 byte boundaries – and automatically assign shared tiles with layouts that minimize bank conflicts for their size and type. Seen in Section 2.3, even the FlashAttention 3 kernels written with CUTLASS templates can face bank conflicts, hurting performance. Our approach helps minimize conflicts, reducing  $C_{bank}$  in Section 2.3.

**3.2 BLOCK PARALLELISM WITH A GENERALIZED ASYNCHRONOUS TEMPLATE**

**THRESHOLDING** helps developers reduce overheads by coordinating how workers in a thread block asynchronously overlap execution. Through the GPU hierarchy might suggest that we need a wide variety of techniques, we propose a single concise template that we find enables high performance on a surprisingly broad range of AI workloads. We first define the template, which has four steps – load-compute-store-finish (LCSF) for short – and build on the classical producer-consumer paradigm (Dijkstra, 1968; Bauer et al., 2011). We show how the LCSF template carefully navigates the tradeoffs between occupancy and efficiency (including CLM,  $C_{bank}$  in Section 2).

Load function:	Compute function:
<pre> 1 load_asyncmem_ptr(0, 0, 0) 2 load_asyncmem_ptr(0, 0, 0) 3 load_asyncmem_ptr(0, 0, 0) 4 load_asyncmem_ptr(0, 0, 0) 5 load_asyncmem_ptr(0, 0, 0) 6 load_asyncmem_ptr(0, 0, 0) 7 load_asyncmem_ptr(0, 0, 0) 8 load_asyncmem_ptr(0, 0, 0) 9 load_asyncmem_ptr(0, 0, 0) 10 load_asyncmem_ptr(0, 0, 0) 11 load_asyncmem_ptr(0, 0, 0) 12 load_asyncmem_ptr(0, 0, 0) 13 load_asyncmem_ptr(0, 0, 0) 14 load_asyncmem_ptr(0, 0, 0) 15 load_asyncmem_ptr(0, 0, 0) 16 load_asyncmem_ptr(0, 0, 0) 17 load_asyncmem_ptr(0, 0, 0) 18 load_asyncmem_ptr(0, 0, 0) 19 load_asyncmem_ptr(0, 0, 0) 20 load_asyncmem_ptr(0, 0, 0) 21 load_asyncmem_ptr(0, 0, 0) 22 load_asyncmem_ptr(0, 0, 0) 23 load_asyncmem_ptr(0, 0, 0) 24 load_asyncmem_ptr(0, 0, 0) 25 load_asyncmem_ptr(0, 0, 0) 26 load_asyncmem_ptr(0, 0, 0) 27 load_asyncmem_ptr(0, 0, 0) 28 load_asyncmem_ptr(0, 0, 0) 29 load_asyncmem_ptr(0, 0, 0) 30 load_asyncmem_ptr(0, 0, 0) 31 load_asyncmem_ptr(0, 0, 0) 32 load_asyncmem_ptr(0, 0, 0) 33 load_asyncmem_ptr(0, 0, 0) 34 load_asyncmem_ptr(0, 0, 0) 35 load_asyncmem_ptr(0, 0, 0) 36 load_asyncmem_ptr(0, 0, 0) 37 load_asyncmem_ptr(0, 0, 0) 38 load_asyncmem_ptr(0, 0, 0) 39 load_asyncmem_ptr(0, 0, 0) 40 load_asyncmem_ptr(0, 0, 0) 41 load_asyncmem_ptr(0, 0, 0) 42 load_asyncmem_ptr(0, 0, 0) 43 load_asyncmem_ptr(0, 0, 0) 44 load_asyncmem_ptr(0, 0, 0) 45 load_asyncmem_ptr(0, 0, 0) 46 load_asyncmem_ptr(0, 