Finding Needles in Images: Can Multimodal LLMs Locate Fine Details?

Parth Thakkar[1*] Ankush Agarwal[1*] Prasad Kasu[1] Pulkit Bansal[1,2]
Chaitanya Devaguptapu[1]
[1]Fujitsu Research of India [2]Indian Institute of Technology, Patna

ACL 2025 VIENNA

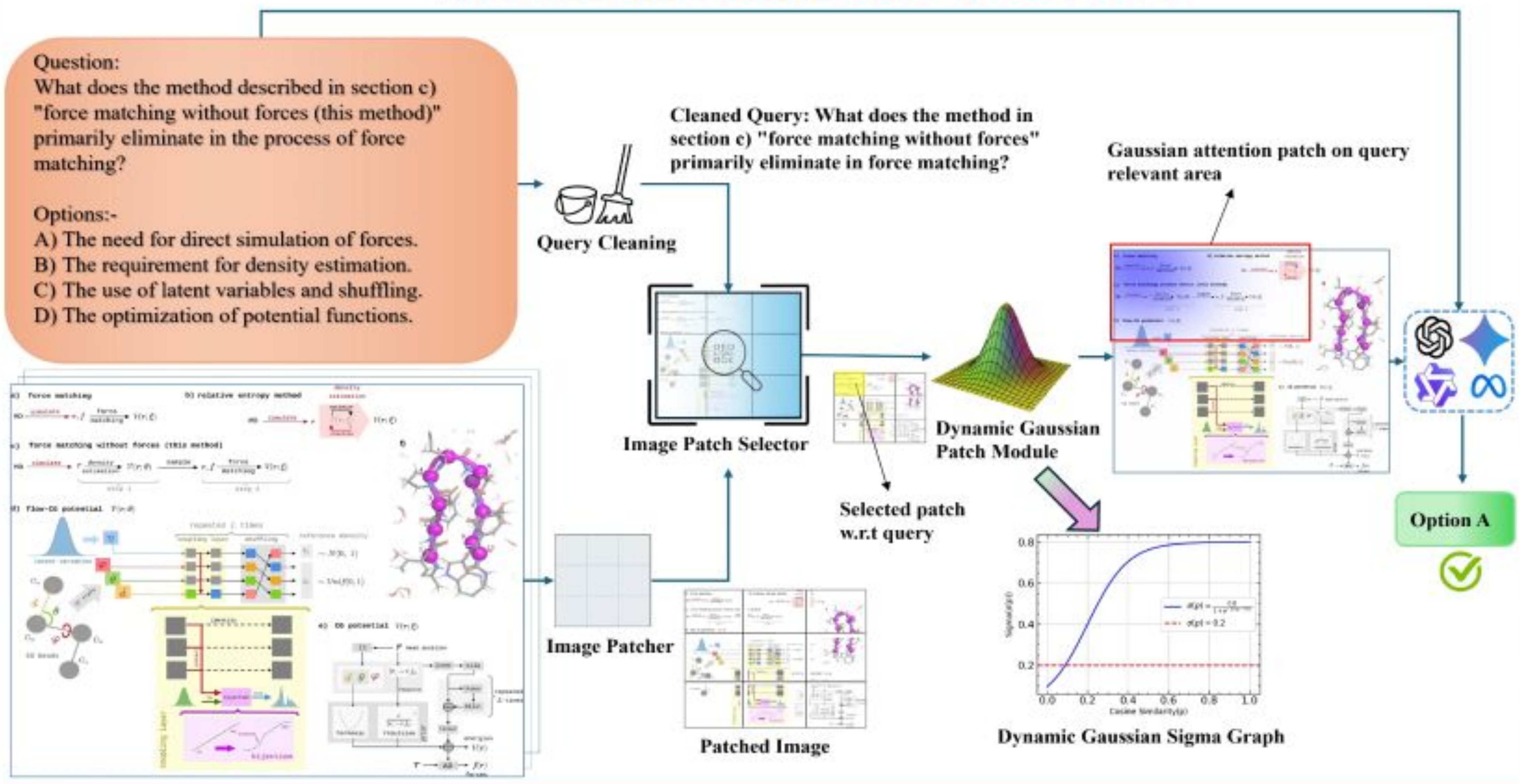
Motivation

•**MLLMs excel at global understanding but miss fine details**
They perform well on overall document comprehension but often fail to locate small, specific regions needed to answer fine-grained queries.
•**Fine-grained information is key in real-world scenarios**
Practical use cases like spotting prices in menus or footnotes in articles, require identifying tiny yet critical details in complex layouts.
•**Current benchmarks overlook fine-grained reasoning**
Existing datasets focus on global understanding and don't explicitly test models' ability to reason about localized, detailed information.

Contributions

•**NiM Challenge & Benchmark:** *We introduce the Needle in an Image (NiM) task and release NiM-Benchmark to evaluate MLLMs on fine-grained detail localization across diverse document types.*
•**Spot-IT Method:** *We propose Spot-IT, a plug-and-play approach that enhances fine-grained reasoning via question-guided dynamic attention, requiring no model changes.*
•**SOTA Results:** *Spot-IT achieves up to* **21.05%** *improvement over GPT-4o, setting new baselines for fine-grained detail extraction in DocVQA.*

Architecture Diagram of Spot-IT



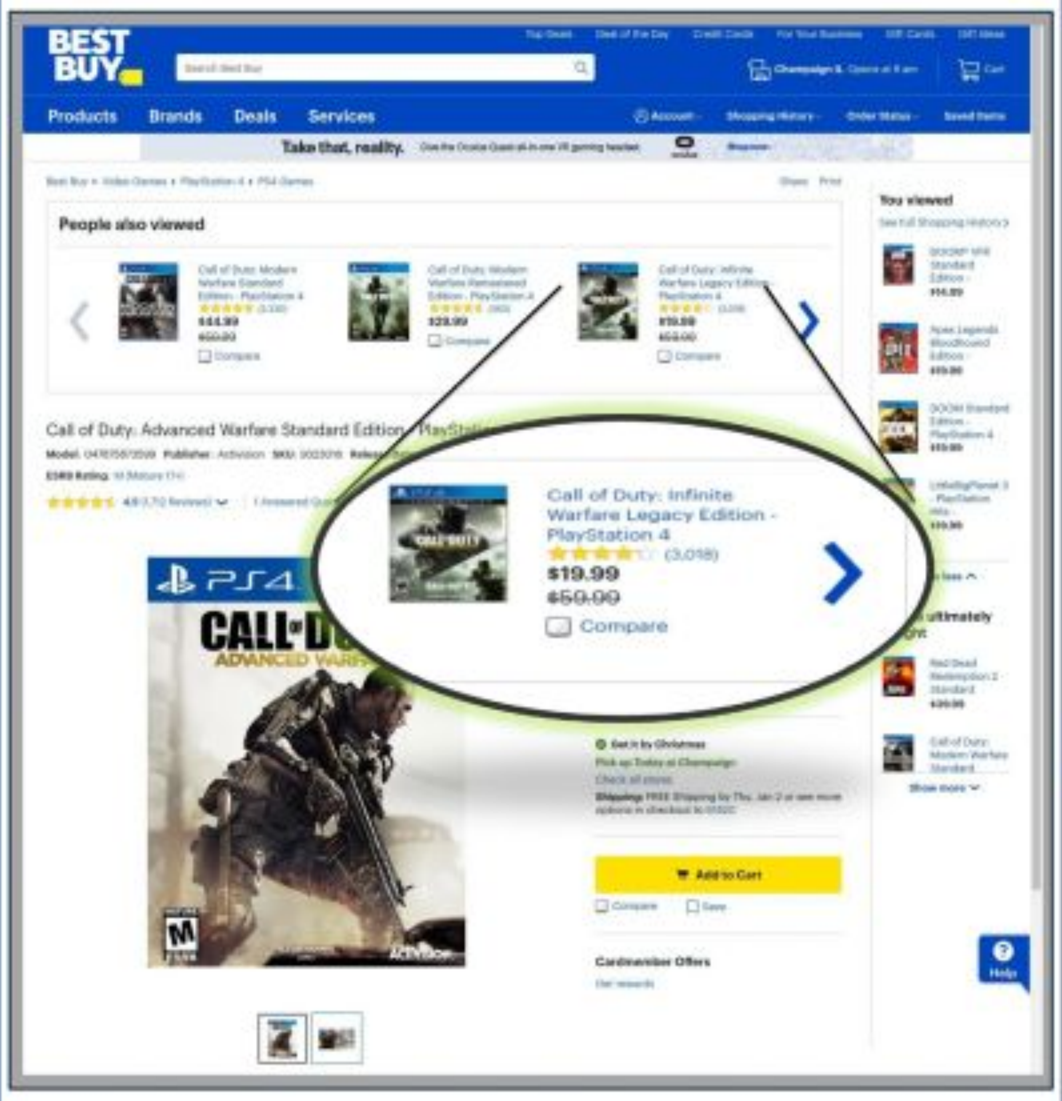
Evaluation of Spot-IT on existing benchmarks

Methods	ArxiVQA	DUDE		
	Acc.(↑)	EM(↑)	F1(↑)	ANLS(↑)
Closed-Source LLMs (zero-shot)				
GPT-4o	0.52	0.42	0.56	0.55
GPT-4o-mini	0.47	0.34	0.50	0.47
Gem-1.5-Flash	0.53	0.30	0.42	0.42
GPT-4o+OCR	0.41	0.34	0.47	0.47
GPT-4o+CoT	0.51	0.43	0.57	0.58
GPT-4o+Ours	**0.60**	**0.45**	**0.60**	**0.60**
GPT-4o-mini+Ours	0.52	0.41	0.55	0.52
Gem-1.5-Flash+Ours	0.54	0.34	0.47	0.45
Open-Source LLMs (zero-shot)				
Llama-3.2-VL-11B	0.41	0.13	0.23	0.18
Qwen2-7B	0.44	0.21	0.32	0.28
Llama-3.2+OCR	0.38	0.05	0.19	0.08
Llama-3.2+CoT	0.42	0.11	0.23	0.17
Llama-3.2+Ours	0.44	0.19	0.29	0.24
Qwen2-7B+Ours	**0.44**	**0.27**	**0.37**	**0.32**

Spot-IT evaluation results compared with baselines adapted from M3DocRAG.