0, 0) 47 load_asyncmem_ptr(0, 0, 0) 48 load_asyncmem_ptr(0, 0, 0) 49 load_asyncmem_ptr(0, 0, 0) 50 load_asyncmem_ptr(0, 0, 0) 51 load_asyncmem_ptr(0, 0, 0) 52 load_asyncmem_ptr(0, 0, 0) 53 load_asyncmem_ptr(0, 0, 0) 54 load_asyncmem_ptr(0, 0, 0) 55 load_asyncmem_ptr(0, 0, 0) 56 load_asyncmem_ptr(0, 0, 0) 57 load_asyncmem_ptr(0, 0, 0) 58 load_asyncmem_ptr(0, 0, 0) 59 load_asyncmem_ptr(0, 0, 0) 60 load_asyncmem_ptr(0, 0, 0) 61 load_asyncmem_ptr(0, 0, 0) 62 load_asyncmem_ptr(0, 0, 0) 63 load_asyncmem_ptr(0, 0, 0) 64 load_asyncmem_ptr(0, 0, 0) 65 load_asyncmem_ptr(0, 0, 0) 66 load_asyncmem_ptr(0, 0, 0) 67 load_asyncmem_ptr(0, 0, 0) 68 load_asyncmem_ptr(0, 0, 0) 69 load_asyncmem_ptr(0, 0, 0) 70 load_asyncmem_ptr(0, 0, 0) 71 load_asyncmem_ptr(0, 0, 0) 72 load_asyncmem_ptr(0, 0, 0) 73 load_asyncmem_ptr(0, 0, 0) 74 load_asyncmem_ptr(0, 0, 0) 75 load_asyncmem_ptr(0, 0, 0) 76 load_asyncmem_ptr(0, 0, 0) 77 load_asyncmem_ptr(0, 0, 0) 78 load_asyncmem_ptr(0, 0, 0) 79 load_asyncmem_ptr(0, 0, 0) 80 load_asyncmem_ptr(0, 0, 0) 81 load_asyncmem_ptr(0, 0, 0) 82 load_asyncmem_ptr(0, 0, 0) 83 load_asyncmem_ptr(0, 0, 0) 84 load_asyncmem_ptr(0, 0, 0) 85 load_asyncmem_ptr(0, 0, 0) 86 load_asyncmem_ptr(0, 0, 0) 87 load_asyncmem_ptr(0, 0, 0) 88 load_asyncmem_ptr(0, 0, 0) 89 load_asyncmem_ptr(0, 0, 0) 90 load_asyncmem_ptr(0, 0, 0) 91 load_asyncmem_ptr(0, 0, 0) 92 load_asyncmem_ptr(0, 0, 0) 93 load_asyncmem_ptr(0, 0, 0) 94 load_asyncmem_ptr(0, 0, 0) 95 load_asyncmem_ptr(0, 0, 0) 96 load_asyncmem_ptr(0, 0, 0) 97 load_asyncmem_ptr(0, 0, 0) 98 load_asyncmem_ptr(0, 0, 0) 99 load_asyncmem_ptr(0, 0, 0) 100 load_asyncmem_ptr(0, 0, 0) </pre>	<pre> 1 compute_asyncmem_ptr(0, 0, 0) 2 compute_asyncmem_ptr(0, 0, 0) 3 compute_asyncmem_ptr(0, 0, 0) 4 compute_asyncmem_ptr(0, 0, 0) 5 compute_asyncmem_ptr(0, 0, 0) 6 compute_asyncmem_ptr(0, 0, 0) 7 compute_asyncmem_ptr(0, 0, 0) 8 compute_asyncmem_ptr(0, 0, 0) 9 compute_asyncmem_ptr(0, 0, 0) 10 compute_asyncmem_ptr(0, 0, 0) 11 compute_asyncmem_ptr(0, 0, 0) 12 compute_asyncmem_ptr(0, 0, 0) 13 compute_asyncmem_ptr(0, 0, 0) 14 compute_asyncmem_ptr(0, 0, 0) 15 compute_asyncmem_ptr(0, 0, 0) 16 compute_asyncmem_ptr(0, 0, 0) 17 compute_asyncmem_ptr(0, 0, 0) 18 compute_asyncmem_ptr(0, 0, 0) 19 compute_asyncmem_ptr(0, 0, 0) 20 compute_asyncmem_ptr(0, 0, 0) 21 compute_asyncmem_ptr(0, 0, 0) 22 compute_asyncmem_ptr(0, 0, 0) 23 compute_asyncmem_ptr(0, 0, 0) 24 compute_asyncmem_ptr(0, 0, 0) 25 compute_asyncmem_ptr(0, 0, 0) 26 compute_asyncmem_ptr(0, 0, 0) 27 compute_asyncmem_ptr(0, 0, 0) 28 compute_asyncmem_ptr(0, 0, 0) 29 compute_asyncmem_ptr(0, 0, 0) 30 compute_asyncmem_ptr(0, 0, 0) 31 compute_asyncmem_ptr(0, 0, 0) 32 compute_asyncmem_ptr(0, 0, 0) 33 compute_asyncmem_ptr(0, 0, 0) 34 compute_asyncmem_ptr(0, 0, 0) 35 compute_asyncmem_ptr(0, 0, 0) 36 compute_asyncmem_ptr(0, 0, 0) 37 compute_asyncmem_ptr(0, 0, 0) 38 compute_asyncmem_ptr(0, 0, 0) 39 compute_asyncmem_ptr(0, 0, 0) 40 compute_asyncmem_ptr(0, 0, 0) 41 compute_asyncmem_ptr(0, 0, 0) 42 compute_asyncmem_ptr(0, 0, 0) 43 compute_asyncmem_ptr(0, 0, 0) 44 compute_asyncmem_ptr(0, 0, 0) 45 compute_asyncmem_ptr(0, 0, 0) 46 compute_asyncmem_ptr(0, 0, 0) 47 compute_asyncmem_ptr(0, 0, 0) 48 compute_asyncmem_ptr(0, 0, 0) 49 compute_asyncmem_ptr(0, 0, 0) 50 compute_asyncmem_ptr(0, 0, 0) 51 compute_asyncmem_ptr(0, 0, 0) 52 compute_asyncmem_ptr(0, 0, 0) 53 compute_asyncmem_ptr(0, 0, 0) 54 compute_asyncmem_ptr(0, 0, 0) 55 compute_asyncmem_ptr(0, 0, 0) 56 compute_asyncmem_ptr(0, 0, 0) 57 compute_asyncmem_ptr(0, 0, 0) 58 compute_asyncmem_ptr(0, 0, 0) 59 compute_asyncmem_ptr(0, 0, 0) 60 compute_asyncmem_ptr(0, 0, 0) 61 compute_asyncmem_ptr(0, 0, 0) 62 compute_asyncmem_ptr(0, 0, 0) 63 compute_asyncmem_ptr(0, 0, 0) 64 compute_asyncmem_ptr(0, 0, 0) 65 compute_asyncmem_ptr(0, 0, 0) 66 compute_asyncmem_ptr(0, 0, 0) 67 compute_asyncmem_ptr(0, 0, 0) 68 compute_asyncmem_ptr(0, 0, 0) 69 compute_asyncmem_ptr(0, 0, 0) 70 compute_asyncmem_ptr(0, 0, 0) 71 compute_asyncmem_ptr(0, 0, 0) 72 compute_asyncmem_ptr(0, 0, 0) 73 compute_asyncmem_ptr(0, 0, 0) 74 compute_asyncmem_ptr(0, 0, 0) 75 compute_asyncmem_ptr(0, 0, 0) 76 compute_asyncmem_ptr(0, 0, 0) 77 compute_asyncmem_ptr(0, 0, 0) 78 compute_asyncmem_ptr(0, 0, 0) 79 compute_asyncmem_ptr(0, 0, 0) 80 compute_asyncmem_ptr(0, 0, 0) 81 compute_asyncmem_ptr(0, 0, 0) 82 compute_asyncmem_ptr(0, 0, 0) 83 compute_asyncmem_ptr(0, 0, 0) 84 compute_asyncmem_ptr(0, 0, 0) 85 compute_asyncmem_ptr(0, 0, 0) 86 compute_asyncmem_ptr(0, 0, 0) 87 compute_asyncmem_ptr(0, 0, 0) 88 compute_asyncmem_ptr(0, 0, 0) 89 compute_asyncmem_ptr(0, 0, 0) 90 compute_asyncmem_ptr(0, 0, 0) 91 compute_asyncmem_ptr(0, 0, 0) 92 compute_asyncmem_ptr(0, 0, 0) 93 compute_asyncmem_ptr(0, 0, 0) 94 compute_asyncmem_ptr(0, 0, 0) 95 compute_asyncmem_ptr(0, 0, 0) 96 compute_asyncmem_ptr(0, 0, 0) 97 compute_asyncmem_ptr(0, 0, 0) 98 compute_asyncmem_ptr(0, 0, 0) 99 compute_asyncmem_ptr(0, 0, 0) 100 compute_asyncmem_ptr(0, 0, 0) </pre>