NiM Benchmark: Examples

Query: For which console is Call of Duty Legacy Edition game available?



Query: "What is The price of Chips & Gravy

Spot-IT: Algorithm

Algorithm 1 Spot-IT: Query-Guided Attention for Document Understanding

Require: Document image I, query q, grid size n, Multi-modal LLM L
Ensure: Answer a to the query
1: Clean q to obtain q_c; Segment I into $n \times n$ grid $\{P_{i,j}\}$
2: $v_q \leftarrow L(q_c)$
3: **for** each patch $P_{i,j}$ **do**
  $v_{i,j} \leftarrow L(P_{i,j})$; $s_{i,j} \leftarrow \frac{v_{i,j} \cdot v_q}{\|v_{i,j}\| \|v_q\|}$
4: **end for**
5: $(i^*, j^*) \leftarrow \arg\max_{i,j} s_{i,j}$; $p \leftarrow \frac{\exp(s_{i^*,j^*})}{\sum_{i,j} \exp(s_{i,j})}$
6: $x^* \leftarrow \frac{(2i^*-1)H}{2n}$, $y^* \leftarrow \frac{(2j^*-1)W}{2n}$
7: $\sigma \leftarrow \frac{1}{1+\exp(-10(p-0.2))}$; $M(x,y) \leftarrow \exp\left(-\sqrt{\frac{(x-x^*)^2+(y-y^*)^2}{2\sigma^2}}\right)$
8: $I'(x,y) \leftarrow (1-\alpha M(x,y))I(x,y) + \alpha M(x,y)H(x,y)$
9: $a \leftarrow L(q, I')$
10: **return** a

Key Takeaways

- **MLLMs lack precision**
Current models struggle to locate and reason about small, detail-rich regions in complex documents.
- **Human-model gap persists** *Humans outperform MLLMs in accuracy for fine-grained document tasks, though with higher latency.*
- **Improvement areas**
Future work should enhance semantic similarity methods, and introduce more fine grained complex reasoning tasks.

NiM Benchmark: Evaluation

Model	GPT-4o			GPT-4o-mini			Gemini-1.5-Flash			Qwen2-7B		
	EM	F1	ANLS	EM	F1	ANLS	EM	F1	ANLS	EM	F1	ANLS
Baseline	0.38	0.48	0.56	0.29	0.38	0.46	0.22	0.28	0.37	0.07	0.10	0.19
Ours	**0.46**	**0.56**	**0.62**	**0.35**	**0.44**	**0.50**	**0.27**	**0.34**	**0.40**	**0.11**	**0.15**	**0.22**

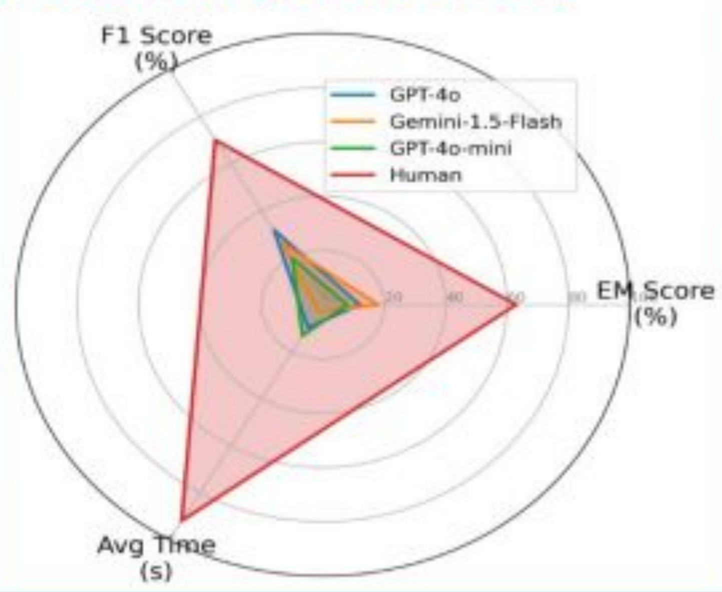
Performance remains modest, underscoring the benchmark's difficulty and the need for improved models.

NiM Benchmark: Statistics

Dataset Statistics

Domains	6	Categories	6
Pages/Images	2,970	Questions	1,180
Document Categories:			
Newspapers	(22)	Academic Papers	(32)
Magazines	(17)	Lecture Shots	(50)
Web Shots	(100)	Menus	(60)
Question Statistics		Answer Statistics	
Max Length	26	Max Length	19
Avg Length	10.96	Avg Length	1.92

NiM Benchmark: Accuracy vs Time



LinkedIn

Paper Blog