Figure 5: A simplified depiction of attention in the LCSF template to highlight the role of different specialized workers. Left is executed by workers that manage HBM to SRAM memory movement, and right by parallel compute workers, which operate in fast memory, registers and SRAM.

**Programming abstractions** As per Section 2.3, AI kernel usually load tiles of large tensors from HBM to SRAM, perform computation in fast memory, store the result for the tile back to HBM, and repeat this for the next tiles. To use the LCSF template, the developer writes four functions:

- Load function.** Specifies the data that load workers should load from HBM to shared memory, and when to signal to compute workers that this memory is ready for use.
- Compute function.** Specifies the kernel instructions that compute workers should execute, using the data structure and operation primitives from Section 2.3.
- Store function.** Specifies what data workers need to store to HBM.
- Finish function.** At the end of the kernel, the workers store any final state and exit.

TK provides abstractions to help the developer manage worker overlapping and synchronization.

M × N × K	Stages	THROPS
4096	1	260
4096	2	484
4096	3	653
4096	4	790

Table 1: Pipeline buffer stages. We would need to wait for all compute workers to finish measuring efficiency in THROPS for our counting before replacing the input tile. A 2-stage buffer GEMM kernels as we vary the number of can hide the HBM load (store) latency since the next pipeline buffer stages in the TK template, tile can asynchronously load, while the compute workers execute on the current tile. Deep buffers can reduce the synchronization required across compute workers, allowing them to operate on multiple tiles concurrently. TK lets the user set a simple number to specify the number of stages, and manages the setup and use of these buffers for the user. In Section 2.3 we vary the number of stages  $N \in \{1, 2, 3, 4\}$  for our GEMM kernel.

Published as a conference paper at ICLR 2025

2. Synchronization barriers. Load/store workers need to alert compute workers when new memory is written to the input buffer. Compute workers need to alert load/store workers when tiles are written to the output buffer, or when input tiles can be evicted from the input buffer. Within the TK template, we provide an `arrive` function for workers to signal that they have finished their stage.

3. Asynchronous IO. We wrap synchronous and asynchronous load and store instructions, including `load_asyncmem_ptr` and `store_asyncmem_ptr`, in the same interface. We abstract tensor map descriptor creation for TMA hardware-accelerated address generation for our global layout descriptors (q.1).

**Tradeoffs between occupancy and efficiency**

TK partitions the number of load/store and compute workers (or occupancy) providing a simple way for developers tune their kernel. As discussed in Section 2.3, higher occupancy increases overlapping, but creates contention over limited hardware resources (e.g., registers). With fewer registers, workers need to operate on smaller tiles of data, resulting in more instruction issues, SRAM to register IO, and potentially higher synchronization costs due to the increased data partitioning across workers.

Figure 6 shows the occupancy tradeoffs for attention kernels. We consider (1) a simple kernel that only uses warp-level parallelism (Listing 2) and (2) a kernel written in the LCSF template (Listing 3). Although with both kernels, performance increases with occupancy until resource contention dominates, LCSF expands the Pareto frontier beyond the naive kernel.

We find the general LCSF template to be effective across a range of AI workloads. We keep the template lightweight and simple by making optional design choices. However, we don't want TK to get in the way of achieving peak GPU performance – TK is embedded, meaning developers can use the full power of CUDA to extend the library as warranted.

**3.3 GRID PARALLELISM WITH BLOCK LAUNCH SCHEDULING**

TK makes it easier for users to quickly try varied grid layouts and coordinate thread block launches. This can help reduce the setup and tear-down costs for each thread block ( $C_{block}$  in Section 2.3), and encourage memory reuse between thread blocks, to avoid slow HBM accesses ( $C_{bank}$  in Section 2.3).

**Block launch costs.** We provide optimizations to minimize launch costs, centered around a persistent grid, where we launch thread blocks on the full set of SMs upfront, and simply load the next task for the thread within the existing block. We further eliminate pipeline bubbles by having load/store workers anticipate the next task and pre-load memory to prepare for future work, while the compute workers run the finish stage for the prior task. Table 2 shows these optimizations for matrix multiplications.

M × N × K	TK No.	TK Yes	CUTLASS
4096 × 4096	921	108	69
4096 × 128 × 161	184	113	113
4096 × 356 × 271	309	282	282
4096 × 512 × 414	450	407	407
4096 × 1024 × 565	600	633	633

Table 2: Persistent block launch THROPS for TK. GEMM kernels with (opt) persistent and without (no) persistent launch as we vary matrix dimension  $N$ .

**L2 reuse and block launch order.** Recall that thread blocks need to communicate via HBM. As introduced in Section 2.3, when thread blocks reuse memory, the data is often available in L2 cache, which is significantly faster than HBM. However, cache evictions mean that these reuse-qualities depend on the order in which blocks get launched. For our attention and GEMM kernels, we measure efficiency as we vary block order, summarized in Table 3. Block order substantially affects L2 reuse (measured through HBM bandwidth), which in turn can control kernel performance.

Figure 5: A simplified depiction of attention in the LCSF template to highlight the role of different specialized workers. Left is executed by workers that manage HBM to SRAM memory movement, and right by parallel compute workers, which operate in fast memory, registers and SRAM.

Figure 6: Occupancy tradeoffs. Left: Attention THROPS as a function of occupancy, benchmarked with lead dimension 64 and context length 4096. We compare synchronous and LCSF kernels.

**Question:** Assess the Methods section for Language & Expression issues.

**Explanation:** The paper introduces a key contribution, the ‘load-compute-store-finish’ template, and its acronym ‘LCSF’. This error introduces inconsistencies in this critical term: it’s defined as ‘LCS-F’ on page 6, called ‘LCSF’ in a figure title on page 7, and written out in full in the conclusion on page 10, while the original ‘LCSF’ acronym remains elsewhere. This terminological inconsistency for a central, paper-defined concept creates ambiguity and undermines the paper’s precision.

**Error Type:** LE (Language & Expression)

**Type:** Within-Generate

49

## E HUMAN-MACHINE CONSISTENCY EVALUATION

To evaluate whether GPT-4.1 accurately extracts detailed information from the model responses, we conduct a human-Machine consistency evaluation. We first randomly sampled 200 questions from the dataset. Then, we invited human experts to analyze the corresponding model-generated responses for these questions and to manually extract key information, including evidence sets, reasoning chains, and the number of unrelated errors. The results are presented in Table 4.

	$S_{total}$	$S_{location}$	$S_{reasoning}$	$P_{unrelated.err}$
Spearman’s correlation coefficients	0.841	0.806	0.842	0.954

Table 4: Spearman’s correlation coefficients for:  $S_{total}$ ,  $S_{location}$ ,  $S_{reasoning}$ , and  $P_{unrelated.err}$ .

In summary, GPT-4.1 can extract relevant evidence and reasoning steps with considerable accuracy, leading to precise evaluation scores.

In addition, we replaced GPT-4.1 with Qwen3-32B and Gemini 2.5 Flash to independently re-evaluate the same 200 samples. The results further confirm that our evaluation framework is not dependent on any particular LLM and exhibits strong robustness.

Table 5: Model performance under Qwen3-32B evaluation (scores scaled by 100).

Models	Avg.	RQD	DI	SG	MO	DHP	CF	IC	RCA	LE
<b>MLLM (Image Input)</b>										
Gemini 2.5 Pro	19.7	15.0	23.2	44.7	13.2	31.4	7.8	17.3	17.8	12.8
GPT-5	21.2	12.6	11.8	33.5	15.4	26.6	16.0	27.1	27.0	6.4
Grok 4	4.2	0.0	1.3	20.5	2.6	5.5	1.2	3.9	2.6	0.8
Doubao-Seed-1.6-Thinking	11.6	5.1	6.9	28.0	7.7	14.3	10.5	15.6	10.8	4.5
Doubao-Seed-1.6	12.0	4.8	5.9	36.2	7.5	13.3	5.7	19.9	9.9	5.5
<b>OCR + LLM (Text Input)</b>										
Gemini 2.5 Pro	35.0	26.6	39.6	53.9	31.4	56.2	15.9	35.6	39.0	10.0
GPT-5	25.3	19.0	28.3	28.3	24.1	38.8	10.3	32.2	31.8	3.1
Claude Sonnet 4	6.3	5.7	2.4	12.4	4.2	8.3	2.8	9.5	6.6	4.4
Grok 4	22.6	10.7	8.9	40.6	14.1	31.3	11.1	22.6	33.6	7.3
Doubao-Seed-1.6-Thinking	17.4	11.0	14.9	31.9	9.5	26.3	9.2	20.5	21.1	4.7
Doubao-Seed-1.6	15.3	6.4	9.8	31.7	10.8	22.4	8.5	21.9	17.8	2.4

Table 6: Model performance under Gemini 2.5 Flash evaluation (scores scaled by 100).

Models	Avg.	RQD	DI	SG	MO	DHP	CF	IC	RCA	LE
<b>MLLM (Image Input)</b>										
Gemini 2.5 Pro	14.6	9.1	11.3	34.2	11.4	28.2	4.9	12.6	14.8	5.3
GPT-5	20.5	11.2	12.0	32.4	17.9	27.4	10.5	25.9	27.4	3.1
Grok 4	3.9	0.5	1.8	15.9	1.7	4.1	1.5	1.1	4.4	0.0
Doubao-Seed-1.6-Thinking	10.3	3.2	4.7	26.4	7.3	14.2	9.2	14.8	9.3	3.1
Doubao-Seed-1.6	10.9	5.3	3.9	34.0	6.1	15.6	6.1	17.9	8.0	4.7
<b>OCR + LLM (Text Input)</b>										
Gemini 2.5 Pro	30.1	20.0	30.6	47.8	27.5	47.8	11.7	30.2	36.6	5.9
GPT-5	23.5	15.3	21.8	26.5	23.1	37.7	7.1	31.9	31.6	2.7
Claude Sonnet 4	5.7	3.8	1.4	11.1	3.9	9.4	2.0	8.7	6.5	3.1
Grok 4	20.2	9.2	7.8	36.1	11.5	32.4	8.0	20.7	30.6	5.8
Doubao-Seed-1.6-Thinking	15.9	8.4	11.1	30.9	10.1	23.7	6.3	19.9	20.2	3.5
Doubao-Seed-1.6	14.0	4.8	8.2	29.1	11.4	23.9	6.4	21.2	16.0	0.8

## F HYPERPARAMETER SENSITIVITY ANALYSIS

We conducted a sensitivity analysis of all 4 hyperparameters involved in scoring. We varied each independently and re-computed  $S_{\text{total}}$  across 11 proprietary model configurations. The results demonstrate that our evaluation metric exhibits strong robustness.

Table 7: Sensitivity under Image input: varying  $\lambda$  and  $\mu$  (scores scaled by 100).

<b>Model</b>	$\lambda=0.6$	$\lambda=0.8$	$\lambda=1.0$	$\mu=0.85$	$\mu=0.9$	$\mu=0.95$
GPT-5	19.3	19.2	19.0	18.5	19.2	19.9
Gemini 2.5 Pro	15.8	15.6	15.3	15.0	15.6	16.1
Doubao-1.6-Thinking	10.4	10.2	10.0	9.9	10.2	10.5
Doubao-1.6	10.1	9.9	9.8	9.7	9.9	10.2
Grok 4	4.0	4.0	3.9	3.9	4.0	4.1

Table 8: Sensitivity under Image input: varying  $\gamma$  and  $q$  (scores scaled by 100).

<b>Model</b>	$\gamma=0.4$	$\gamma=0.6$	$\gamma=0.8$	$q=1.0$	$q=1.5$	$q=2.0$
GPT-5	19.4	19.2	19.0	19.3	19.2	19.1
Gemini 2.5 Pro	15.7	15.6	15.5	15.6	15.6	15.5
Doubao-1.6-Thinking	10.4	10.2	10.0	10.3	10.2	10.1
Doubao-1.6	10.1	9.9	9.8	10.0	9.9	9.9
Grok 4	4.0	4.0	4.0	4.0	4.0	4.0

Table 9: Sensitivity under Text input: varying  $\lambda$  and  $\mu$  (scores scaled by 100).

<b>Model</b>	$\lambda=0.6$	$\lambda=0.8$	$\lambda=1.0$	$\mu=0.85$	$\mu=0.9$	$\mu=0.95$
Gemini 2.5 Pro	30.6	30.2	29.9	29.1	30.2	31.5
GPT-5	22.7	22.5	22.3	21.4	22.5	23.7
Grok 4	21.1	20.8	20.6	20.2	20.8	21.4
Doubao-1.6-Thinking	15.6	15.3	15.0	14.8	15.3	15.8
Doubao-1.6	14.1	13.9	13.7	13.6	13.9	14.3
Claude Sonnet 4	6.0	5.9	5.8	5.6	5.9	6.1

Table 10: Sensitivity under Text input: varying  $\gamma$  and  $q$  (scores scaled by 100).

<b>Model</b>	$\gamma=0.4$	$\gamma=0.6$	$\gamma=0.8$	$q=1.0$	$q=1.5$	$q=2.0$
Gemini 2.5 Pro	30.7	30.2	29.9	30.5	30.2	30.1
GPT-5	23.4	22.5	21.9	23.0	22.5	22.2
Grok 4	21.0	20.8	20.7	20.9	20.8	20.8
Doubao-1.6-Thinking	15.6	15.3	15.1	15.5	15.3	15.2
Doubao-1.6	14.2	13.9	13.7	14.1	13.9	13.8
Claude Sonnet 4	6.3	5.9	5.6	6.1	5.9	5.7